THESIS FOR THE DEGREE OF DOCTOR OF ENGINEERING

Statistics in Gene Expression, Metabolomics, and Comparative Genomics in Evolution

Alexandra Jauhiainen

CHALMERS | UNIVERSITY OF GOTHENBURG

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg
Göteborg, Sweden 2010

Statistics in Gene Expression, Metabolomics, and Comparative Genomics in Evolution
Alexandra Jauhiainen
ISBN 978-91-7385-419-1

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie nr 3100
ISSN 0346-718X

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX.
Printed in Göteborg, Sweden 2010

Abstract

This thesis contains four papers concerning (I) the evolutionary conservation of drug targets and its potential use in environmental risk assessments, (II) RNA degradation as a control mechanism during osmotic stress in the yeast *S. cerevisiae*, (III) the localization and effects of the gene DDIT3 encoding a key regulator of stress response, and (IV) the integration and analysis of transcriptional and metabolic data to identify active metabolic pathways.

Environmental risk assessments are needed for the approval of new pharmaceutical compounds. To date, the risk assessments have mainly been focused on organisms like algae and *Daphnia*. The conservation of drug targets in species relevant for ecotoxicity testing is a key aspect in developing more targeted test strategies on higher organisms like fish or amphibians. With information on predicted proteomes for a wide range of species it is possible to extract data on evolutionary conservation for drug targets. In paper I, orthology data is compiled and analyzed for a set of human drug targets in several species, and the result evaluated based on an extensive literature search.

mRNA degradation can be investigated on a genome-wide scale with the use of a transcriptional inhibitor and subsequent hybridization of RNA pools, isolated at a set of time-points, to microarrays. Due to the complexity of the microarray methodology in this context, the data are in need of processing and transformation to deduce relevant information on changes in degradation rates. In paper II, mRNA degradation is investigated as a post-transcriptional control effect in connection to hyperosmotic stress. We conclude that mRNA degradation mechanisms are important regulatory keys in the stress response.

The gene *DDIT3* encodes a protein acting as a regulator of the stress response within human cells. For example DNA damage, hypoxia, and starvation are stress types inducing *DDIT3* transcription. DDIT3 is a transcription factor and has mainly been reported as a nuclear protein. In paper III, the effects and target genes of DDIT3 are investigated using techniques like microarrays, RT-qPCR, and various bioinformatical and statistical methods. We report that DDIT3 also can be localized to the cytoplasm, and induces or represses different genes compared to the nuclear form. The cytoplasmic form of DDIT3 is involved in migration, and inhibits the migratory effects of fibrosarcoma cells.

The development of different 'omics' technologies in molecular biology has resulted in several methods to characterize cells and tissues, for example microarrays to characterize the transcriptome (collection of gene transcripts) and spectrometry techniques like NMR to describe the metabolome (collection of small molecules). Interpretation of different 'omics' data is usually done sepa-

rately, and often with respect to pathways, which are sets of reactions involving genes, metabolites, and proteins. A common research question is to deduce which pathways are active (regulated) when comparing two or several conditions. In paper IV, we propose a model to make such pathway level decisions by integrating transcriptomic and metabolomic data.

Acknowledgments

I would like to thank my supervisor Olle Nerman for guiding me in, but also letting me on my own explore, the field of applied statistics and bioinformatics. I am grateful for all the support, valuable ideas, and comments you have provided throughout my PhD.

I want to thank all my co-authors, particularly my main co-authors Claes Molin, Pierre Åman, and Lina Gunnarsson. I also want to express my thanks to Rebecka Jörnsten and to George Michailidis for our collaboration and the advice and ideas both of you provided for the final paper of this thesis.

I wish to extend my thanks to Simon Tavaré, for letting me visit his computational biology research group at the University of Cambridge, and for all the interesting and valuable research collaborations that sprung from those visits.

Further, I want to thank all current and former colleagues at the department. Many thanks to friends far and wide, and finally, I want to thank all my family, especially Kai and Helena, for their support and friendship.

List of Papers

The thesis includes the following papers.

I. Gunnarsson, L.*, **Jauhiainen, A.***, Kristiansson, E., Nerman,O., and Larsson, D.G.J. (2008). Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments. *Environ. Sci. Technol.* **2008**, *42*, 5807-5813.
   * Equal contribution.

II. Molin, C., **Jauhiainen, A.**, Warringer, J., Nerman, O., and Sunnerhagen, P. (2009). mRNA Stability Changes Precede Changes in Steady-State mRNA Amounts During Hyperosmotic Stress. *RNA* **2009**, *15*, 600-614.

III. **Jauhiainen, A.**, Thomsen, C., Strömbom, L., Grundevik, P., Andersson, C., Danielsson, A., Andersson, M.K., Nerman, O., Rörkvist, L., Ståhlberg, A., and Åman, P. (2010). Subcellular Localization and Effects of DDIT3/GADD153/CHOP. *Submitted.*

IV. **Jauhiainen, A.**, Nerman, O., Michailidis, G., Jörnsten, R. (2010). Transcriptional and Metabolic Data Integration and Modeling for Pathway Identification. *Working Paper.*

Contents

1 Introduction **1**

1.1 Microarrays 1

1.1.1 The Microarray Technology 2

1.1.2 Microarray Data Preprocessing and Analysis 3

1.1.3 Applications 5

1.2 Bioinformatics 7

1.2.1 Biological Databases 7

1.2.2 Sequence Analysis 8

1.3 Metabolomics 10

1.3.1 Separation and Detection Techniques 10

1.3.2 Analysis Methods 11

1.3.3 Applications 12

2 Technical Aspects of mRNA Decay Measurements **13**

2.1 Agents used for transcriptional arrest 14

2.2 mRNA Decay Models 15

2.3 Using Microarrays to Measure mRNA Decay 16

2.3.1 Experimental setups 16

2.3.2 Hybridization schemes 18

2.3.3 Normalization 18

2.3.4 Data transformation 19

3 Evolutionary Conservation Studies 21

3.1 Homology, Orthology, and Paralogy 21

3.2 Ortholog Prediction 25

3.2.1 Phylogeny-based Methods 25

3.2.2 BLAST-based Methods 25

3.2.3 Methods based on Evolutionary Distance 26

3.2.4 Synteny information 26

4 Regression Models and Shrinkage Methods 27

4.1 Regression Models 27

4.2 Selection Methods 28

4.3 Shrinkage Methods 28

4.3.1 Nonnegative Garrote and Ridge Regression 29

4.3.2 The Lasso 30

4.3.3 Other Shrinkage Methods 33

4.3.4 Dantzig 34

4.4 Model Selection via Rate-Distortion 34

5 Summary of Papers 37

Paper I: Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments 37

Paper II: mRNA Stability Changes Precede Changes in Steady-State mRNA Amounts During Hyperosmotic Stress 40

Paper III: Subcellular Localization and Effects of DDIT3/GADD153/CHOP 42

Paper IV: Transcriptional and Metabolic Data Integration and Modelling for Pathway Identification 44

Bibliography 46

Chapter 1

Introduction

This chapter will introduce concepts in microarray analysis, bioinformatics, and metabolomics. The information can be useful as a reference in reading the forthcoming chapters.

1.1 Microarrays

The microarray technology was introduced fifteen years ago in a landmark paper (Schena et al., 1995). The array described in the paper contained merely 45 genes, although the authors recognized the potential of the technology and mentioned arrays with over 20,000 spots. The spots on the glass slide contained complementary DNA (cDNA) from *Arabidopsis*, as well as three control cDNAs from other organisms. Differential expression of the *Arabidopsis* genes under investigation were detected with competitive two-color fluorescence hybridization of two RNA sources.

The microarray technology has developed fast since its introduction and several different microarray platforms are available today, many of which offer arrays with very high spot density. Parallel to the technological development of microarrays, a lot of data analysis methods have been presented to deal with the statistical problems connected to microarray data analysis.

1.1.1 The Microarray Technology

The array briefly described above was a spotted cDNA array, which today still are commonly used. The microarray itself is a glass slide with material (sometimes referred to as probes), usually cDNA clones or PCR products, spotted onto the slide in a grid pattern. mRNA from two sources of interest (e.g., cancer and normal cells) are reverse transcribed into cDNA and labeled with two different fluorescent cyanine dyes, commonly Cy3 and Cy5. The two labeled cDNA sources are then hybridized to the microarray in a competitive manner.

After hybridization, the microarray is scanned with a laser scanner at two different wavelengths designed to excite the fluorophores Cy3 and Cy5 into emission of green and red light, respectively. The two images produced in the scanning step are overlaid, giving a pseudo-image, which shows spots as either yellow, red, or green.

To extract more detailed information on the expression status for each gene, further analysis of the microarray images is needed. The processing can be divided into three parts: gridding, segmentation, and intensity extraction (Yang et al., 2001). Briefly, gridding, or addressing, concerns locating and giving coordinates to each spot, segmentation deals with assigning pixels as either within the spot or as background, and intensity extraction means calculating red and green intensities for each spot. In the last step, quality measures of the spot are usually also extracted.

The two-color spotted microarray technique is frequently used throughout the research community. Many research groups manufacture in-house arrays which usually are cheaper than commercialized platforms, but may suffer from quality problems (Bammler et al., 2005). Such quality issues are for example dye-biases, degradation of fluorescense signals, and poor reproducibility between replicates.

As a contrast to the the cheaper spotted arrays, several commercial microarray platforms are available. The perhaps most popular type is the single channel oligonucleotide Affymetrix arrays. Oligonucleotides, or oligos for short, are short stretches of nucleotides designed to match a specific sequence or group of sequences, most common genes or gene families. The array is built up of multiple probes (typically 11-20) of length 25 for each gene, which are synthesized directly on the array. With the Affymetrix technology, only one RNA source is hybridized onto an array, using only one fluorescent dye, and comparisons between RNA sources have to be made with several hybridizations onto different arrays.

The more customizable Agilent technology is based on single 60-mer oligonucleotides and work in the two-color competitive hybridization fashion. The probes are synthesized on the array using inkjet printing, which is cheaper than the method employed by Affymetrix.

Another commercial technology is the Illumina bead array platform. The support for the DNA material consists of microscopic beads, instead of a solid (glass) surface. In general, the strength of the commercial platforms are their ready-made protocols and support, as well as overall high quality (Bammler et al., 2005; Irizarry et al., 2005).

1.1.2 Microarray Data Preprocessing and Analysis

For a two-color array, the red and green intensities for each spot on the array are further processed to extract relevant information on differential expression.

Preprocessing of the microarray data involves background correction and different normalization steps. Background intensities for each spot are extracted in the image analysis step with the foreground red and green intensities. A common assumption is that the background intensities are additive to the foreground intensities:

$$R_g = R_g^f - R_g^b \quad \text{and} \quad G_g = G_g^f - G_g^b$$

R_g^f and R_g^b symbolize the red foreground and background intensities for gene g respectively (and similarly for the green intensities). In a review paper (Ritchie et al., 2007), eight different background correction methods were compared with different estimates for R_g^b and G_g^b and variants of subtraction. To simply subtract the background intensities from the foreground intensities proved to be even worse than no background correction at all. An improvement of the stabilization of the variance as a function of intensity characterized the best performing methods.

A plethora of normalization methods exists for microarray data. The normalization step is crucial to remove intensity dependent trends and to achieve comparability between arrays. To apply different normalization methods, the R and G intensities are usually transformed to $M = \log_2 R - \log_2 G$ and $A = (\log_2 R + \log_2 G)/2$. M is called the $\log_2$-fold change and A the average $\log_2$-intensity.

The most common method for within-array normalization is probably the loess normalization. In its simplest form, a locally-weighted polynomial regression is used to adapt a smoothing curve to the A-values and each M-value is normalized by subtracting the value of the loess-curve at the given average intensity.

To ensure comparability between arrays, or sets of arrays, between-array normalization is sometimes applied. A quite simple normalization of this type is scale-normalization, which results in a series of arrays having the same median absolute deviation.

A large set of different methods for normalization and background correction exists for single-channel microarray data. Different methods, including MAS5 and gcRMA, are compared in Qin et al., 2006.

Ensuing the preprocessing steps, different ranking methods can be used to assess differential expression for the genes on the arrays. Assume that we are interested in finding differentially expressed genes between two conditions. A set of n replicate arrays are hybridized to assess this difference. Let μ_g denote the true $\log_2$-fold change in expression for gene g. We want to test the hypothesis

$$H_0 : \mu_g = 0 \quad \text{vs.} \quad H_1 : \mu_g \neq 0$$

for each gene. The number of genes, N, is typically large (thousands) while the number of replicate arrays n rarely exceeds ten. Let M_{gj} denote the $\log_2$-fold change for gene g on array j and s_g the gene specific standard deviation over the replicate arrays. The ordinary t-statistic for differential expression is in this setting

$$t_g = \frac{\overline{M}_{g.}}{s_g/\sqrt{n}},$$

where $\overline{M}_{g.}$ is the average $\log_2$-fold change over the replicate arrays. The problem in using the ordinary t-statistic in microarray experiments is that the number of replicates is quite small, and a small variance might occur by chance even if the $\log_2$-fold changes M_{gj} are small. Several methods have been proposed to circumvent this problem. The Efron t-statistic is a slight modification of the ordinary t-statistic:

$$t_g = \frac{\overline{M}_{g.}}{(a_0 + s_g)/\sqrt{n}}$$

where a_0 is the value of the 90%-quantile of all the gene specific standard deviations $s_1, \ldots, s_N$ (Efron et al., 2001). Other offset statistics have also been proposed, for example the statistic used in the S-test which is implemented in the SAM software (Tusher et al., 2001).

The development of the regularized t-test (Baldi and Long, 2001) and the B-test (Lönnstedt and Speed, 2002) marked a more model-based approach to assess differential expression. The B-test was a log-odds test and was later reformulated and generalized into a moderated t-test (Smyth, 2004). To deduce the moderated t-statistic, an empirical Bayes approach is adopted with the following distributional assumptions:

$$M_{gj}|\mu_g, \sigma_g^2 \sim N(\mu_g, \sigma_g^2)$$
$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$$

A model is fit to every gene, and to use this parallel structure, a hierarchical model is adapted with priors for the hyperparameters. The prior distribution on σ_g^2, describing how the variances vary across genes, is

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_o}^2,$$

with d_0 degrees of freedom and s_0^2 the prior estimate of the variance (using Γ-distributions, d_0 may be interpolated). Under this hierarchical model, the posterior mean of σ_g^{-2} given s_g^2 is

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

and the moderated t-statistic is defined as

$$\tilde{t}_g = \frac{\overline{M}_{g.}}{\tilde{s}_g/\sqrt{n}}$$

Under the null hypothesis, the moderated t-statistic follows a t-distribution with $d_g + d_0$ degrees of freedom. An effect of this model is a shrinkage of the observed variances towards the prior values, depending on the observed and prior degrees of freedom. The moderated t-test and the underlying hierarchical model is formulated in a more general context as linear combinations of parameters from a linear model in Smyth, 2004.

The moderated t-statistic is implemented in the package LIMMA (Smyth, 2005) in R as a part of the Bioconductor suite. Several different normalization methods and other functions developed to deal with microarray data are also implemented in the package.

1.1.3 Applications

Gene expression analysis and transcriptional profiling were the starting points of the microarray technology, for example in Schena et al., 1995, where the expression (levels of mRNA) between root and leaf were compared in *Arabidopsis* for 45 genes. Another example of early usage of the microarray technology was the identification of cell cycle regulated transcripts in yeast (Spellman et al., 1998). Today, microarrays are routinely used in many studies to address a wide range of scientific questions.

Besides 'standard' gene expression profiling, adaptations of the microarray technology have been made to measure other transcriptional features like mRNA stability (e.g., Wang et al., 2002) and polysomal association to transcripts as a measure of translational activity (e.g., Preiss et al., 2003). General problems with mRNA decay measurements using microarrays are discussed in Chapter 2.

A different version of a microarray is a tiling array, where a whole-genome, or a segment of the genome, is represented by oligonucleotide probes at a certain spacing (resolution). Both the Affymetrix, Agilent, and Nimblegen platforms offer commercial tiling arrays. A characterization of the transcriptome (boundaries, structures, and levels of coding and non-coding transcripts) have been made in yeast using tiling arrays covering the whole yeast genome (David et al., 2006). The study identified complex transcriptional patterns, and also mapped the untranslated regions (UTRs) of expressed transcripts. Whole-genome tiling arrays have also been used in several studies on *Arabidopsis*, for example to investigate the transcriptional pattern associated with circadian rhythms (Hazen et al., 2009).

The ENCODE (ENCyclopedia Of DNA Elements) project was launched in 2003 to identify all functional elements in the human genome, and a pilot study was published three years ago (The ENCODE Project Consortium, 2007). The pilot study, involving (among many other methods) tiling arrays covering 1% of the human genome, brought more understanding on transcription start sites and regulatory mechanisms, as well as further understanding of chromatin structure.

Array-based comparative genomic hybridization (array-CGH) is another microarray format which is used to analyze DNA copy number variations (CNVs) and larger genomic gains and losses. The array can be either a spotted, or a tiling array, and genomic material from sample and control are hybridized together. The array-CGH technology was for example combined with whole-genome sequencing data to profile 30 Asian females in order to develop a comprehensive common CNV map for Asian populations (Park et al., 2010).

In contrast to CNVs, which are large (1000 bp or more) structural variations in the genome, single nucleotide polymorphisms (SNPs) are variations that occur in a single nucleotide. Different versions of SNP arrays are available, for example from Illumina. The Illumina SNP array technology is based on a probe which targets the genomic location adjacent to a SNP site (LaFramboise, 2009).

Chromatin immunoprecipitation (ChIP) combined with microarray analysis (so called ChIP-on-chip), is another application of the microarray technology (Ren et al., 2000). ChIP-on-chip is used to investigate interactions between DNA and proteins, and can be employed to identify transcription factors. Briefly, the method consists of extracting proteins by epitope binding via an antibody and the DNA bound to the protein is purified, labeled, and hybridized to an array. Replacing the array technology by next generation sequencing techniques, creating a technique called ChIP-seq, has proved to be more cost-efficient. The ChIP-seq technology generally also requires less input material and offers a rapid analysis pipeline (Mardis, 2007).

1.2 Bioinformatics

In short, bioinformatics is the development and application of computational and statistical tools to biological or medical data in some form. Areas falling under the bioinformatics umbrella include sequence analysis, structural bioinformatics, comparative genomics, and phylogenetic studies, just to mention a few.

1.2.1 Biological Databases

Biological databases play a crucial role in the bioinformatics field. Nucleic Acids Research publishes a yearly database issue (Cochrane and Galperin, 2010), now listing over 1200 publicly available databases, of which only a few are briefly described below. The database issue covers databases on nucleotide sequences, protein sequences, structures, metabolomic and signaling pathways, and whole genomes.

GenBank is a large database (Benson et al., 2008) which has been available for more than 25 years. The database, hosted by the National Center for Biotechnology Information (NCBI), contains nucleotide sequences for more than 260 000 different species (the June 15 release of 2010 reports that 120 604423 sequences are available). GenBank can be accessed through a sophisticated retrieval system called Entrez, which also integrates data from other major databases.

Other large and frequently used databases are for example Ensembl (Hubbard et al., 2009) and Uniprot (The UniProt Consortium, 2010). Ensembl is a genome database, with information on more than 50 eukaryotic species, and with the pre-release of the baboon genome its newest addition. The UniProt database is a resource on protein sequences and functional annotation. A part of UniProt is a manually curated protein knowledge database.

The Gene Ontology (GO) (The Gene Ontology Consortium, 2010) is a widely used resource of controlled and consistent vocabularies for annotation of gene products. The three main ontologies consist of a terminology describing the molecular function of gene products, their associated biological processes, and their cellular localization. The GO Consortium also manages and integrates annotation information of gene products, i.e., connecting gene products to different terms in the ontologies, and develops tools for GO annotation.

The Kyoto Encyclopedia of Genes and Genomes, KEGG for short (Kanehisa et al., 2010), is a collection of pathway information for a large set of species. A pathway is made up of a set of reactions involving genes, metabolites, and proteins. The pathways can be viewed as functional modules describing the network of molecular interactions in the cell.

1.2.2 Sequence Analysis

Automated genome annotation and comparative genomics are heavily reliant on sequence analysis. A large part of sequence analysis is concerned with sequence alignments.

The Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1997; Altschul et al., 1990) was first introduced twenty years ago. It has become a hugely popular tool to search and compare both protein and DNA sequences against query databases. BLAST searches for segments of sequences in a database that matches segments of the query sequence by optimizing a local similarity measure. The search algorithm uses a modified version of the Smith-Waterman algorithm (Smith and Waterman, 1981) and is considerably faster, which has greatly contributed to its popularity.

The BLAST algorithm can shortly be divided into three steps:

1. Find word pairs of a specified length W with scores above a certain threshold T. A word pair is a 'hit' between a subsequence of length W in the query sequence and a subsequence in the database with score larger than T.
2. Represent the query sequence and the database hit in a matrix, with a word hit on the diagonal. Multiple hits on the same diagonal within a

certain distance will be extended into high scoring pairs (HSPs).

3. The highest scoring HSPs are extended, including gaps, until their overall score fall under a threshold.

Mutation of individual nucleotides, and in effect changes of amino acids, combined with insertions and deletions, make up the basis for sequence evolution. Substitution matrices are commonly used to measure similarity between diverged sequences. Such a matrix gives a score to every substitution that is possible in a protein or DNA sequence. For proteins, some substitutions of amino acids are more likely than others, due to similar chemical properties of the amino acids.

The scoring scheme in BLAST, used to assess scoring for word pairs and extended hits, is based on substitution matrices and also includes penalties for gaps. The default matrix for protein comparisons used in the current versions of BLAST is the BLOSUM62 matrix (Henikoff and Henikoff, 1992) but other versions of the BLOSUM matrices may be used, as well as PAM matrices (Dayhoff et al., 1978). The similarity matrix for comparisons between DNA sequences is of a much simpler type, with all mismatches scored identically (Altschul et al., 1990).

In contrast to the local similarity algorithm in BLAST, which identifies high-scoring subsequences in a database to a query sequence, global similarity algorithms aim to maximize a similarity score for an overall alignment between two sequences. A global alignment algorithm for pairs of sequences is the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) which uses dynamic programming and guarantees an optimal alignment given a substitution matrix and gap penalties.

Clustal is one of the most used tools in bioinformatics to perform multiple alignments. Although it is not guaranteed to find the highest scoring alignment, the method is quite fast. The most recent version of Clustal is in the 2.0 series (Larkin et al., 2007), but the original Clustal program was introduced more than 20 years ago (Higgins and Sharp, 1988). The multiple alignment algorithm is a more heuristic approach than the Needleman-Wunsch algorithm: first, all pairs of sequences are aligned pairwise to calculate a matrix of distances between the sequences. Second, a phylogenetic tree is calculated based on the distance matrix, and thirdly, the sequences are progressively aligned in turn (using consensus sequences in the intermediate steps), according to the branching order in the tree.

Clustal can be downloaded as command-line version (ClustalW) or a graphical user interface (ClustalX). Clustal is also available as a web-service via several different mirrors.

1.3 Metabolomics

In analogy to other 'omics' fields, like genomics and proteomics, metabolomics is the analysis and study of the set of metabolites in a cell, organ, or tissue. A metabolic profile is, in parallel to a transcriptional profile, a characterization of concentrations of different molecular species. Instead of measuring levels of mRNA transcripts from genes, a metabolic profile holds information on concentrations of different small molecules (metabolites) in the cell, e.g., sugars, amino acids, organic acids, and vitamins.

1.3.1 Separation and Detection Techniques

To detect and quantify metabolites, separation techniques like gas or liquid chromatography, followed by quantification by mass spectrometry (GC-MS, or LC-MS) are often used. Nuclear magnetic resonance spectroscopy (NMR) is also frequently employed and has some appealing properties. The NMR technique is non-destructive, in the sense that it does not "destroy" the samples during the analysis process. Hence, NMR is useful when analyzing tissues or when sequential analysis of samples is required (Nicholson and Lindon, 2008).

In order to use the separation based technique GC-MS, the samples need to be derivatized (gaseous form), and may thus introduce some bias towards detection of certain types of compounds. The LC-MS technique offers very good sensitivity, but is unfortunately not quantitative in the same was as NMR or GC-MS (Wishart, 2010). In general, the mass spectrometry based methods do offer more sensitivity than NMR.

The result from detection of metabolites in a sample, with either of the techniques, is a spectrum. In the case of NMR, the spectrum is the Fourier transform of a mixture of decaying sinus like waves. The spectrum consists of peaks of varying height and width. A specific molecule can correspond to one or several peaks, and the peaks are sometimes overlaid onto each other. The area under the different peaks correspond to, in an ideal setting, the concentration of the different metabolites.

The identification of metabolites in a spectrum is not trivial, especially since the positions (shifts) of the peaks can vary relative to each other with varying experimental conditions. These non-linear peak shifts need to be adjusted for in order to compare different samples. Preprocessing in the form of baseline and phase corrections, and reference peak adjustment need to be performed as well.

Mapping the different peaks uniquely to metabolites involves matching against reference spectra of some kind. Such spectra can either be derived separately in additional experiments, or collected from databases like the Human Metabolome Database (HMDB), which contains information on over 7900 metabolite entries (HMDB, 2010; Wishart et al., 2009). Although the database is extensive, the identification of individual metabolites is usually a very time-consuming task. Due to this, the number of uniquely identified metabolites in an experiment ranges from 50 to 200 metabolites, significantly less than for example the number of transcripts identified in a microarray experiment.

1.3.2 Analysis Methods

Since the identification of metabolites is such a difficult task, a common way to process metabolomics data is instead to analyze a compressed version of the original spectrum with multivariate methods like PCA or PLS. The spectrum is compressed by binning (integration of the spectrum signal in small segments) resulting in data of lower dimensionality, but also with lower resolution. The purpose of a PCA or PLS analysis is to prove that there are differences between different groups of samples, and to try to identify a smaller set of variables that explain differences between groups. One problem with this type of analysis is that the smaller set of variables identified by for example PCA is hard to interpret.

The interpretation of variables differing between samples that have been profiled is important in order to map them to interesting functional groups or biochemical pathways. This can be achieved in part by the use of correlation maps (Steuer, 2006). The correlations between the metabolites (within a sample) can be used to elucidate if certain enzymes are highly variable (and therefore induce negative correlations between their substrate and product metabolites). Comparisons of several experimental groups can be made with the correlation maps, as different groups should show different correlation fingerprints A drawback to this methodology is that the resulting correlation maps are very sensitive to data analysis issues, and spurious high correlations can appear if the different samples are not normalized properly against each other.

1.3.3 Applications

In plants, and especially the model organism *Arabidopsis*, metabolomics data has been used to address a number of scientific questions, among them characterization of the freezing tolerance response, and the downstream effects of transgenic manipulation of the transcription factor MYB1 (Last et al., 2007). Studies have also been published in which transcription data and metabolite data are correlated in order to achieve a more global understanding of the cellular processes (Gibon et al., 2006). This is in fact a very interesting research problem, which we address in paper IV.

An novel application of metabolite data is to genome-wide association (GWA) (Gieger et al., 2008). By measuring around 300 metabolites in male subjects, association of single nucleotide polymorphisms were found to variations in metabolite homeostasis, suggesting that common polymorphisms generate differentiations in the metabolome of the human population.

Recently, more efforts have been focused on understanding the metabolome of cancer cells, in order to gain understanding of the cellular physiology and biochemical activity of tumors (Abate-Shen and Shen, 2009). The prostate cancer metabolome was for example characterized by comparing normal prostate to prostate cancer in its metastatic form, and the analysis identified sarcosine as a potential biomarker (Sreekumar et al., 2009).

Chapter 2

Technical Aspects of mRNA Decay Measurements

Many papers have been published on mRNA degradation pathways and mechanisms in different model organisms during the last few years. Examples include the characterization of important deadenylases in yeast (Tucker et al., 2001) and the discovery of a functional link between decapping and deadenylation of mRNAs in mammalian cells (Yamashita et al., 2005).

A comprehensive review (Garneau et al., 2007) highlights the complexity and variation in the mRNA degradation signals and machinery. Six, partly overlapping, pathways of mRNA decay have been described. Three pathways monitor and execute the degradation of normal mRNA molecules, while the remaining three pathways degrade aberrant mRNA transcripts. The effect of RNA binding proteins (RBPs) on mRNA decay is also touched upon.

The importance of mRNA degradation as a mechanism of post-transcriptional regulation has been investigated on a genome-wide scale in a handful of studies. Studies on decay rates for mRNA molecules are usually carried out by treating cells with a transcriptional inhibitor followed by harvesting of mRNA at a set of time points, and hybridization onto microarrays or some other blotting method.

Quite a few articles have been published analyzing global decay rates in *Bacteria* and *Archae.* Studies have been performed with the antibiotic agent actinomycin D as a transcriptional inhibitor on *Plasmodium falciparum* (Shock et al., 2007), *Halobacterium salinarum* (Hundt et al., 2007), and *Sulfolobus* (Andersson et al., 2006), whereas the bacterial RNA polymerase inhibitor rifampicin has been employed in studies on *Escherichia coli* (Bernstein et al., 2002; Selinger et al., 2003), *Bacillus subtilis* (Hambraeus et al., 2003), and *Lactococcus lactis* (Redon et al., 2005).

Actinomycin D has also been used as a transcriptional inhibitor in the same type of study on human T lymphocytes (Raghavan et al., 2002), as well as on human HepG2 cells and Bud8 primary cells (Yang et al., 2003).

In the yeast *Saccharomyces cerevisiae*, the chelating agent thiolutin and the temperature-sensitive *rpb1-1* allele, with a mutation in the catalytic subunit of Polymerase II, have been used in a couple of studies (Guan et al., 2006; Wang et al., 2002). Other transcriptional inhibitors, such as 1,10-phenanthroline, as well as the two mentioned above were used in Grigull et al., 2004.

Finally, a similar study in *Arabidopsis thaliana* employed the RNA synthesis inhibitor 3'deoxyadenosine (cordycepin) (Gutiérrez et al., 2002). Decay rates in *Arabidopsis* have also been investigated with Actinomycin D (Narsai et al., 2007).

At a first glance, these different methods of transcriptional inhibition pose obvious problems. Using the temperature sensitive *rpb1-1* allele is dependent on a temperature shift, which inevitably triggers the heat shock response in yeast cells (Grigull et al., 2004). Chelating agents like phenanthroline and thiolutin also have a stress effect on the cells, in addition problems with non-specificity.

Another problem with all these approaches is that, at transcriptional inhibition, the pool of mRNA consists of both young and old transcripts (Meyer et al., 2004). The effect of this is a heterogeneity of the pool of mRNA as well as a non-stationarity in population localization.

An indirect approach to estimating mRNA decay rates, in connection to oxidative stress in yeast, has been presented (Molina-Navarro et al., 2008). With this methodology, steady state levels of mRNAs are measured, as well as transcription rates using Genomic run-on, which allows for indirect estimates of mRNA decay rates.

2.1 Agents used for transcriptional arrest

Global analysis of mRNA decay rates is performed in order to investigate the mRNA degradation processes and its effects on gene expression. One goal is to find out which types of transcripts, for example grouped in Gene Ontology categories, that have fast respectively slow decay rates under normal conditions.

One problem with the methods based on transcriptional arrest is that we cannot exclude the possibility that the transcriptional shut-down does not itself affect the decay rates of transcripts, as reported with thiolutin in yeast (Pelechano and Pérez-Ortín, 2008). Another problem is that the conditions under which decay rates usually are examined are far from normal.

Decay rates for transcripts in yeast were examined with the use of several transcriptional inhibitors, and compared accordingly in Grigull et al., 2004. The authors also compared the expression for decay rate profiles ($\log_2$-ratios ranked at the different time points) in their data set with a large set of microarray profiles from the literature. The conclusion was that the ranked mRNA decay expression patterns were most similar to expression patterns from microarray experiments on heat shock and nutrient starvation.

Since the use of the *rpb1-1* allele in mRNA decay experiments in yeast requires a temperature shift, the similarity with heat shock expression profiles is not surprising. The stress response in the cells might be post-transcriptional, but may also be transcriptional due to the fact that the shut-down of transcription is incomplete (which has been shown to occur for other transcriptional inhibitors as well, for example phenanthroline).

In many of the studies on mRNA decay, conclusions are drawn that stress response genes connected to heat shock are stable, while ribosomal transcripts are highly unstable. Under the stress response that the transcriptional inhibitor elicits, this is very likely, but not necessarily so in non-stressed conditions. Generally it is hard to make a statement, that stress genes are stable, and ribosomal genes are unstable under these circumstances. What might be of interest, is that the stress genes become stabilized during a stress response, as described in paper II.

2.2 mRNA Decay Models

Several models for the process of mRNA decay are plausible. Most mRNA decay analysis experiments are analyzed using a simple exponential decay model.

Assume that we are observing a single mRNA species, with $N(0)$ copies in normal conditions. We are interested in observing the change of the number of copies over time, denoting this with $N(t)$. A simple exponential decay model is

$$N(t) = N(0)e^{-\lambda t}$$

where λ is a decay constant. The half-life, $t_{1/2}$, of an mRNA transcript is often referred to in the literature. The meaning of the half-life is the amount of time that passes until the amount of a transcripts has dropped to half of its initial value. For our model we have $t_{1/2} = \ln(2)/\lambda$ and substituting the half-life for the decay constant in the expression above will produce

$$N(t) = N(0)2^{-t/t_{1/2}} \, .$$

Ideally, in a decay experiment of a competitive hybridization fashion (and indirectly in one-channel experiments), the quantity wanted is $N(t)/N(0)$, giving

$$\frac{N(t)}{N(0)} = 2^{-t/t_{1/2}} \, .$$

In a microarray experiment, transformations on a $\log_2$-scale are often used, and transforming the above quantity would give

$$\log_2 \left(\frac{N(t)}{N(0)} \right) = -\frac{t}{t_{1/2}} \, . \tag{2.1}$$

2.3 Using Microarrays to Measure mRNA Decay

The microarray technology has been used in all of the studies mentioned earlier to investigate decay rates on a genome-wide scale. Different microarray technologies were used, for example Affymetrix microarrays, two-color cDNA arrays, and nylon filter arrays.

2.3.1 Experimental setups

A common experimental setup for an mRNA decay experiment using cDNA arrays is illustrated in Figure 2.1.

Cell samples are harvested before, and at a set of time points, here denoted $t_1, \ldots, t_4$, after the addition of a transcriptional inhibitor, alternatively after a temperature shift when the *rpb1-1* allele is used. Ideally, this setup allows for direct comparisons of decay slope coefficients between different mRNAs, and also between two time series experiments under different conditions. As an exception, in Wang et al., 2002, genomic DNA was used as a reference with filter arrays.

mRNA degradation experiments using single-channel techniques like Affymetrix are naturally different, with no competitive hybridization scheme. Instead, cells

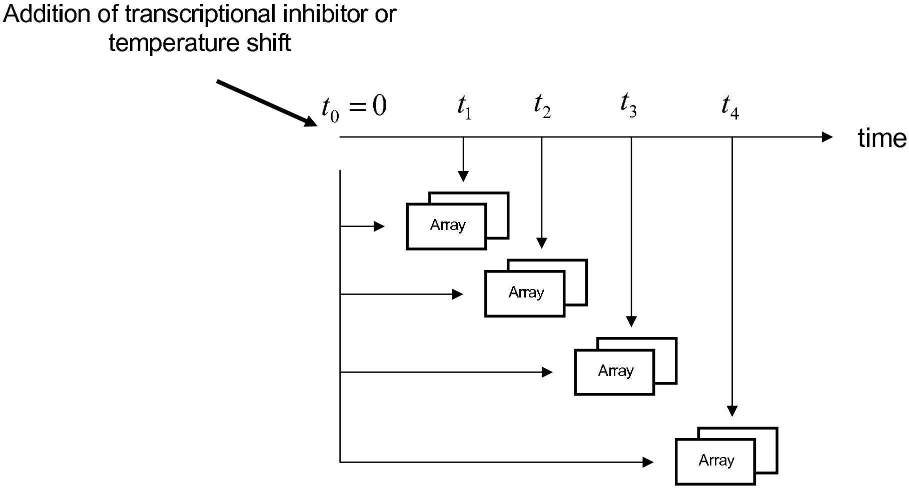


Figure 2.1: Experimental setup

harvested before and after transcriptional inhibition are separately hybridized onto arrays, e.g., in Guan et al., 2006; Raghavan et al., 2002.

In an ideal setup with measurements of the expression before and after the addition of a transcriptional inhibitor at a set of time points, our measurements of the quantity given in Equation 2.1 would produce decay profiles, from which we could quite easily extract a median half-life. Unfortunately, a lot of noise is added to each of the time point measurements, and the normalization methods and/or hybridization scheme lead to a shift of the expression values in each time point.

The situation is not hopeless, however. By hybridizing replicate arrays, reproducible signals will be amplified despite the presence of noise. In Paper II, we also chose to rely on the strength of such multi-parallel time series comparisons. When using this approach, the global behavior over each time series is assumed to be unchanged (which is not an overall unreasonable assumption). To detect transcripts with changing stability profiles, the profiles were compared over the time series to produce a change in stability index. By performing this analysis, systematic errors within each series are likely to cancel out. By using parallel series the number of experiments are increased, giving more degrees of freedom and hence power in the statistical procedures.

2.3.2 Hybridization schemes

With two-color microarrays, cDNA from two sources is labelled and co-hybridized to an array. The same amount of RNA should be hybridized from the two sources onto the array. In order to achieve this, one measures the amount of total RNA, of which rRNA constitutes the major part, and hybridizes the same amount from the two sources. Another method is to measure the amount of mRNA and hybridize equal amounts from the two sources onto the array.

In mRNA decay investigations, the first method of hybridizing the same amount of total RNA onto the arrays, really means that substantially different amounts of mRNA are used in the two pools. If the rRNA remains relatively stable, the RNA pool from the cells treated with a transcriptional inhibitor will contain less mRNA than the reference pool.

The problem with this approach is that two-color mRNAs are not made for comparisons of this type. All normalization methods, designed for example to remove intensity dependent effects, are based on the prerequisite that there is no dependence between $\log_2$-ratios of the two channels, to the mean intensity in both channels, i.e., that an MA-plot has a cloud centered around zero.

In an experiment measuring mRNA decay, performing these types of normalization procedures, will remove trends between mRNA length and decay rate, if such trends exist. On the other hand, not performing any normalization might mean that technical artifacts are mistaken for true signals. All this leads up to the fact that mRNA decay rates, for example half-lives, cannot easily be computed from the experimental data. Substantial data transformation and calculations have to be performed in order to extract the interesting information, and even then it becomes highly dependent on the methods used.

2.3.3 Normalization

As mentioned above, the standard normalization methods might not be appropriate for mRNA degradation data. Dependencies, which of course are hard to investigate, may exist that disappears upon data preprocessing.

Another aspect is that arrays from different time points (e.g., 5 and 30 minutes) have very different orders of magnitude for $\log_2$-ratios (M-values). Performing global scale normalization will then remove this information, and give erroneous estimates of decay rates. Using Affymetrix arrays and normalizing with for example RMA will give the same problem.

2.3.4 Data transformation

To extract approximate half-lives for individual mRNAs from the type of microarray experiments described here, some kind of transformation of the data is required. Several methods have been used in the studies published in the literature.

Spike-in controls can be used in both one-color and two-color microarray experiments. For a two-color array, a set of control sequences (e.g., clones) are printed onto the array. The control sequences are commonly rDNA or gene fragments from a different organism than the array is designed for. Labeled control genes (sequences meant to hybridize to the control sequences) are also added, in equal known concentrations, to the biological sample pools. In one-color experiments, control oligos are added to the array, and known concentrations of one or several control genes are hybridized to the array with the biological sample.

Spike-ins are quite common in the standard microarray experiments, for example to evaluate different normalization methods, or to use directly as a means for normalization. In experiments designed to measure mRNA decay, spike-ins can be used to estimate a common base-line for the arrays in the time series measuring the expression after the addition of the transcriptional inhibitor. The base-line is used to shift the M-values for time points to achieve a decay profile.

The spike-in data transformation method was used in for example *E. coli* with 64 rDNA spots to compute the base-line (here referred to as the normalization constant) (Bernstein et al., 2002) and in *H. salinarium* with a single *E. coli* gene fragment as a control spot.

As a general rule, the more spiked-in control spots that are used, the better the information concerning the base-line expression. If the spike-ins also are to be used as basis for normalization of the array, for example to remove intensity dependent trends in the data, it is important that the spike-ins have a good spread over the intensity range. However, spike-in data cannot solve the inherent problem of pooling of genomic material from several cells that is needed in the microarray experiments. The microarray data is interpreted on a per cell basis, although several cells were pooled to achieve the effect. The fact that cells are pooled to collect mRNA also induces a heterogeneity since the sampled cells can be in different phases of the cell cycle.

To achieve a decay profile, it is also possible to use computational methods which can be applied under certain assumptions concerning the decay profiles for a subset of genes. In Andersson et al., 2006, a given proportion (10%) of the transcripts were assumed to be stable. The individual decay profile slopes were normalized by subtraction of the average profile slope in the stable group. A similar computational method was applied in Guan et al., 2006, but the set of stable transcripts was instead identified from an external data set and additional information from Northern Blot data was used.

Instead of assuming that a group of transcripts are stable, a mean half-life can be assumed for all the transcripts. In yeast, this method was applied by a reweighting of the decay profile slopes to a mean half-life of 23 minutes (Grigull et al., 2004). The mean half-life value was taken from (Wang et al., 2002), but since the setup of the experiments were somewhat different, this assumption might not be valid.

Finally, a transformation of the data can be achieved by choosing a base-line given by an internal gene (Hambraeus et al., 2003; Raghavan et al., 2002). For this method to work, the base-line gene has to be stable over the time course experiment, which is of course always an approximation.

Chapter 3

Evolutionary Conservation Studies

Studies of evolutionary conservation are important tools in genomics. Here, terminology and methods are introduced that are utilized in this field.

3.1 Homology, Orthology, and Paralogy

In its most general definition, homology describes common evolutionary descent between entities. Concerning genes, two genes are called homologous if they derive from a single ancestral gene. Orthologs are genes in different species that derive from a single common ancestral gene, i.e. resulting from splitting of lineage via a speciation event (also called vertical descent). Paralogs, on the other hand, originate from gene duplication events within a genome (Webber and Ponting, 2004; Koonin, 2005; Sonnhammer and Koonin, 2002; Jensen, 2001).

Orthologs are *not* simply genes with gene products that have the same catalytic function in different species and neither are paralogs simply homologs within an organism (Sonnhammer and Koonin, 2002; Jensen, 2001). In fact, paralogs are also defined between species, and can be divided into in-paralogs and out-paralogs (Sonnhammer and Koonin, 2002; Koonin, 2005). Out-paralogs are genes resulting from duplication events *preceding* a speciation event. In-paralogs, on the other hand, are paralogous genes that are formed by a lineage specific duplication event, i.e. a duplication event *subsequent* to the last speciation event.

In Figure 3.1 a quite complex situation with paralogs and orthologs in human, worm, and yeast is depicted (Sonnhammer and Koonin, 2002). After the speciation (lineage-split) of animals and fungi, a duplication event has taken place, giving two forms A and B of a gene. After the lineage split of worms and humans, additional duplications in the A form took place in both the human and worm lineages. From the figure we can deduce the following orthologous and paralogous relationships:

i) The yeast gene is homologous to all genes in both the worm and human lineages, and these genes are called co-orthologous to the yeast gene.

ii) Comparing the human and worm genes, the H_A genes are co-orthologous to the W_A set of genes.

iii) The H_A genes are in-paralogs when comparing to the worm lineage.

iv) The H_B gene and the H_A genes are out-paralogs when comparing human to worm.

v) The H_A set and H_B, and the W_A set and W_B, are all in-paralogs in comparing with yeast.

The division of paralogs into subgroups may be useful to exactly describe evolutionary events and connections but it also becomes quite complex to understand. In Figure 3.2, a simpler case is depicted. Orthologs of the α-version and β-versions of the globin gene are present in all three species of interest (frog, chicken, and mouse). The paralogs in the figure (mouse lineage only indicated) are all out-paralogs.

Although speciation and duplication can be thought of as the primary forces of evolution (Koonin, 2005), several other events complicate the picture of genome evolution. Horizontal gene transfer is a phenomenon where genes have been transferred between species. A gene and its transferred version in another species are called xenologs. In addition, gene loss and gene fusion or fission events complicate the process of identifying orthologs and paralogs in gene families.

Sequence similarity is not necessarily evidence of orthologous relationships between genes, but is in many cases a good indicator. The same is true for structural similarity - similarities in three-dimensional structure of proteins, including binding sites - may as well be a good indicator. The problem with

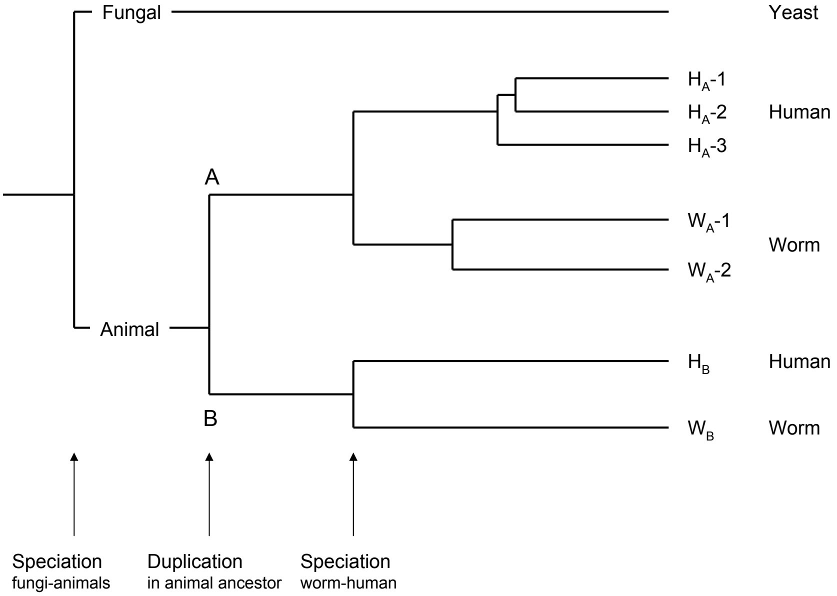


Figure 3.1: Concepts of homology (adapted from Sonnhammer and Koonin, 2002).

structural similarity is that relatively few proteins have known 3D-structures and computational prediction is difficult. Working with two-dimensional structure components like loops and sheets might also be useful.

Purpose of Orthology Prediction

Although homologous relationships are evolutionary connections and not functional connections, a highly used method for functional prediction is homology (Gabaldón and Huynen, 2004). The reason for this is naturally that orthologous genes (sharing a single common ancestor) are likely to have similar function.

The Ensembl database (Hubbard et al., 2009) utilizes a pipeline to assign both orthologous and paralogous relationships as an aid in their automated annotation procedure. As an example, if an orthologous group of genes is found between three species, and one of these genes has functional annotation information, this function can be inferred to the genes in the other species in the orthologous group as well. Inferred functional annotation is also common between

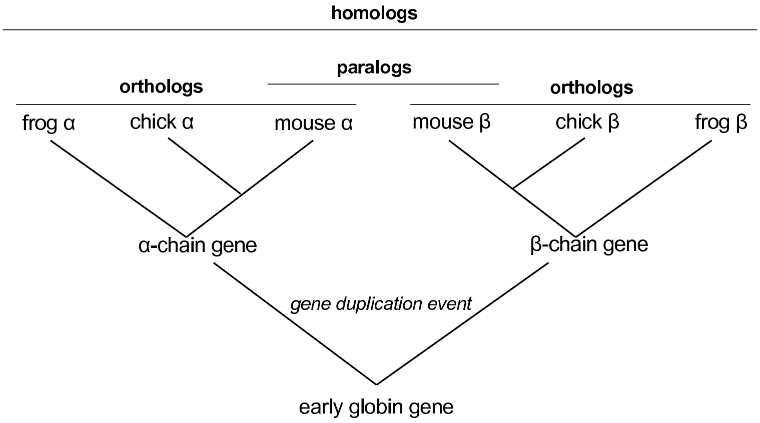


Figure 3.2: Concepts of homology, simpler case.

yeast species, where, for example, functional annotation for the fission yeast *S. pombe* is often inferred from the more studied budding yeast *S. cerevisiae.* Although the evolutionary distance between the two yeasts is large, orthologs between the two species are present and possible to identify. As of July 2010, 2163 of the 5036 protein coding genes for *S. pombe* in GeneDB (Hertz-Fowler et al., 2004) have a functional role inferred from homology (compared to 1873 protein coding genes with annotation status experimentally characterized or published).

After a speciation event, the different orthologous proteins are independently subject to mutation events. The mutation events must be weighted against functional constraints, i.e., the function of the protein must be retained. In a study on evolutionary rates for orthologous proteins in the three domains of life (Jordan et al., 2001), a small percentage of the genes exhibit accelerated evolution. Classification of orthologs into groups, together with rate of evolution studies may thus be used to indicate adaptive diversification.

In the comparative genomics field, orthology is essential since it allows for comparison of genomes in terms of gene content (Gabaldón and Huynen, 2004). In general, comparisons of genomes over orthologous groups is an important tool in evolutionary biology, for example in the tree of life project (Delsuc et al., 2005).

3.2 Ortholog Prediction

A large set of different ortholog prediction methods and procedures have been published. A comparison and categorization of several methods into phylogeny-based methods, BLAST-based methods, and methods based on evolutionary distance measures is presented in (Chen et al., 2007).

In predicting orthologs, the sequence information used is of vital importance. Predicting orthologs for non-fully sequenced organisms will produce incomplete clusters with possibly missing orthologs and paralogs. Even if the genome of an organism is fully sequenced, the genome information can be of varying quality, which will affect the ortholog prediction.

The information usually needed are lists of proteins or nucleotide sequences from two or several genomes in an easy format. If the species used in the orthology prediction procedure are distantly related, protein sequences are used.

3.2.1 Phylogeny-based Methods

RIO (Resampled Inference of Orthologs) (Zmasek and Eddy, 2002) and OrthoStrapper (Storm and Sonnhammer, 2002) are methods using phylogenetic trees to infer groups of orthologous proteins or genes within a family. To build phylogenetic trees, high quality multiple sequence alignments are usually required. RIO uses for example alignments and profile HMMs from the Pfam database.

3.2.2 BLAST-based Methods

RBH (Reciprocal Best BLAST Hit) is a common first step in the BLAST based ortholog identification procedure, as well as a part in the RSD method mentioned below. BLAST is a commonly used algorithm for local alignments of protein and DNA sequences (see Section 1.2.2). A reciprocal best hit between two sequences (e.g., genes) in two genomes means that a BLAST search for the first sequence in the second genome produces the second sequence as the best hit and vice versa.

The Inparanoid program (Remm et al., 2001) uses a reciprocal best hit strategy as a first step to identify putative ortholog pairs, which in a second cluster step is used to add-on additional orthologs and paralogs. The cluster step is rule based, and assumes that all orthologs have an equal evolutionary rate.

COG (Clusters of Ortholog Groups), see for example (Tatusov et al., 2000) is a very well known database of ortholog groups. The procedure used to identify the COGs starts with an all-against-all BLAST search followed by identification of genome-specific best hits (BeTs). The BeTs are triangles of mutually consistent genome-specific best hits where related BeTs are merged to produce COGs. The COGs are manually curated in different ways to exclude false-positives and find groups with multidomain proteins.

The OrthoMCL algorithm (Li et al., 2003) is also based on an all-against-all BLAST search as a first step, in which putative ortholog pairs as well as recent paralogs are identified. The connections between sequences, both within and between genomes, are represented as a graph and clustered. The Markov clustering step is based on the MCL algorithm (Enright et al., 2002).

3.2.3 Methods based on Evolutionary Distance

The RSD (reciprocal smallest distance) algorithm was developed to improve upon the reciprocal best BLAST hit procedure to identify orthologs (Wall et al., 2003). The improvement consists of an additional step of pairwise global sequence alignments and estimation of evolutionary distances between putative (reciprocal) ortholog pairs using maximum likelihood.

3.2.4 Synteny information

Homologene is a database hosted by NCBI which contains homology information for a large set of sequenced genomes. The detection procedure is partly based on BLAST searches but information concerning synteny, i.e. co-localization of genes (or other chromosomal regions) on chromosomes is also utilized.

Chapter 4

Regression Models and Shrinkage Methods

4.1 Regression Models

The aim of the univariate linear regression model is to explain the variability in a response variable $\mathbf{y}$ with a set of predictors $\mathbf{x}_1, \dots, \mathbf{x}_p$. The response variable has n data points; $\mathbf{y} = (y_1, \dots, y_n)^T$, and n is hence also the dimension of each predictor $\mathbf{x}_j$. The model explains the responses with a linear combination of the predictor variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

for $i = 1, \dots, n$, and ε_i denoting the error random variables, often assumed to be Gaussian. In matrix form, with $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$, the equations can be formulated as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

When minimizing the squared deviations of the observations from the model we get the least-squares solution

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

under the restriction that $\mathbf{X}^T\mathbf{X}$ is non-singular, and the solution in matrix form is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (Hastie et al., 2009). We observe that the number of replicates n needs to exceed the number of predictors p in the model for a unique least-squares solution to exist. In fact, the closed form solution to the ridge regression problem (see below) was first introduced to alleviate the problem of a singular $\mathbf{X}^T\mathbf{X}$ matrix.

In regression modeling, situations can occur when we want to, or have to, improve upon the least-squares solution to the regression problem. Two such reasons are prediction and interpretation (Hastie et al., 2009). The least-squares estimates have low bias but large variance. The prediction can sometimes be improved by accepting a larger bias for a gain in the form of lower variance. Interpreting models with a large set of predictors can be hard, and it is sometimes preferable to select a smaller subset of variables that show the strongest effects. Also, in many applications, for example biology with genome-wide association studies, the number of predictors p can be much larger than the number of observations n. Since the full model is not estimable in this setting, some kind of regularization or selection on the predictor matrix is necessary to find influential predictors.

4.2 Selection Methods

To resolve the issue with $p \geqslant n$ or improving the interpretability of the model, some kind of subset selection on the predictors can be done.

Best-subset selection, involves for each size $k = 1, \ldots, p$, to find the best subset of predictors with size k (e.g., the set that minimizes the quadratic deviations). As p increases the problem becomes more difficult to handle, since the number of subsets increases rapidly (Hastie et al., 2009). An alternative to searching through all subsets, is to search for a path of predictors by sequentially adding the predictors one by one. Such a method is forward stepwise selection (Hastie et al., 2009) in which a predictor is added if it is the one that improves most upon the fit. A related method is the backward stepwise regression in which predictors instead are eliminated from the full model. Based on this description, it is easy to conclude that backward stepwise regression is only possible if $n > p$.

4.3 Shrinkage Methods

An alternative to subset-selection methods is to use shrinkage methods. In a regression model, constraints can be imposed on the sizes of the regression coefficients, which induces shrinkage.

4.3.1 Nonnegative Garrote and Ridge Regression

The nonnegative garrote was introduced as a better alternative to subset selection (Breiman, 1995). The least-squares estimates (OLS estimates) of the regression coefficients are shrunk by nonnegative factors, under the condition that the sum of these factors are constrained.

$$\hat{\beta}_{garrote} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} c_j \hat{\beta}_j^{OLS})^2$$

$$\text{subject to } c_j \geq 0, \sum_{j=1}^{p} c_j \leq t$$

The nonnegative garrote cannot be used in the $p > n$ setting since it requires the least-squares estimates of the regression coefficients. If some of the predictors are highly correlated, the OLS estimates of the coefficients (if they can be found) may perform badly, and has an equally bad influence on the garrote (Tibshirani, 1996).

A related shrinkage method is ridge regression, in which the sizes of the squared regression coefficients are constrained. The resulting constrained optimization problem can be written

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

which can be rewritten in Lagrangian form as (see below for an explanation of the connection between t and λ)

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \Big\}.$$

The coefficients in the ridge regression are shrunk towards zero (and consequently towards each other). The ridge regression has an explicit solution given by $\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ (Hastie et al., 2009). The ridge regression is not scale invariant, and usually, the predictors are scaled to have mean 0 and variance 1.

4.3.2 The Lasso

By replacing the $\mathcal{L}_2$ constraint on the regression coefficients in the ridge regression by an $\mathcal{L}_1$ penalty constraining the absolute values of the coefficients, $\sum_{j=1}^{p} |\beta_j| \leq t$, we get the lasso (Tibshirani, 1996):

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}$$

The constraint on the parameters, t, can be mapped one-to-one to λ in the Lagrangian form of the lasso problem. Provided that $\hat{\boldsymbol{\beta}}(\lambda)$ is a solution to the Lagrangian form, then it also solves the constrained version of the lasso with $t = \sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|$ (Friedman et al., 2007).

In contrast to ridge regression, the lasso penalty will often force some coefficients to be set to zero. Figure 4.1 (adapted from (Tibshirani, 1996)) illustrates why this occurs in the two-predictor case. The ridge constraint region $\beta_1^2 + \beta_2^2 \leq t$ is spherical, while the lasso constraint region $|\beta_1| + |\beta_2| \leq t$ is rhomboid. The elliptical contours of the least-squares solution will more often hit a point where some coefficients are exactly zero with the lasso constraint.

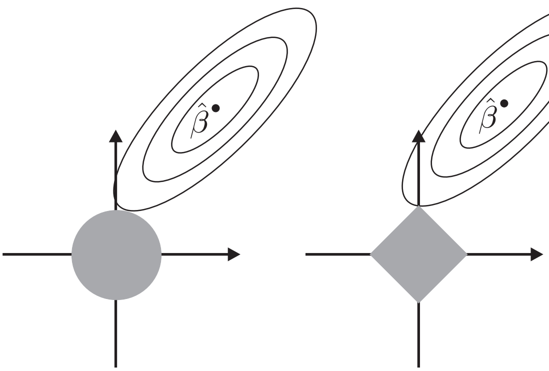


Figure 4.1: Ridge constraint region (left) and lasso constraint region (right) with the elliptical contours of the least-squares solution with two predictors.

The lasso (unless the predictors are orthonormal) does not have a closed form solution, as opposed to ridge regression, but it is a convex optimization problem, and hence has a unique minimum. The Least-Angle Regression (LARS) algorithm was introduced as a new model selection algorithm (Efron et al., 2004) and is related to the forward stepwise regression in which variables are entered sequentially. LARS, in contrast to forward stepwise, does not necessarily enter the full variable in the fit, but just enough until another variable has the same correlation with the current residual. A small modification to the LARS algorithm produces lasso solutions to a regression problem and is based on the fact that the lasso paths are piecewise linear.

The lasso problem is usually approached by estimating the coefficients as the penalty factor λ is varied across a relevant range, producing what we refer to as the lasso paths. The paths are the trajectories of the regression coefficients as the penalty parameter λ changes, and an illustration is given in Figure 4.2. The penalty parameter λ is large for small values of the $\mathcal{L}_1$ norm of the coefficient vector and decreases along the x-axis until it is zero, and we end up with the OLS estimates of the coefficients (since we here have the $p < n$ situation).

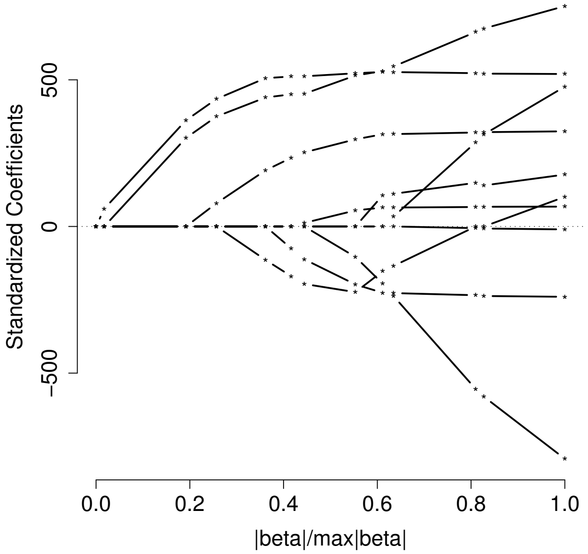


Figure 4.2: Lasso paths for the regression coefficients in the diabetes data (Efron et al., 2004). The data set includes a response vector and ten predictor variables for 442 patients. The x-axis shows the $\mathcal{L}_1$ norm of the coefficient vector, as a fraction of the maximal $\mathcal{L}_1$ norm: $\sum_{j=1}^{p} |\beta_j| / \max\{\sum_{j=1}^{p} |\beta_j|\}$.

The solution paths can be easily produced for relevant choices of lambda via the LARS algorithm or by coordinate descent (Friedman et al., 2010). The problem is then to choose a suitable value of λ to define the final model. Small values of lambda means little regularization and hence a complex model, while large values indicate a heavy penalty and a sparse model.

Model selection criteria like Mallow's C_p, the Bayesian Information Criterion (BIC) or Akaike's Information Criterion (AIC) can be used to select λ in the lasso under certain circumstances. The AIC and BIC select models based on the log-likelihood under a trade-off with the number of parameters (the degrees of freedom) needed to achieve the likelihood. For a fitted model with responses $\mathbf{y}$ and fitted values $\hat{\boldsymbol{\mu}}$ we define

$$\begin{aligned}\mathrm{BIC}(\hat{\boldsymbol{\mu}}) &= -2\cdot\mathrm{loglik} + \log(n)\, df(\hat{\boldsymbol{\mu}})\\ \mathrm{AIC}(\hat{\boldsymbol{\mu}}) &= -2\cdot\mathrm{loglik} + 2\, df(\hat{\boldsymbol{\mu}})\end{aligned}$$

and choose the model with the minimum BIC or AIC. In Gaussian models, the criteria can be rewritten to illustrate the dependence on the residual sum of squares and the sample variance, where the log-likelihood term (and the constant) can be shown to be $\|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2/\sigma^2$, and the resemblance to the C_p criterion is clear.

$$C_p(\hat{\boldsymbol{\mu}}) = \|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2 + 2\, df(\hat{\boldsymbol{\mu}})\,\sigma^2$$

Choosing one of the model selection criteria is not trivial. The BIC has attractive properties like consistency when the sample size $n \to \infty$, which is not the case for AIC, as its choices tend to be too complex for large n. However, BIC has a tendency to choose too sparse models for smaller n (Hastie et al., 2009).

In the lasso context, using any of the criteria is a bit more complicated, especially when it comes to estimating the degrees of freedom of the fitted model, since the lasso is a technique that in a way searches through the set of predictors. An unbiased estimate of the degrees of freedom of a lasso fitted model with parameter λ is the number of predictors included in the model for that particular λ (Zou et al., 2007).

Another problem that arises is in the $p \geqslant n$ context when the lasso selects as many variables as possible (giving a saturated model), and then terminates. Naturally the models having as many parameters as (or close to) the number of observations n will have very small residual sum of squares, and hence a very big likelihood. The likelihood will dominate in the BIC or AIC criterion and hence the most complex model will almost always be chosen. The C_p is difficult to apply in the $p > n$ context as well, since it includes an estimate of the model variance. In the BIC with the lasso this usually means the variance for the OLS model, but if $p > n$ this is not available.

An alternative which seems to work well for the lasso in general is cross-validation. By dividing up the data into K bins, and in turn excluding the kth bin, adapting the models, and estimating the predictive power on the excluded bin, an average prediction error for a sequence of λ's can be estimated.

The λ showing the smallest prediction error is chosen. One drawback with the cross validation technique is that it can be unstable, especially if the sample size is small.

4.3.3 Other Shrinkage Methods

Since the publication of the original lasso paper, and especially the LARS paper, many more papers have been published that explore, apply, or develop shrinkage and lasso type methods.

The elastic net (Zou and Hastie, 2005) is a compromise between ridge regression and the lasso, in which the penalty on the regression coefficients is a linear combination of the ridge and lasso penalties: $\sum_{i=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \leq t$. The penalty is designed to shrink variables in a ridge-like fashion, while selecting variables according to the lasso. The elastic net has some attractive properties, like the ability to include more predictors than the number of replicates, unlike the lasso. If there are highly correlated predictors present, the lasso tends to pick one of them (and ignore the others), while the elastic net allows for highly correlated predictors to be included in the model together by shrinking their coefficients towards each other. One additional issue with the elastic net is the need to choose the parameter α determining the trade-off between the ridge and the lasso penalties (Zou and Hastie, 2005; Hastie et al., 2009).

The grouped lasso (Yuan and Lin, 2006) was introduced to address the problem of adding predictors in a grouped fashion in the lasso. The predictors are gathered into predefined groups, and if such a group is chosen to be in the model, all of its members will be included. This type of variable selection is imposed by a penalty combining a lasso type $\mathcal{L}_1$ penalty between groups, and ridge type $\mathcal{L}_2$ penalty within groups. In order to compensate for different group sizes, a scaling constant can be included in the penalty. It turns out that the solution paths in the grouped lasso are not piecewise linear, so a LARS type algorithm cannot be used. Instead algorithms based on block coordinate descent can be used to solve the group lasso (and other more complicated regression models with lasso type penalties) (Friedman et al., 2007; Friedman et al., 2010).

The Composite Absolute Penalties (CAP) family of penalties is a generalization of the group lasso penalty, in which the between groups and within groups penalties can be used with different norms. One important application they present is hierarchical selection of variables, which may be useful in regression models with interaction terms. Hierarchical selection of variables can be useful when it is reasonable to include an interaction term only if both of the main effects also are included (Zhao et al., 2009).

In the graphical lasso (Friedman et al., 2008) estimation of sparse graphs is done via a lasso penalty on the inverse covariance matrix in a Gaussian model. The reasoning behind this is that if the ij^{th} element of the inverse covariance matrix is zero, the variables i and j are conditionally independent.

4.3.4 Dantzig

The Dantzig selector is somewhat different in flavor to the lasso based methods. The constrained optimization problem solved by the Dantzig is

$$\begin{aligned} &\min_{\beta} \|\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t \end{aligned}$$

in which the maximum inner product of the current residual with the predictors is minimized (Candes and Tao, 2007). The constraint region is identical to the constraint region in the lasso, but the objective function is not ellipsoid but box-shaped (Meinshausen et al., 2007), which in certain circumstances results in multiple solutions to the minimization problem. The Dantzig selector has some appealing theoretical properties but has been shown to sometimes produce erratic coefficient paths and suffer from poor predictive accuracy (Hastie et al., 2009).

4.4 Model Selection via Rate-Distortion

In the information theory field, rate-distortion theory builds the foundation for data compression, and has many uses in for example video and image encoding. Generally, the aim in applying rate-distortion in image compression is to represent a source with as few bits as possible (Ortega and Ramchandran, 1998). In the compression of an image, the total number of bits R_T (i.e., the coding rate) is restricted. For efficient compression, it makes sense to code separate parts (units) of the image with different number of bits, for instance in allocating few bits to background parts of the image. The distortion of the image, measured with for example mean squared error, naturally decreases with increasing number of bits allowed in the compression for the different units.

Suppose we have N units with M different operating points (different ways to allocate bits). For each unit i and operating point j, we have a rate r_{ij}

and a distortion d_{ij}. The optimization problem that is to be solved consists of minimizing the overall distortion $D = \sum_{i=1}^{N} d_{ij}$ under the constraint that the total number of bits allocated to different units must not exceed $R_T = \sum_{i=1}^{N} r_{ij}$. We wish to find the operating points of each unit that satisfies the optimization problem.

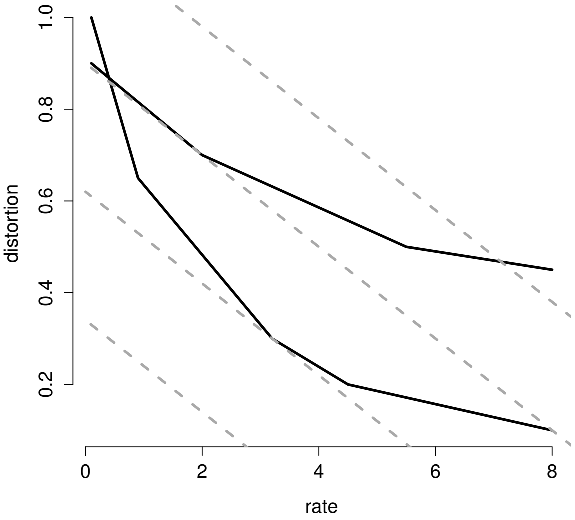


Figure 4.3: Rate distortion curves for two units. The dashed lines correspond to a fixed slope constraint.

Figure 4.3 illustrates the rate-distortion curves for two units with the dashed lines corresponding to a fixed slope constraint Δ. For a fixed Δ, it can be shown that the optimization problem is satisfied by selecting the points for each unit that is first "hit" by the slope constraint wave with the total rate $R_T = \sum_{i=1}^{N} r_{ij(\Delta)}$ (Ortega and Ramchandran, 1998). By closer examination, the resemblance to the optimization problem solved by the lasso is apparent, and we can also formulate the rate-distortion problem in a Lagrangian form.

The rate-distortion theory can be adjusted to work as a model selection criterion in significance testing or cluster analysis on high-dimensional data like gene expression (Jörnsten, 2009). Suppose we have N genes that act as responses, each observed in a replicated microarray experiment. For each gene we have a set of predictors from which we would like to select a subset of features that influences the gene expression the most. The predictors can for example be binding site affinities for transcription factors. Employing the lasso or some other shrinkage method of choice, we produce a set of different models for each gene by varying the penalty parameter λ. For each λ we can calculate the residual sum-of-squares (SS_R) for the model for each of the genes.

In the rate-distortion setting, we let the genes be units, the distortions be the SS_R for the genes indexed by lambda, and the rate for each gene be model complexity (proportional to the inverse of lambda). Simultaneous model selection for all the genes can now be performed by choosing a slope constraint Δ. In effect, we wish to minimize the overall residual sum-of-squares under the restriction that the number of parameters allowed is limited. The slope constraint can be chosen by for example cross-validation.

Chapter 5

Summary of Papers

Paper I: Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments

This paper deals with identifying orthologs for human drug targets in a set of species in order to assess if human drugs potentially can affect these species by means of a drug-drug target interaction. The paper proposes the use of more targeted test strategies in ecotoxicity testing based on predicted orthology information in 16 different species.

Unintentional exposure of pharmaceutics is a risk for a wide range of species in their natural habitats, although the pharmaceutical residues appear in quite low concentrations. In the approval of new pharmaceutical products, environmental risk assessments are compulsory. However, the testing procedures required are mainly based on Daphnia (waterflea) and algae, and not on higher organisms, like fish or amphibians, in a specific way.

The conservation of drug targets in species relevant for ecotoxicity testing is a key aspect in developing such targeted test strategies. Since complete genome information, with predicted proteomes, is available for a wide range of species, it is possible to extract and compile data on evolutionary conservation for drug targets. One method to do this is to predict orthologs for the human drug targets in a set of species of interest.

With the OrthoMCL algorithm (see Section 3.2.2 and Li et al., 2003), orthologs were predicted in 16 species. The parameters of the prediction method were chosen with respect to control cases with known orthologs.

The drug targets were classified into five quite large groups based on the Gene Ontology (GO). Since the GO classifies genes and gene products into non-exclusive groups, some drug targets were annotated into several groups, see Figure 5.1. Over-representation and under-representation for the GO-groups were performed in the ortholog groups in all the species, using one-sided versions of Fisher's exact test. The general trend was that proteins annotated as receptors were statistically under-represented, while proteins annotated as enzymes were over-represented.

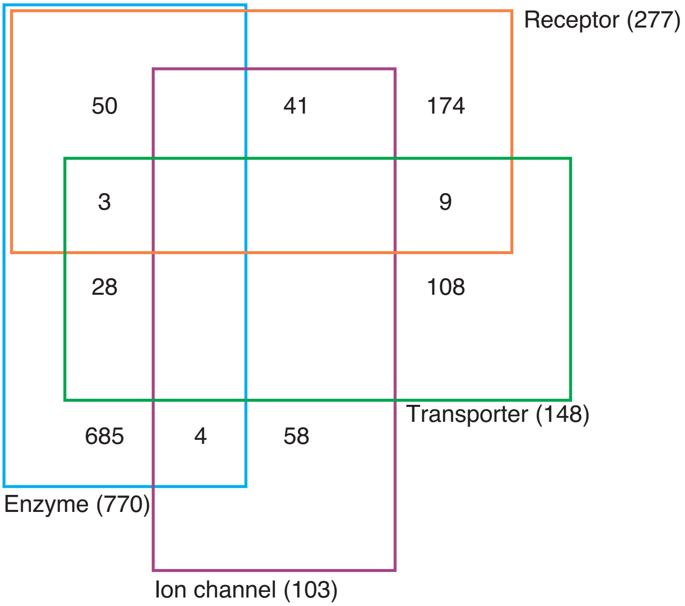


Figure 5.1: Venn diagram over the drug target inclusions in the different GO groups defined in Paper I. The 'other' category is not included.

Supporting information in the literature is presented in the paper, indicating reliability in the ortholog prediction procedure. The presented orthology predictions can be used as a guide to prioritize test species for a certain drug, interpret the relevance of existing ecotoxicity data or to deduce which pharmaceuticals may pose an increased risk to a certain organism group.

Genome annotation is, for the most part, performed automatically, giving predictions of genes, proteins, and transcripts. Errors may arise during the automated annotation process, such as failure to recognize proper start sites for open reading frames or problems in the identification of intron-exon boundaries. Also, the sequence coverage might not be adequate in the sequencing process, resulting in errors in the annotation. Searching for proteins with similarity in other species, might therefore fail because a putative ortholog has been wrongly annotated as a pseudogene, or missed all together in the sequencing and annotation pipeline. The genome sequence reliability for the different species used in the analysis is varying. Some species have quite stable genome sequences, while the genomes for other species were quite recently determined. The differing quality of the sequence data must be considered in interpreting the data from this study.

Contributions to Paper I

The current author contributed to the work by performing the data preprocessing, including proteome data processing and drug target annotations, the ortholog prediction procedures, the GO annotations, and the statistical analysis. The current author prepared the artwork, the supplementary material, and contributed to the writing and discussion in the paper together with the other authors.

Paper II: mRNA Stability Changes Precede Changes in Steady-State mRNA Amounts During Hyperosmotic Stress

This paper concerns the hyperosmotic stress response in *S. cerevisiae* and its dependence upon the stabilization and destabilization of certain transcripts.

During the different phases of the stress response; shock, adaptation, and recovery, cells need to optimize the amount and activity of large numbers of gene products. The regulation of transcription rates must be harmonized with turnover and translation efficiencies of mRNAs.

Using microarrays with the transcriptional inhibitor 1,10-phenanthroline, we analyzed mRNA turnover (decay) changes in relation to mRNA steady-state levels following hyperosmotic shock. The changes in decay rates between stress conditions were modeled with differences in decay slope coefficients (in the paper called stability indices, and denoted k_S), as described in Section 2.2. Although the method does not provide absolute estimates of mRNA half-lives, valuable information can be extracted based on the slope coefficient differences, as exemplified below.

Based on the measurements on both steady state levels and decay changes, we found that regulation of mRNA stability precedes the changes in steady state levels, both regarding the early upregulation and the later downregulation of stress induced genes. The corresponding but inverse behavior was observed for the downregulated genes.

At the peak of the stress response (in Paper II estimated to be approximately 30 minutes after the addition of salt), we observed a stability decrease for the 100 most upregulated genes. Assume that the difference in stability between 6 minutes and 30 minutes after stress is estimated to Δk_S. How does this change affect the removal of transcripts? Let $N(t)^i$ indicate the number of transcripts t minutes after the peak time point if the stability index is unchanged. Let similarly $N(t)^{ii}$ indicate the number of transcripts when we have an altered stability index. Assuming an exponential decay model, the $\log_2$ fold-change of the number of transcripts after t minutes between the two scenarios is

$$\log_2 \left[\frac{N(t)^i}{N(t)^{ii}} \right] = \Delta k_S \cdot t$$

Despite the naive model, the measurements on the decay indexes are in quite good correspondence with the steady state level data presented in the paper. At the 60 minute time-point after the addition of salt, the amount of transcripts (in reference to unstressed conditions) for the 100 most upregulated mRNAs is reduced in half compared to the peak levels. The model indicates a $\log_2$ fold-change of approximately $0.038 * 30 \simeq 1.1$ (based on data from Figure 3 in the paper).

In addition to using wild-type strains under different stress conditions, two mutant strains, $hog1\Delta$ and $rck2\Delta$, which lack components of the HOG pathway, were included. The MAP kinase Hog1 was seen to affect stability as well as steady state levels of stress-responsive transcripts while the Hog1-activated kinase Rck2 mainly played a part in the regulation of steady state levels.

Using functional categorization with both the Gene Ontology (GO) and the MIPS functional catalogue, several functional mRNA groups were found to be regulated during the phases of the stress response, allowing timely tuning of their steady-state levels. The destabilization of mRNAs at the peak steady-state levels during adaptation, allows the cell to prepare for the subsequent recovery phase during which these transcripts are down-regulated. Conversely, stabilization of stress-repressed mRNAs permits their rapid accumulation in the late phase of the stress response.

Please note that there is a methodological error in the paper connected to Figure 3. Of course, a t-test is not valid for the comparison of wild-type and $hog1\Delta$ cells for the 30 minute time point, as the genes were selected for differential expression in these time points. The comparisons for the other time points however, are valid.

Contributions to Paper II

The current author contributed to this paper by performing the microarray preprocessing and analysis, as well as the modeling of mRNA stability. The comparisons with data from other publications, the GO annotations, and most of the artwork and supplementary material was also prepared by the author. The current author contributed to the writing and discussion in the paper together with the other authors.

Paper III: Subcellular Localization and Effects of DDIT3/ GADD153/CHOP

This paper deals with the molecular effects induced by the protein DDIT3, also known as GADD153, or CHOP. DDIT3 has been indicated as a regulator in stress response, where for example stress types like ER stress and starvation induce transcription of *DDIT3*. Previously, DDIT3 has mainly been reported as a nuclear protein, but this paper shows that DDIT3 also can be localized to the cytoplasm, and that the two forms regulate different sets of genes.

We analyzed cultured normal human fibroblasts and sarcoma cells carrying amplified DDIT3 or tamoxifen inducible DDIT3 expression constructs. The effects of DDIT3 were investigated via a range of methods, including expression profiling with microarrays, real-time qPCR, immunoblotting, confocal microscopy and live imaging, and a migration assay. The design of the microarray experiment was such that both analysis of cytoplasmic and nuclear DDIT3 effects could be investigated. The genes differentially expressed due to cytoplasmic DDIT3 effects, were involved in cell movement, cell death, and cellular growth and proliferation processes.

The cell migration effects were tested in a migration assay experiment in which cells of different types were allowed to grow and migrate together with wild type (control) cells. The novelty of this assay is the use of co-migration with control cells, aimed at reducing variability and proliferation giving rise to migration rate differences. By using a mathematical model and subsequently a Wilcoxon test, differences in migration rates for the different cell types could be deduced, as described in the supplementary information accompanying the paper. We plan to further investigate interesting properties of the method, like the effect of the approximation with the use of the Wilcoxon test, and the sensitivity to the different estimated rate ratios, in a future paper.

For the nuclear form of DDIT3, we found that categories like cell death and cellular growth and proliferation, as well as cell cycle were predominant among the regulated genes. Indeed, we could observe a transient cell cycle arrest and an accumulation of cells in the G1 phase of the cell cycle in cells treated with tamoxifen (which induces the nuclear translocation of DDIT3).

DDIT3 is a leucine-zipper transcription factor of the C/EBP family, but cannot bind DNA in a homodimer form. Instead it has been reported that DDIT3 functions as a negative factor (i.e., blocking transcription) when binding to DNA in a heterodimer form with other C/EBP factors. DDIT3 can also induce gene transcription when binding with other lecuine-zipper factors, and we were hoping to discover such interaction partners in this study. We attempted to predict transcription factor binding sites and identify enriched sites among our regulated genes, both with a novel and a traditional scoring method (see supplementary information to the paper). However, we were not successful, perhaps due to the fact that our set of regulated genes is quite small, but more likely due to the fact that DDIT3 may divergently bind several different factors, and induce transcription by binding several different sites.

Contributions to Paper III

The current author contributed to the paper by performing the microarray data preprocessing and analysis, as well as the functional annotation and network analysis of differentially expressed genes. Comparisons of array data with data from qPCR experiments were also performed by the author. Prediction of transcription factor binding sites and the subsequent enrichment analysis, as well as the migration assay modeling was developed and performed by the author. The artwork and supplementary material was prepared by the author, and the manuscript and discussions therein written together with the other authors.

Paper IV: Transcriptional and Metabolic Data Integration and Modelling for Pathway Identification

This paper deals with two research questions (i) how to combine transcriptional and metabolic data, and (ii) how to use these data to draw conclusions about activities of (metabolic) pathways. Pathways are sets of reactions involving genes, metabolites, and proteins, highlighting functional modules in the cell. The pathways of course partly overlap and are connected to each other, but it may still be very important to find modules of the cell metabolism that appear to be perturbed when comparing two conditions with each other (for example comparing cancer tissue with matched normal). Figure 5.2 shows a small part of the overall metabolic pathway system described and depicted in the KEGG database (Kanehisa et al., 2010).

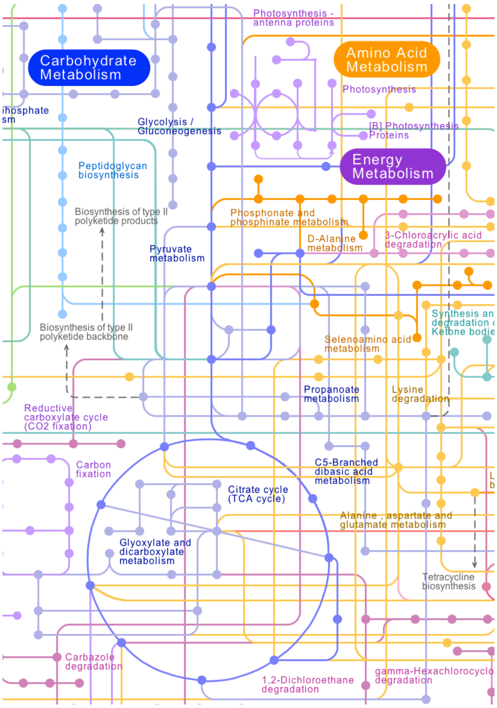

Photosynthesis - antenna proteins
Carbohydrate Metabolism
Amino Acid Metabolism
Photosynthesis
hosphate sm
Glycolysis / Gluconeogenesis
[B] Photosynthesis Proteins
Peptidoglycan biosynthesis
Energy Metabolism
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
D-Alanine metabolism
3-Chloroacrylic acid degradation
Pyruvate metabolism
Synthesis an degradation Ketone bodie
Biosynthesis of type II polyketide backbone
Selenoamino acid metabolism
Propanoate metabolism
Lysine degradation
Reductive carboxylate cycle (CO2 fixation)
Carbon fixation
C5-Branched dibasic acid metabolism
Citrate cycle (TCA cycle)
Alanine , aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Tetracycline biosynthesis
Carbazole degradation
1,2-Dichloroethane degradation
gamma-Hexachlorocyclo degradation

Figure 5.2: Part of the map over metabolic pathways provided in KEGG. Different pathways make up the overall metabolic system.

We propose a model that integrates gene and metabolite expression data in order to make pathway level based decisions. We adopt a modeling scheme in which we view the gene expression data as explanatory for the metabolite data. Modeling the data like this makes sense since metabolites can be seen as end-points of perturbations happening on the gene level (Abate-Shen and Shen, 2009).

We adapted a stepwise procedure to sequentially select pathways that the data and the model indicate are the most perturbed, or active, when comparing two conditions.

For the genes, the model and estimation procedure considers how a certain pathway explains the differences between the two conditions for its member genes. If the expression for a gene is sufficiently explained by the treatment-group difference, we estimate a coefficient for this effect, and let the gene be active, i.e. contribute to the "explanatory" power of the pathway. We choose which genes to include by rate-distortion under a Bayesian Information Criterion (BIC) (see Section 4.4).

For the metabolites in a given pathway, we allow genes to influence the expression if the gene is active and a member of the pathway under consideration. We select the genes to include as predictors for the metabolites by an elastic net penalty, and choose the model complexity via cross-validation (CV).

The purpose of using a rate-distortion selection criterion in combination with either BIC or CV, is to choose models that do not overfit to the data, so that large pathways with many genes automatically explain the metabolite data well. Even though there are many genes in a pathway, all of them may be excluded as predictors for the metabolites in the pathway if they do not significantly contribute to explain the data.

The benefits of the proposed model model compared to methods like enrichment on the individual data sets, is that in trying to incorporate the pathway decision into the modeling, we pick up a specific causal relationship between the different data sources. We avoid the problem of combining p-values from separate analyses, and can handle moderate amounts of missing data in the modeling. The method seems to perform slightly better in detecting pathways containing genes and metabolites with moderate differential expression, which can be a problem with rank-based enrichment methods.

For future versions of the paper we intend to test the model a sharp data set (not yet publicly released), as well as improving the penalty parameter selection by using an FDR based criterion instead of cross-validation. The overall pathway scoring scheme, now based on R^2 values, is also planned to be investigated further.

Contributions to Paper IV

The current author contributed to the paper by implementing and jointly developing the model, performing the literature review, designing the simulations, and writing up the paper.

Bibliography

Abate-Shen, C. and Shen, M. M. (2009). Diagnostics: The prostate-cancer metabolome. *Nature*, 457:799–800.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402.

Andersson, A., Lundgren, M., Eriksson, S., Rosenlund, M., Bernander, R., and Nilsson, P. (2006). Global analysis of mRNA stability in the archaeon *Sulfolobus*. *Genome Biol.*, 7(10):R99.

Baldi, P. and Long, A. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519.

Bammler, T., Beyer, R., Bhattacharya, S., Boorman, G., Boyles, A., Bradford, B., Bumgarner, R., Bushel, P., Chaturvedi, K., Choi, D., Cunningham, M., Deng, S., Dressman, H., Fannin, R., Farin, F., Freedman, J., Fry, R., Harper, A., Humble, M., Hurban, P., Kavanagh, T., Kaufmann, W., Kerr, K., Jing, L., Lapidus, J., Lasarev, M., Li, J., Li, Y., Lobenhofer, E., Lu, X., Malek, R., Milton, S., Nagalla, S., O'malley, J., Palmer, V., Pattee, P., Paules, R., Perou, C., Phillips, K., Qin, L., Qiu, Y., Quigley, S., Rodland, M., Rusyn, I., Samson, L., Schwartz, D., Shi, Y., Shin, J., Sieber, S., Slifer, S., Speer, M., Spencer, P., Sproles, D., Swenberg, J., Suk, W., Sullivan, R., Tian, R., Tennant, R., Todd, S., Tucker, C., Van Houten, B., Weis, B., Xuan, S., and Zarbl, H. (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods*, 2:351–356.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. (2008). GenBank. *Nucleic Acids Res.*, 36:25–30.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, 99(15):9697–9702.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351.

Chen, F., Mackey, A., Vermunt, J., and Roos, D. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2:e383.

Cochrane, G. R. and Galperin, M. Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res.*, 38:1–4.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C., Bofkin, L., Jones, T., Davis, R., and Steinmetz, L. (2006). A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5320–5325.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In Dayhoff, M., editor, *Atlas of Protein Sequences and Structure*, volume 5, pages 345–352. National Biomedical Research Foundation, Washington DC.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, 6:361–375.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32:407–499.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160.

Enright, A., Van Dongen, S., and Ouzounis, C. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30:1575–1584.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, 1:302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the grahical lasso. *Biostatistics*, 9:432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-prints*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.*, 33(1):1–22.

Gabaldón, T. and Huynen, M. (2004). Prediction of protein function and pathways in the genome era. *Cell. Mol. Life Sci.*, 61:930–944.

Garneau, N., Wilusz, J., and Wilusz, C. (2007). The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.*, 8:113–126.

Gibon, Y., Usadel, B., Blaesing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. *Genome Biol.*, 7:R76.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabe de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H. W., Wichmann, H. E., Weinberger, K. M., Adamski, J., Illig, T., and Suhre, K. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.*, 4:e1000282.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M., and Hughes, T. (2004). Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol. Cell. Biol.*, 24:5534–5547.

Guan, Q., Zheng, W., Tang, S., Liu, X., Zinkel, R., Tsui, K., Yandell, B., and Culbertson, M. (2006). Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast. *PLoS Genet.*, 2(11):e203.

Gutiérrez, R., Ewing, R., Cherry, J., and Green, P. (2002). Identification of unstable transcripts in arabidopsis by cDNA microarray analysis: rapid decay is associated with a group of touch- and specific clock-controlled genes. *Proc. Natl. Acad. Sci. U.S.A.*, 99(17):11513–8.

Hambraeus, G., von Wachenfeldt, C., and Hederstedt, L. (2003). Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. *Mol. Genet. Genomics*, 269:706–714.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA.

Hazen, S. P., Naef, F., Quisel, T., Gendron, J. M., Chen, H., Ecker, J. R., Borevitz, J. O., and Kay, S. A. (2009). Exploring the transcriptional landscape of plant circadian rhythms using genome tiling arrays. *Genome Biol.*, 10:R17.

Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919.

Hertz-Fowler, C., Peacock, C., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A., Rajandream, M., and Barrell, B. (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, 32:D339–343.

Higgins, D. and Sharp, P. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73:237–244.

HMDB (2010). The Human Metabolome Database (HMDB). `http://www.hmdb.ca/`.

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Res.*, 37:D690–697.

Hundt, S., Zaigler, A., Lange, C., Soppa, J., and Klug, G. (2007). Global analysis of mRNA decay in *Halobacterium salinarum* NRC-1 at single-gene resolution using DNA microarrays. *J. Bacteriol.*, 189:6936–6944.

Irizarry, R., Warren, D., Spencer, F., Kim, I., Biswal, S., Frank, B., Gabrielson, E., Garcia, J., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S., Hoffman, E., Jedlicka, A., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S., and Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, 2:345–350.

Jensen, R. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol.*, 2:INTERACTIONS1002.

Jordan, I., Kondrashov, F., Rogozin, I., Tatusov, R., Wolf, Y., and Koonin, E. (2001). Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.*, 2:RESEARCH0053.

Jörnsten, R. (2009). Simultaneous model selection via rate-distortion theory, with applications to cluster and significance analysis of gene expression data. *J. Comput. Graph. Statist*, 18:613–639.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, 38:D355–360.

Koonin, E. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.*, 37:4181–4193.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., and Higgins, D. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948.

Last, R. L., Jones, A. D., and Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. *Nat. Rev. Mol. Cell Biol.*, 8:167–174.

Li, L., Stoeckert, C., and Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13:2178–2189.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statist. Sinica*, 12:31–46.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat. Methods*, 4:613–614.

Meinshausen, N., Rocha, G., and Yu, B. (2007). Discussion: a tale of three cousins: Lasso, L2Boosting and Dantzig. *Ann. Statist.*, 35:2373–2384.

Meyer, S., Temme, C., and Wahle, E. (2004). Messenger RNA turnover in eukaryotes: pathways and enzymes. *Crit. Rev. Biochem. Mol. Biol.*, 39:197–216.

Molina-Navarro, M., Castells-Roca, L., Bellí, G., García-Martínez, J., Marín-Navarro, J., Moreno, J., Pérez-Ortín, J., and Herrero, E. (2008). Comprehensive transcriptional analysis of the oxidative response in yeast. *J. Biol. Chem.*, 283:17908–17918.

Narsai, R., Howell, K. A., Millar, A. H., O'Toole, N., Small, I., and Whelan, J. (2007). Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. *Plant Cell*, 19:3418–3436.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Nicholson, J. K. and Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature*, 455:1054–1056.

Ortega, A. and Ramchandran, K. (1998). Rate-distortion methods for image and video compression. *IEEE Signal Process. Mag.*, 15:23–50.

Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S., Lee, S., Suh, D., Hong, D., Kang, H. P., Yoo, Y. J., Shin, J. Y., Kim, H. J., Yavartanoo, M., Chang, Y. W., Ha, J. S., Chong, W., Hwang, G. R., Darvishi, K., Kim, H., Yang, S. J., Yang, K. S., Kim, H., Hurles, M. E., Scherer, S. W., Carter, N. P., Tyler-Smith, C., Lee, C., and Seo, J. S. (2010). Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, 42:400–405.

Pelechano, V. and Pérez-Ortín, J. (2008). The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast. *Yeast*, 25:85–92.

Preiss, T., Baron-Benhamou, J., Ansorge, W., and Hentze, M. (2003). Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat. Struct. Biol.*, 10:1039–1047.

Qin, L., Beyer, R., Hudson, F., Linford, N., Morris, D., and Kerr, K. (2006). Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics*, 7:23.

Raghavan, A., Ogilvie, R., Reilly, C., Abelson, M., Raghavan, S., Vasdewani, J., Krathwohl, M., and Bohjanen, P. (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.*, 30(24):5529–38.

Redon, E., Loubière, P., and Cocaign-Bousquet, M. (2005). Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *J. Biol. Chem.*, 280:36380–36385.

Remm, M., Storm, C., and Sonnhammer, E. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041–1052.

Ren, B., Robert, F., Wyrick, J., Aparicio, O., Jennings, E., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T., Wilson, C., Bell, S., and Young, R. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309.

Ritchie, M., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23:2700–2707.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Selinger, D., Saxena, R., Cheung, K., Church, G., and Rosenow, C. (2003). Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.*, 13:216–223.

Shock, J., Fischer, K., and DeRisi, J. (2007). Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.*, 8:R134.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.

Smyth, G. (2005). LIMMA: Linear Models for Microarray Data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

Sonnhammer, E. and Koonin, E. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, 18:619–620.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297.

Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R. J., Li, Y., Nyati, M. K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G. S., Ghosh, D., Pennathur, S., Alexander, D. C., Berger, A., Shuster, J. R., Wei, J. T., Varambally, S., Beecher, C., and Chinnaiyan, A. M. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914.

Steuer, R. (2006). Review: on the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinformatics*, 7:151–158.

Storm, C. and Sonnhammer, E. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18:92–99.

Tatusov, R., Galperin, M., Natale, D., and Koonin, E. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, 28:33–36.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816.

The Gene Ontology Consortium (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, 38:D331–335.

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38:D142–148.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288.

Tucker, M., Valencia-Sanchez, M., Staples, R., Chen, J., Denis, C., and Parker, R. (2001). The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in *Saccharomyces cerevisiae*. *Cell*, 104:377–386.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5116–5121.

Wall, D., Fraser, H., and Hirsh, A. (2003). Detecting putative orthologs. *Bioinformatics*, 19:1710–1711.

Wang, Y., Liu, C., Storey, J., Tibshirani, R., Herschlag, D., and Brown, P. (2002). Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, 99(9):5860–5865.

Webber, C. and Ponting, C. (2004). Genes and homology. *Curr. Biol.*, 14:R332–333.

Wishart, D. S. (2010). Computational approaches to metabolomics. *Methods Mol. Biol.*, 593:283–313.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, 37:D603–610.

Yamashita, A., Chang, T., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C., and Shyu, A. (2005). Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nat. Struct. Mol. Biol.*, 12:1054–1063.

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., and Darnell, J. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.*, 13:1863–1872.

Yang, Y., Buckley, M., and Speed, T. (2001). Analysis of cDNA microarray images. *Brief. Bioinformatics*, 2:341–349.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped varaibles. *J. Roy. Statist. Soc. Ser. B*, 68:49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37:3468–3497.

Zmasek, C. and Eddy, S. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35:2173–2192.

Paper I

Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments

LINA GUNNARSSON,[†]
ALEXANDRA JAUHIAINEN,[‡,⊥]
ERIK KRISTIANSSON,[†,‡,§]
OLLE NERMAN,[‡,⊥] AND
D. G. JOAKIM LARSSON[*,†]

Department of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Box 434, SE-405 30 Göteborg, Sweden, Mathematical Statistics, Chalmers University of Technology, Mathematical Statistics, University of Gothenburg, SE-412 96 Göteborg, Sweden, and Department of Zoology, University of Gothenburg, SE-405 30 Göteborg, Sweden

Received February 20, 2008. Revised manuscript received May 13, 2008. Accepted May 19, 2008.

Pharmaceuticals are typically found in very low concentrations in the aquatic environment. Accordingly, environmental effects clearly assigned to residual drugs are consistent with high affinity interactions with conserved targets in affected wildlife species rather than with a general toxic effect. Thus, evolutionarily well-conserved targets in a given species are associated with an increased risk. In this study orthologs for 1318 human drug targets were predicted in 16 species of which several are relevant for ecotoxicity testing. The conservation of different functional categories of targets was also analyzed. Zebrafish had orthologs to 86% of the drug targets while only 61% were conserved in Daphnia and 35% in green alga. The predicted presence and absence of orthologs agrees well with published experimental data on the potential for specific drug target interaction in various species. Based on the conservation of targets we propose that aquatic environmental risk assessments for human drugs should always include comprehensive studies on aquatic vertebrates. Furthermore, individual targets, especially enzymes, are well conserved suggesting that tests on evolutionarily distant organisms would be highly relevant for certain drugs. We propose that the results can guide environmental risk assessments by improving the possibilities to identify species sensitive to certain types of pharmaceuticals or to other contaminants that act through well defined mechanisms of action. Moreover, we suggest that the results can be used to interpret the relevance of existing ecotoxicity data.

Introduction

A range of organisms are unintentionally exposed to a large number of pharmaceutical residues in their natural habitats (*1–4*). Since many drugs are designed to affect specific protein targets at relatively low doses, pharmaceuticals may become a potential environmental hazard even at low concentrations. To date there are only a few examples where the presence of pharmaceuticals in the environment has been clearly linked to adverse effects on wildlife. For example, the analgesic diclofenac fed to livestock has caused a dramatic decline in the populations of vultures in India and Pakistan (*5, 6*). Furthermore, the synthetic estrogen in human contraceptives, 17-α-ethinylestradiol (EE_2), strongly contributes to the feminization of fish observed downstream from sewage treatment works (*7–12*). The environmental effects of EE_2 and diclofenac (feminization and renal failure) would hardly have been predicted from the current standard tests applied in the human or veterinary risk assessment procedures (*13–15*). On the contrary, both these examples were identified in retrospect through field observations of specific, known effects of these drugs (*5, 7, 12, 16*). This suggests that prospective testing could be made more powerful by including *targeted test strategies*, i.e., a selection of species, tests, and end points based on the known pharmacological properties of the tested pharmaceutical (*17*).

For about ten years, environmental risk assessments have been compulsory for the approval of new pharmaceutical products in the United States and the European Union (*13, 15*). Only acute tests are required in the United States and tests on fish are not mandatory (*15*). In 2006, acute tests were abandoned from the European risk assessment procedures for human drugs, and replaced by chronic toxicity tests with Daphnia and algae, and semichronic early life stage tests with fish (*13*). Despite these new EU guidelines requiring more relevant chronic toxicity data, there is still little focus on targeted test strategies.

Pharmaceuticals are typically found in the aquatic environment at ng/L to low μg/L concentrations. Adverse effects in nontarget species are therefore most likely to occur as a consequence of specific drug target interactions rather than via an unspecific mode of action, such as narcosis. Consequently, evolutionarily well-conserved drug targets are likely associated with an increased risk for pharmacological effects of a given drug in exposed aquatic organisms. Indeed, several authors have pointed out that information on conservation of drug targets is a key aspect for the development of more efficient test strategies (*17–19*). For the majority of newly developed pharmaceuticals, the molecular human drug targets are known (*20*). Since complete genomes for a wide range of species are now available, it is possible to compile data on the evolutionary conservation of drug targets. Prediction of orthologs, i.e., proteins derived from a common ancestral protein at the time of speciation, is a common way to link functionally similar proteins among different species. Although orthology does not guarantee common function, the value of predicting orthologs has already been recognized in the field of pharmacology (*21*).

We propose that orthology data of human drug targets can add important information in order to direct future research efforts to assess the ecotoxicological risks posed by pharmaceuticals. In this study, we have therefore predicted orthologs for 1318 human drug targets in seven species commonly used for ecotoxicity testing (*Xenopus laevis, Xenopus tropicalis, Danio rerio, Gasterosteus aculeatus, Daphnia pulex, Chlamydomonas reinhardtii,* and *Synechococcus elongatus*) and in nine diverse species with well-known genomes (*Mus musculus, Gallus gallus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Dictyostelium discoideum, Tetrahymena thermophila, Saccharomyces cerevisiae,* and *Escherichia coli*).

[*] Corresponding author phone: +46-31-7863589; fax: +46-31-7863512; e-mail: joakim.larsson@fysiologi.gu.se.
[†] Department of Neuroscience and Physiology.
[‡] Mathematical Statistics, Chalmers University of Technology.
[⊥] Mathematical Statistics, University of Gothenburg.
[§] Department of Zoology.

10.1021/es8005173 CCC: $40.75
Published on Web 06/25/2008

TABLE 1. Protein Information for the Species Used in the Ortholog Identification Procedure

species (common name)	reported proteins in database	database	reference
Arabidopsis thaliana (thale cress)	32825	TAIR 8[a]	(*54*)
Caenorhabditis elegans (nematode)	23693	Wormpep 188[b]	(*55*)
Chlamydomonas reinhardtii (unicellular green alga)	14598	JGI Chlamy v3.0[c]	(*56*)
Danio rerio (zebrafish)	31743	Ensembl release 49[d]	
Daphnia pulex (water flea)	30940	JGI *Daphnia pulex* v1.0[e]	
Dictyostelium discoideum (cellular slime mold)	13488	DictyBase release 080330[f]	(*57*)
Drosophila melanogaster (fruit fly)	21017	FlyBase r5.6[g]	(*58*)
Escherichia coli K12 (enterobacterium)	4285	CMR data release 1[h]	(*59*)
Gallus gallus (chicken)	22195	Ensembl release 49[d]	(*60*)
Gasterosteus aculeatus (three-spined stickleback)	27577	Ensembl release 49[d]	
Mus musculus (mouse)	39276	Ensembl release 49[d]	(*61*)
Saccharomyces cerevisiae (baker's yeast)	5801	GENEDB version 2.1[i]	(*62*)
Synechococcus elongatus PCC 6301 (cyanobacterium)	2524	CMR data release 16[h]	(*63*)
Tetrahymena thermophila (ciliate)	24071	TGD release Oct 07[j]	(*64*)
Xenopus laevis (African clawed frog)	29943	Xenbase 2.1[k]	
Xenopus tropicalis (western clawed frog)	27711	Ensembl release 49[d]	

[a] http://www.arabidopsis.org. [b] http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml. [c] http://genome.jgi-psf.org/Chlre3. [d] http://www.ensembl.org. [e] http://genome.jgi-psf.org/Dappu1. [f] http://dictybase.org. [g] http://flybase.bio.indiana.edu. [h] http://cmr.tigr.org. [i] http://www.genedb.org/. [j] http://www.ciliate.org. [k] http://xenbase.org/.

Materials and Methods

Definition of Human Drug Targets. Information concerning all human drugs approved by the U.S. Food and Drug Administration (FDA) with specified drug targets were downloaded from DrugBank (*22*) (April 2008). Annotations for the drug targets were updated using the Uniprot/SwissProt (*23*) collection of human proteins (18886 sequences from release 55.1), resulting in 1318 unique human targets for a total of 1152 drugs. Consequently, many drugs have more than one target. The order of targets listed in DrugBank generally reflects their importance regarding therapeutic indication or physiological effect (*22*).

Ortholog Prediction. The OrthoMCL algorithm (*24, 25*) was used to predict orthologs for all human drug targets in the selected species (Table 1). All protein sequence data were downloaded in April 2008. With the OrthoMCL algorithm (*25*), putative ortholog pairs and paralog pairs (sequences within the same genome that are reciprocally most similar to each other) were first identified via BLASTP (*26*) as reciprocal best hits in and between the species of interest. The MCL algorithm (*27*) was then used to cluster the different protein sequences. The E-value score, used as a cutoff in the similarity searches, was the OrthoMCL default value 10^{-5}. The parameter governing cluster tightness, the Markov inflation index, was set to 1.3. Next, the clusters containing the human drug targets were extracted. For each cluster, all orthologs were aligned with the corresponding human drug target using Clustal W (v. 2.0) (*28*). Based on the alignments, a sequence similarity score (percent similarity) was calculated by a count of all matching amino acids divided by the length of the human drug target. For the species represented in each selected cluster, the orthologs with the highest sequence similarity to the human drug target were extracted. For groups of proteins within a species, the median sequence similarity is reported.

Simple Best BLAST Hits. One-way similarity searches for all drug targets were also performed with BLASTP to identify possible high-scoring nonorthologous matches. Each drug target was compared against the protein collections for all the chosen species in turn. Similarity scores were calculated for the best BLAST hit to each target using Clustal W alignments in the same way as in the ortholog prediction procedure.

Gene Ontology Classifications. The human drug targets were classified into five different functional categories: enzyme, receptor, ion channel, transporter, and "other" based on the Gene Ontology (GO) (*29*). The enzyme group consisted of drug targets annotated to six different GO categories and their children: oxidoreductase activity (GO:0016491), transferase activity (GO:0016740), hydrolase activity (GO:0016787), lyase activity (GO:0016829), isomerase activity (GO:0016853), and ligase activity (GO:0016874). The receptor group consisted of all drug targets assigned to the receptor activity category (GO:0004872) and its children, while the transporter group was defined as all drug targets annotated to transporter activity (GO:0005215) and its children, excluding ion channel activity (GO:0005216), which was defined as a separate group in a similar manner. The remaining drug targets not annotated to any of the given groups were assigned as "other". In total, 135 of the drug targets were assigned to more than one category. The two largest intersections were found between the receptor and enzyme groups and between the receptor and ion channel groups, respectively, sharing 53 and 41 drug targets. Annotations to two child terms of receptor activity, ligand dependent nuclear receptor activity (GO:0004879) and rhodopsin-like receptor activity (GO:0001584), were also studied. All GO assignments for the drug targets were made using the Gene Ontology Annotation (GOA) Database at EBI (version 62.0 of GOA Human from http://www.ebi.ac.uk/GOA/).

Statistical Analysis. Tests for over-representation and under-representation were performed for the conservation rates of the five functional GO categories (comparing each group with the overall conservation rate of the other drug targets in the study) in all of the investigated species. One-tailed versions of Fisher's exact test were used (*30*).

Results and Discussion

Ortholog Prediction. Figure 1 gives an overview of the number of predicted orthologs and their median sequence similarity compared to the human drug targets for all the investigated species. Orthologs for 1292 drug targets were identified in mouse with a similarity of 87% while chicken had 1151 orthologs with a similarity of 70%. In the aquatic vertebrates *X. tropicalis*, *D. rerio*, and *G. aculeatus*, 1137, 1136, and 1160 orthologs were predicted, respectively, with a similarity above 60%. Although the orthologs in *X. laevis* had a similarity of 65% (in agreement with the similarities in the other aquatic vertebrates), the number of orthologs was less than expected [940], probably due to the lack of full genome sequence information. The number of orthologs in Daphnia and Drosophila [808 and 755, respectively] were roughly the same but fewer than in the vertebrates. The similarity for the

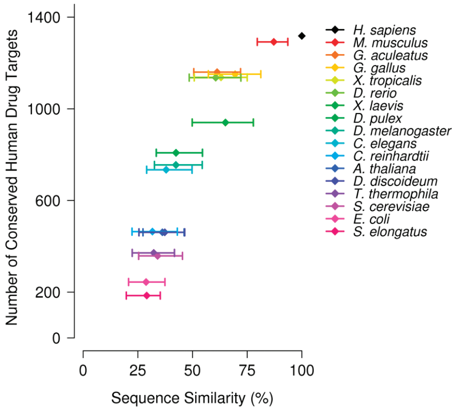


FIGURE 1. Median similarity and the number of predicted orthologs in all investigated species compared to the human drug targets. The boxes indicate 25% and 75% quantiles. Note that the ortholog prediction in *X. laevis* is based on comprehensive EST data and not on a fully sequenced genome.

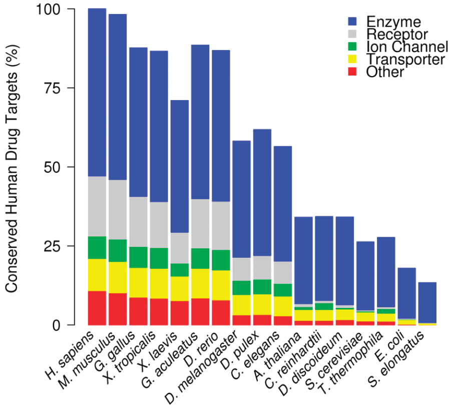


FIGURE 2. GO categories for the drug targets in human and in all the investigated species. Some drug targets were annotated to more than one functional category. Note that the ortholog prediction in *X. laevis* is based on comprehensive EST data and not on a fully sequenced genome.

orthologs in the two arthropods to the human drug targets was about 42%. *C. elegans* had slightly fewer orthologs [734] and a lower similarity of 38%. *A. thaliana* [461], the green alga [464], *D. discoideum* [461], the yeast [358], the ciliate [371], *E. coli* [244], and the cyanobacterium [185] all had relatively few orthologs, with a similarity of less than 37% (Figure 1 and Table S1 in the Supporting Information).

Among the evaluated species currently used for aquatic environmental risk assessments, fish and frog were predicted to have by far the greatest number of human drug target orthologs, with the highest degree of similarity. Therefore, it is more likely that low levels of residual drugs would result in a specific pharmacological target interaction in aquatic vertebrates than in the other species used for aquatic environmental risk assessments. Hence, we propose that aquatic toxicity testing for human pharmaceuticals should always include comprehensive tests with fish or amphibians. Tests on nonvertebrates would be particularly valuable for pharmaceuticals that have drug target orthologs in species more distantly related to man. In Table S1 in the Supporting Information we have compiled orthology predictions for 1318 human drug targets in species commonly used for environmental risk assessments and some additional species for comparative purpose. These data can be used to prioritize environmental risk assessment efforts and to interpret the relevance of existing ecotoxicity data.

Classification of Orthologs. The distribution of orthologs among different functional groups (enzymes, receptors, ion channels, transporters, and "others") for all the evaluated species as well as the human drug targets is illustrated in Figure 2. The most notable characteristic is that the proportion of receptors decreases while the proportion of enzymes increases with the evolutionary distance to man. In fact, drug targets annotated as receptors are under-represented within the ortholog groups in all tested vertebrates except mouse (*G. gallus* $p = 0.02$, *X. tropicalis* $p = 4 \times 10^{-7}$, *X. laevis* $p = 3 \times 10^{-15}$, *G. aculeatus* $p = 1 \times 10^{-3}$, *D. rerio* $p = 2 \times 10^{-3}$). The under-representation of ortholog receptors is even more clear in the remaining species (all p-values $< 3 \times 10^{-12}$). Proteins annotated as enzymes are over-represented in the aquatic vertebrates (*X. tropicalis* $p = 3 \times 10^{-6}$, *X. laevis* $p = 9 \times 10^{-13}$, *G. aculeatus* $p = 4 \times 10^{-7}$, *D. rerio* $p = 8 \times 10^{-7}$) and weakly over-represented in mouse and chicken ($p = 0.03$ for both groups). The over-representation of ortholog enzymes is highly significant in all the remaining species (all p-values $< 2 \times 10^{-16}$).

Pharmacologically important subgroups of the receptor drug targets were further studied. For example, 101 of the drug targets had rhodopsin-like G-protein coupled receptor (GPCR) activity in human and at least 72 of these targets had orthologs in the aquatic vertebrates, while roughly 30 orthologs were found in the nematode or in the arthropods. Only one rhodopsin-like GPCR ortholog was found in the green alga and in the cyanobacterium, respectively. Similarly, almost all of the 23 drug targets with ligand-dependent nuclear receptor activity had orthologs in the aquatic vertebrates, whereas the nematode and the two arthropods had less than six orthologs for this category. The plant, the green alga, *D. discoideum*, *E. coli*, and the cyanobacterium all had one nuclear receptor target ortholog.

Receptors, especially GPCRs, constitute the most prominent family of validated pharmacological targets. Approximately 40% of all FDA approved drugs elicit their therapeutic effects by targeting rhodopsin-like GPCRs and nuclear receptors (*31*). The receptors were the least conserved drug target type (with low p-values for under-representation in all of the investigated species used for ecotoxicity testing, Figure 2) which compares favorably with previous results. For example, rhodopsin-like receptors occur in vertebrate genomes twice as frequently as they do in invertebrates, with the exception of odorant receptors (*32*). Thus, the choice of test species is particularly important for drugs that have receptors as targets.

Environmental Risk Assessments. The evolutionary distances between the studied species and man indicated in our results are not reflected in the current ecotoxicity testing relevant for new authorizations of pharmaceuticals within either the United States or the EU. No tests on vertebrates are required in the tier one ecotoxicity test in the United States (*15*). The assessment factor used to extrapolate toxicity to other species could be too small if a drug-target ortholog is absent in the test species. In the EU, semichronic, early life stage tests but no life cycle tests, are required on fish (*13*). However, chronic tests are demanded for Daphnia (reproduction test) and green alga or cyanobacteria (*13*). Ankley et al. (*17*) suggest that exotoxicity testing should be focused in two ways: (1) identification of drugs with the most potential to elicit adverse effects, and (2) determination of which species and end points should be used for testing.

Support for the Ortholog Prediction in Published Experimental Data. We argue that orthology prediction can be used as a method to deduce possible interactions between a human drug and target orthologs in species distantly related to man. We have listed examples from the literature (Table 2 and below), adding empirical support to this presumption.

Aldehyde dehydrogenase and inosine-5′-monophosphate dehydrogenase (IMPDH) are drug targets that were predicted to be well conserved. Accordingly, disulfiram inhibits aldehyde dehydrogenase in man as well as in a bacterium and in a plant (*33*). Similarly, functional IMPDH is described in a range of species from bacteria to mammals. Mycophenolic acid inhibits microbial IMPDH, albeit poorly (*34*), whereas it is a potent inhibitor in a fungus (*35*) and other eukaryotes.

Orthologs were predicted in almost all of the eukaryotic species for the statin drug target, HMG-CoA reductase. HMG-CoA reductase activity is described in eukaryotes, archaea, and in certain true bacteria. The drugs simvastatin and atorvastatin strongly inhibit growth of some fungal species and their effect can be reversed by providing metabolites (downstream from HMG-CoA-reductase in the cholesterol biosynthesis) to the growth-medium (*36*). Statins also inhibit HMG-CoA reductase activity in plants (*37*) and in invertebrates (*38*).

The 5HT transporter (sodium dependent serotonin transporter) and the voltage dependent sodium channels are drug targets for which orthologs were predicted in the two arthropods but for example not in the plant or the bacteria. The 5HT transporter is inhibited by selective serotonin reuptake inhibitors (SSRIs), e.g., in zebrafish (*39*), in Drosophila (*40*), and in a flatworm (*41*). To our knowledge, no 5HT transporter has been described in any fungi, plant, or algae. Indeed, a very high concentration of fluoxetine (1 mg/L) has no effect on either wet weight, frond number, chlorophyll, or cartenoids in a plant (*42*). The neuroactive compound lidocaine, acting on voltage dependent sodium channels in man, increases intracellular sodium concentration through the same mechanism also in a pond snail neuron (*43*). Lidocaine also blocks nerve conduction in crayfish (*44*) and affects sodium channels in eel (*45*).

The estrogen receptor and the type-1 angiotensin II receptor are examples of drug targets for which orthologs were predicted only in the vertebrates. EE_2 potently binds to the estrogen receptor (*46*) and induces effects at concentrations below 1 ng/L (*47, 48*). To our knowledge, no estrogen receptor has conclusively been characterized in any crustacean and a gene loss of the estrogen receptor is predicted in the arthropod lineage (*49*). The effects of high concentrations of estrogen in different crustaceans do not seem to be mediated through a specific steroid receptor target (*50*). A type-1 angiotensin II receptor (AT1) is described in teleost fish but experiments with the specific AT1 receptor antagonist losartan gave inconsistent results in fish, often acting as a partial agonist or as an inhibitor at high concentrations (*51*). Little is known about angiotensin receptors in invertebrates. In insects, angiotensin II receptors appear to be absent or at least very different from those in mammals (*51*).

The presence of a drug target ortholog in a species does not guarantee that a functional interaction with the drug can occur. Vice versa, functional interactions between a drug and other nonorthologous proteins are also possible. For example, an estrogen receptor ortholog has been described in a few mollusks but it is not activated by estradiol or other vertebrate steroids (*49*). However, EE_2 can induce reproductive responses in mollusks at low ng/L concentrations (*52*). This suggests that EE_2 could mediate its effect via either a nonortholog receptor or possibly via a noncharacterized ortholog. A more precise prediction of a potential drug target interaction might be possible with better knowledge about drug binding domains and the three-dimensional structure of the target proteins. However, drug binding domains are not extensively characterized and predicting the three-

TABLE 2. **Examples of Predicted Orthologs with Specific Drug Interactions in Distantly Related Species Supported in the Literature**[a]

human target	drug or class of drugs	*M. musculus*	*G. gallus*	*X. tropicalis*	*X. laevis*	*G. aculeatus*	*D. rerio*	*D. melanogaster*	*D. pulex*	*C. elegans*	*A. thaliana*	*C. reinhardtii*	*D. discoideum*	*S. cerevisiae*	*T. thermophila*	*E. coli*	*S. elongatus*
IMDH2: Inosine-5′-monophosphate dehydrogenase 2	mycophenolate mofetil	99	95	93	93	89	90	66	74	52	46	52	60	61	29	37	18
ALDH2: Aldehyde dehydrogenase	disulfiram	94	85	80	79	78	77	69	74	66	59	57	50	49	47	39	-
HMDH: HMG-CoA reductase	statins	93	86	74	93	75	77	44	41	25	30	-	29	32	10	-	-
SC6A4: 5HT transporter	SSRIs	93	76	57	-	69	66	49	49	42	-	-	-	-	-	-	-
SCN5A: Sodium channel protein type 5 subunit alpha	neuroactive drugs blocking voltage-gated sodium channels	94	76	56	16	61	62	41	40	-	-	19	-	-	14	-	-
ESR1: Estrogen receptor	estrogens	89	77	69	69	46	46	-	-	-	-	-	-	-	-	-	-
AGTR1:Type-1 angiotensin II receptor	certain antihypertensive agents	94	76	61	62	38	47	-	-	-	-	-	-	-	-	-	-

[a] Sequence similarities to the human drug target are given as percentages. Note that the ortholog prediction in *X. laevis* is based on comprehensive EST data and not on a fully sequenced genome. SSRI: Selective serotonin reuptake inhibitor. HMG-CoA: 3-hydroxy-3-methylglutaryl-coenzyme A reductase. -: ortholog absent.

dimensional structure from an amino acid sequence alone is a nontrivial undertaking.

Methodological Considerations. The physiological function of drug targets and the detoxification systems differ between species. The risk of a significant drug exposure may also vary between organisms as they occupy separate ecological niches. Therefore, it is a complex task to predict which species are the most sensitive. Thus, if orthologs are present in several groups of organisms, there is an incentive to test species from all groups if a significant exposure is plausible. Although not within the scope of this study, it should be pointed out that pharmaceuticals which have drug targets in bacteria or eukaryotic parasites should be comprehensively tested on organisms with a higher degree of similarity to those organisms. Accordingly, cyanobacteria, instead of green algae, are recommended by the European Medicines Agency for testing of antimicrobials, as they are more sensitive to these compounds (*13*).

The prediction of orthologs can be difficult, especially when the true ortholog has been lost or duplicated (paralogs) since speciation. The orthology data should therefore be interpreted carefully on an individual basis, especially when several paralogous genes are present in humans (Tables S1 and S2 in the Supporting Information). However, the OrthoMCL algorithm has previously been shown to perform well on a divergent set of eukaryotic genomes (*53*). Another uncertainty is the varying genome sequence reliability for the species used in our analysis. For example, the genome of *X. laevis* is not fully sequenced and the Daphnia's and the green alga's genomes have quite recently been sequenced. The annotations might therefore be less reliable in these genomes. To add confidence to the results, species with better studied genomes were included. A different challenge is that pharmaceuticals can have multiple targets at their therapeutic level. Some drugs have several targets through which their intended therapeutic activities are mediated, while other drugs show unintended polypharmacology ("dirty drugs"). The exact definition of a drug's targets is therefore debated. For example, Overington et al. (*31*) uses a strict definition, resulting in 266 targets for all FDA-approved drugs, while DrugBank, which uses a wider concept, contains more than five times as many targets (*22*).

We have compiled orthology predictions for all human drug targets defined by DrugBank in species commonly used for environmental risk assessments and some additional species for comparative purpose (Table S1 in the Supporting Information). When evaluating possible sensitive species for a certain drug both the primary and alternative drug targets should be considered. We suggest that the presented orthology predictions can be used as a guide to prioritize test species for a certain drug, to interpret the relevance of existing ecotoxicity data, or to deduce which pharmaceuticals may pose an increased risk to a certain organism group. A more comprehensive understanding of the mechanism of actions of drugs in wildlife at environmentally relevant concentrations would be valuable to assess the full potential of the proposed approach.

Acknowledgments

Authors L.G. and A.J. contributed equally to this work. This research was supported by the Foundation for Strategic Environmental Research (MISTRA), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS), the Swedish Research Council, the Gothenburg Stochastic Centre, Gothenburg Mathematical Modeling Centre and the Swedish Foundation for Strategic Research. Valuable help was received from the Daphnia Genomics Consortium (DGC; http://daphnia.cgb.indiana.edu) supported by the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231, and Los Alamos National Laboratory under Contract W-7405-ENG-36. We also thank Jenny Andersson. The authors declare that they have no competing financial interests.

Supporting Information Available

Ortholog predictions for the human drug targets in all the investigated species, sequence similarities, drugs targeting the human proteins, cluster information (recent paralogs), GO annotations and simple best BLAST hits, which includes nonorthologs, for the human drug targets in all investigated species. This information is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Heberer, T. Occurrence, fate, and removal of pharmaceutical residues in the aquatic environment: a review of recent research data. *Toxicol. Lett.* **2002**, *131*, 5–17.

(2) Kolpin, D. W.; Furlong, E. T.; Meyer, M. T.; Thurman, E. M.; Zaugg, S. D.; Barber, L. B.; Buxton, H. T. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999−2000: a national reconnaissance. *Environ. Sci. Technol.* **2002**, *36*, 1202–1211.

(3) Larsson, D. G.; de Pedro, C.; Paxeus, N. Effluent from drug manufactures contains extremely high levels of pharmaceuticals. *J. Hazard. Mater.* **2007**, *148*, 751–755.

(4) Boxall, A.; Long, C. Veterinary medicines and the environment. *Environ. Toxicol. Chem.* **2005**, *24*, 759–760.

(5) Oaks, J. L.; Gilbert, M.; Virani, M. Z.; Watson, R. T.; Meteyer, C. U.; Rideout, B. A.; Shivaprasad, H. L.; Ahmed, S.; Chaudhry, M. J.; Arshad, M.; et al. Diclofenac residues as the cause of vulture population decline in Pakistan. *Nature* **2004**, *427*, 630–633.

(6) Green, E. R; Newton, I.; Shultz, S.; Cunningham, A. A.; Gibert, M.; Pain, J. D.; Prakash, V. Diclofenac poisoning as a cause of vulture population declines across the Indian subcontinent. *J. Appl. Ecol.* **2004**, *41*, 793–800.

(7) Desbrow, C.; Rutledge, E. J.; Brighty, G. C.; Sumpter, J. P.; Waldock, M. Identification of estrogenic chemicals in STW effluent: 1. Chemical fractionation and in vitro biological screening. *Environ. Sci. Technol.* **1998**, *32*, 1549–1558.

(8) Jobling, S.; Beresford, N.; Nolan, M.; Rodgers-Gray, T.; Brighty, G. C.; Sumpter, J. P.; Tyler, C. R. Altered sexual maturation and gamete production in wild roach (*Rutilus rutilus*) living in rivers that receive treated sewage effluents. *Biol. Reprod.* **2002**, *66*, 272–281.

(9) Larsson, D. G. J.; Adolfsson-Erici, M.; Parkkonen, J.; Pettersson, M.; Berg, A. H.; Olsson, P. E.; Förlin, L. Ethinyloestradiol - an undesired fish contraceptive? *Aquat. Toxicol.* **1999**, *45*, 91–97.

(10) Purdom, C. E.; Hardiman, P. A.; Bye, V. J.; Eno, N. C.; Tyler, C. R.; Sumpter, J. P. Estrogenic effects of effluents from sewage treatment works. *Chem. Ecol.* **1994**, *8*, 275–285.

(11) Kidd, K. A.; Blanchfield, P. J.; Mills, K. H.; Palace, V. P.; Evans, R. E.; Lazorchak, J. M.; Flick, R. W. Collapse of a fish population after exposure to a synthetic estrogen. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 8897–8901.

(12) Parrott, J. L.; Blunt, B. R. Life-cycle exposure of fathead minnows (*Pimephales promelas*) to an ethinylestradiol concentration below 1 ng/L reduces egg fertilization success and demasculinizes males. *Environ. Toxicol.* **2005**, *20*, 131–141.

(13) EMEA. *Guideline on the Environmental Risk Assessment of Medical Products for Human Use*; EMEA/CHMP/SWP/4447/00; London, 2006.

(14) EMEA. *Guideline on Environmental Impact Assessment for Veterinary Medicinal Products in support of the VICH Guidelines GL6 and GL 38*; EMEA/CVMP/ERA/418282/05; London, 2007.

(15) FDA. *Guidance for Environmental Assessment of Human Drug and Biologics Applications*; Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research: Rockville, MD, 1998; available at http://www.fda.gov/cder/guidance/1730fnl.pdf.

(16) Meteyer, C. U.; Rideout, B. A.; Gilbert, M.; Shivaprasad, H. L.; Oaks, J. L. Pathology and proposed pathophysiology of diclofenac poisoning in free-living and experimentally exposed oriental white-backed vultures (*Gyps bengalensis*). *J. Wildl. Dis.* **2005**, *41*, 707–716.

(17) Ankley, G. T.; Brooks, B. W.; Huggett, D. B.; Sumpter, J. P. Repeating history: pharmaceuticals in the environment. *Environ. Sci. Technol.* **2007**, *41*, 8211–8217.

(18) Kostich, M. S.; Lazorchak, J. M. Risks to aquatic organisms posed by human pharmaceutical use. *Sci. Total Environ.* **2008**, *389*, 329–339.

(19) Seiler, J. P. Pharmacodynamic activity of drugs and ecotoxicology-can the two be connected? *Toxicol. Lett.* **2002**, *131*, 105–115.

(20) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–672.

(21) Searls, D. B. Pharmacophylogenomics: genes, evolution and drug targets. *Nat. Rev. Drug Discov.* **2003**, *2*, 613–623.

(22) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–906.

(23) UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2007**, *35*, D193–197.

(24) Chen, F.; Mackey, A. J.; Stoeckert, C. J., Jr.; Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **2006**, *34*, D363–368.

(25) Li, L.; Stoeckert, C. J., Jr.; Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *13*, 2178–2189.

(26) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(27) Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584.

(28) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.

(29) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **2000**, *25*, 25–29.

(30) Agresti, A. *Categorical Data Analysis*; Wiley-Interscience: Hoboken, NJ, 2002.

(31) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **2006**, *5*, 993–996.

(32) Rompler, H.; Staubert, C.; Thor, D.; Schulz, A.; Hofreiter, M.; Schoneberg, T. G protein-coupled time travel: evolutionary aspects of GPCR research. *Mol. Interv.* **2007**, *7*, 17–25.

(33) Velasco-Garcia, R.; Zaldivar-Machorro, V. J.; Mujica-Jimenez, C.; Gonzalez-Segura, L.; Munoz-Clares, R. A. Disulfiram irreversibly aggregates betaine aldehyde dehydrogenase-a potential target for antimicrobial agents against *Pseudomonas aeruginosa*. *Biochem. Biophys. Res. Commun.* **2006**, *341*, 408–415.

(34) Digits, J. A.; Hedstrom, L. Species-specific inhibition of inosine 5′-monophosphate dehydrogenase by mycophenolic acid. *Biochemistry* **1999**, *38*, 15388–15397.

(35) O'Gara, M. J.; Lee, C. H.; Weinberg, G. A.; Nott, J. M.; Queener, S. F. IMP dehydrogenase from *Pneumocystis carinii* as a potential drug target. *Antimicrob. Agents Chemother.* **1997**, *41*, 40–48.

(36) Macreadie, I. G.; Johnson, G.; Schlosser, T.; Macreadie, P. I. Growth inhibition of Candida species and *Aspergillus fumigatus* by statins. *FEMS Microbiol. Lett.* **2006**, *262*, 9–13.

(37) Brain, R. A.; Reitsma, T. S.; Lissemore, L. I.; Bestari, K.; Sibley, P. K.; Solomon, K. R. Herbicidal effects of statin pharmaceuticals in *Lemna gibba*. *Environ. Sci. Technol.* **2006**, *40*, 5116–5123.

(38) Zapata, R.; Piulachs, M. D.; Belles, X. Inhibitors of 3-hydroxy-3-methylglutaryl-CoA reductase lower fecundity in the German cockroach: correlation between the effects on fecundity in vivo with the inhibition of enzymatic activity in embryo cells. *Pest. Manage. Sci.* **2003**, *59*, 1111–1117.

(39) Wang, Y.; Takai, R.; Yoshioka, H.; Shirabe, K. Characterization and expression of serotonin transporter genes in zebrafish. *Tohoku J. Exp. Med.* **2006**, *208*, 267–274.

(40) Demchyshyn, L. L.; Pristupa, Z. B.; Sugamori, K. S.; Barker, E. L.; Blakely, R. D.; Wolfgang, W. J.; Forte, M. A.; Niznik, H. B. Cloning, expression, and localization of a chloride-facilitated, cocaine-sensitive serotonin transporter from *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 5158–5162.

(41) Patocka, N.; Ribeiro, P. Characterization of a serotonin transporter in the parasitic flatworm, *Schistosoma mansoni*: cloning, expression and functional analysis. *Mol. Biochem. Parasitol.* **2007**, *154*, 125–133.

(42) Brain, R. A.; Johnson, D. J.; Richards, S. M.; Sanderson, H.; Sibley, P. K.; Solomon, K. R. Effects of 25 pharmaceutical compounds to *Lemna gibba* using a seven-day static-renewal test. *Environ. Toxicol. Chem.* **2004**, *23*, 371–382.

(43) Onizuka, S.; Kasaba, T.; Hamakawa, T.; Ibusuki, S.; Takasaki, M. Lidocaine increases intracellular sodium concentration through voltage-dependent sodium channels in an identified lymnaea neuron. *Anesthesiology* **2004**, *101*, 110–120.

(44) Yano, T.; Ibusuki, S.; Takasaki, M. A comparison of intracellular lidocaine and bupivacaine concentrations producing nerve conduction block in the giant axon of crayfish in vitro. *Anesth. Analg.* **2006**, *102*, 1734–1738.

(45) Salazar, B. C.; Flash, D. O.; Walewski, J. L.; Recio-Pinto, E. Lidocaine has different effects and potencies on muscle and brain sodium channels. *Brain Res.* **1995**, *699*, 305–314.

(46) Denny, J. S.; Tapper, M. A.; Schmieder, P. K.; Hornung, M. W.; Jensen, K. M.; Ankley, G. T.; Henry, T. R. Comparison of relative binding affinities of endocrine active compounds to fathead minnow and rainbow trout estrogen receptors. *Environ. Toxicol. Chem.* **2005**, *24*, 2948–2953.

(47) Gunnarsson, L.; Kristiansson, E.; Forlin, L.; Nerman, O.; Larsson, D. G. J. Sensitive and robust gene expression changes in fish exposed to estrogen-a microarray approach. *BMC Genomics* **2007**, *8*, 149.

(48) Örn, S.; Holbech, H.; Madsen, H. T.; Norrgren, L.; Petersen, I. G. Gonad development and vitellogenin production in zebrafish (*Danio reio*) exposed to ethinylestradiol and methyltestosterone. *Aquat. Toxicol.* **2003**, *65*, 397–411.

(49) Thornton, J. W.; Need, E.; Crews, D. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* **2003**, *301*, 1714–1717.

(50) Clubbs, R. L.; Brooks, B. W. *Daphnia magna* responses to a vertebrate estrogen receptor agonist and an antagonist: a multigenerational study. *Ecotoxicol. Environ. Saf.* **2007**, *67*, 385–398.

(51) Russell, M. J.; Klemmer, A. M.; Olson, K. R. Angiotensin signaling and receptor types in teleost fish. *Comp. Biochem. Physiol., A* **2001**, *128*, 41–51.

(52) Jobling, S.; Casey, D.; Rogers-Gray, T.; Oehlmann, J.; Schulte-Oehlmann, U.; Pawlowski, S.; Baunbeck, T.; Turner, A. P.; Tyler, C. R. Comparative responses of molluscs and fish to environmental estrogens and an estrogenic effluent. *Aquat. Toxicol.* **2004**, *66*, 207–222.

(53) Chen, F.; Mackey, A. J.; Vermunt, J. K.; Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2007**, *2*, e383.

(54) Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **2000**, *408*, 796–815.

(55) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science***1998**, *282*, 2012–2018.

(56) Merchant, S. S.; Prochnik, S. E.; Vallon, O.; Harris, E. H.; Karpowicz, S. J.; Witman, G. B.; Terry, A.; Salamov, A.; Fritz-Laylin, L. K.; Marechal-Drouard, L.; et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **2007**, *318*, 245–250.

(57) Eichinger, L.; Pachebat, J. A.; Glockner, G.; Rajandream, M. A.; Sucgang, R.; Berriman, M.; Song, J.; Olsen, R.; Szafranski, K.; Xu, Q.; et al. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **2005**, *435*, 43–57.

(58) Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; et al. The genome sequence of *Drosophila melanogaster*. *Science* **2000**, *287*, 2185–2195.

(59) Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277*, 1453–1474.

(60) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution *Nature***2004**, *432*, 695–716.

(61) Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*, 520–562.

(62) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; et al. Life with 6000 genes. *Science* **1996**, *274*, 536–547.

(63) Sugita, C.; Ogata, K.; Shikata, M.; Jikuya, H.; Takano, J.; Furumichi, M.; Kanehisa, M.; Omata, T.; Sugiura, M.; Sugita, M. Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth. Res.* **2007**, *93*, 55–67.

(64) Eisen, J. A.; Coyne, R. S.; Wu, M.; Wu, D.; Thiagarajan, M.; Wortman, J. R.; Badger, J. H.; Ren, Q.; Amedeo, P.; Jones, K. M.; et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **2006**, *4*, e286.

ES8005173

Paper II

mRNA stability changes precede changes in steady-state mRNA amounts during hyperosmotic stress

CLAES MOLIN,[1] ALEXANDRA JAUHIAINEN,[2,3] JONAS WARRINGER,[1] OLLE NERMAN,[2,3] and PER SUNNERHAGEN[1]

[1]Department of Cell and Molecular Biology, Lundberg Laboratory, University of Gothenburg, SE-405 30 Göteborg, Sweden
[2]Department of Mathematical Statistics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
[3]Department of Mathematical Statistics, University of Gothenburg, SE-412 96 Göteborg, Sweden

ABSTRACT

Under stress, cells need to optimize the activity of a wide range of gene products during the response phases: shock, adaptation, and recovery. This requires coordination of several levels of regulation, including turnover and translation efficiencies of mRNAs. Mitogen-activated protein (MAP) kinase pathways are implicated in many aspects of the environmental stress response, including initiation of transcription, translation efficiency, and mRNA turnover. In this study, we analyze mRNA turnover rates and mRNA steady-state levels at different time points following mild hyperosmotic shock in *Saccharomyces cerevisiae* cells. The regulation of mRNA stability is transient and affects most genes for which there is a change in transcript level. These changes precede and prepare for the changes in steady-state levels, both regarding the initial increase and the later decline of stress-induced mRNAs. The inverse is true for stress-repressed genes, which become stabilized during hyperosmotic stress in preparation of an increase as the cells recover. The MAP kinase Hog1 affects both steady-state levels and stability of stress-responsive transcripts, whereas the Hog1-activated kinase Rck2 influences steady-state levels without a major effect on stability. Regulation of mRNA stability is a wide-spread, but not universal, effect on stress-responsive transcripts during transient hyperosmotic stress. By destabilizing stress-induced mRNAs when their steady-state levels have reached a maximum, the cell prepares for the subsequent recovery phase when these transcripts are to return to normal levels. Conversely, stabilization of stress-repressed mRNAs permits their rapid accumulation in the recovery phase. Our results show that mRNA turnover is coordinated with transcriptional induction.

Keywords: mRNA turnover; *Saccharomyces cerevisiae*; stress-activated MAP kinase; HOG pathway

INTRODUCTION

In order to maintain viability and proliferation during increasing turgor and concomitant molecular crowding, cells need to recognize and rapidly adapt to changes in extracellular osmolarity. In the yeast *Saccharomyces cerevisiae*, the required adaptation to hyperosmosis is mainly initiated by the high osmolarity glycerol (HOG) pathway. The HOG pathway consists of a stress-activated mitogen-activated protein (MAP) kinase (SAPK) core module upon which two independent upstream branches, the Sln1 and the Sho1 branches, converge. Activation of any of these branches by hyperosmotic stress leads to rapid phosphorylation of the MAP kinase kinase Pbs2, which in turn phosphorylates and activates the MAP kinase Hog1. In the initial shock phase following mild salt stress, a major fraction of phosphorylated Hog1 quickly (within 1 min) translocates to the nucleus (Maeda et al. 1994), where it resides for about 10 min. In the adaptation phase, nuclear phosphorylated Hog1 in turn induces, as well as represses, transcription of osmoresponsive genes (Gasch et al. 2000; Posas et al. 2000; Rep et al. 2000; Causton et al. 2001; Krantz et al. 2004) through interactions with different transcription factors, including Hot1, Sko1, and Smp1 (Rep et al. 2000; Proft et al. 2001; de Nadal et al. 2003), as well as through recruitment of the Rpd3 histone deacetylase (de Nadal et al. 2004). The genes induced by Hog1 are involved in stress defense processes, such as production of the osmolyte glycerol (Albertyn et al. 1994), ion homeostasis (Marquez and Serrano 1996), and redox metabolism (Schüller et al. 1994; Krantz et al. 2004). By contrast, genes repressed by Hog1 are mainly involved in

Reprint requests to: Per Sunnerhagen, Department of Cell and Molecular Biology, Lundberg Laboratory, University of Gothenburg, Box 462, Se-405 30 Göteborg, Sweden; e-mail: per.sunnerhagen@cmb.gu.se; fax: 46-31-7863801.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1403509.

translation, ribosome biogenesis, and amino acid synthesis (Mager and Varela 1993; Gasch et al. 2000). These responses essentially reflect a redirection of cellular resources from growth to stress adaptation. The levels of the stress-responsive mRNAs change transiently with a timing that depends on the severity of the stress (Rep et al. 1999). After treatment with 0.7 M NaCl, the levels of induced mRNAs peak after 45 min (Rep et al. 2000), while after 0.4 M of NaCl stress the peak appears as early as after 10 min (Posas et al. 2000). After adaptation, cells eventually resume growth in the recovery phase, which is characterized by lower levels of stress proteins and increased translational activity (Warner 1999).

Under hyperosmotic stress, Hog1 also phosphorylates substrates not known to have a role in transcription. Within 5 min of 0.4 M NaCl treatment, activated Hog1 targets the plasma membrane ion channels Nha1 and Tok1 (Proft and Struhl 2004). Cell cycle arrest is mediated by activated Hog1 after 10 min of treatment with 0.4 M NaCl in G1 through phosphorylation of the CDK inhibitor Sic1, or in G2 (Escote et al. 2004) through phosphorylation of the protein kinase Hsl1 (Clotet et al. 2006), respectively. Furthermore, activated Hog1 phosphorylates the MAP kinase activated kinase (MAPKAP kinase) Rck2 (Bilsland-Marchesan et al. 2000), which has been implicated in regulation of translation (Teige et al. 2001; Swaminathan et al. 2006). Both Hog1 and Rck2 contribute to cellular hyperosmotic and oxidative stress tolerance (Brewster et al. 1993; Bilsland et al. 2004; Swaminathan et al. 2006). In other eukaryotes, there is also evidence for the involvement of SAPK homologs in translational control. Thus, activation of the mammalian SAPK p38 stimulates translation of tumor necrosis factor α mRNA (Kontoyiannis et al. 2001; Hitti et al. 2006). The fission yeast SAPK, Sty1, binds to translation factors (Asp et al. 2008), and *sty1* mutants have defects in recovery of translation after stress (Dunand-Sauthier et al. 2005; Asp et al. 2008).

Messenger RNA levels are determined not only by transcription rates but also by degradation rates. Hence, regulation of degradation rate in response to different stimuli is a potentially important stress response mechanism. For example, mammalian gene transcripts encoding interleukins and containing A/U-rich elements (AREs), which are involved in inflammation, are stabilized following signaling from the p38 or JNK SAPK pathways (Winzen et al. 1999; Chen et al. 2000). Similarly, the mammalian Rck2 homolog MAPKAPK-2 regulates mRNA stability through phosphorylation of the ARE-binding protein tristetraproline (Hitti et al. 2006). The stability of large functional mRNA groups can be co-regulated via mRNA sequence elements, such as the iron responsive element (IRE) in yeast (Puig et al. 2005).

Genome-wide analysis of intrinsic mRNA stability has been performed in *S. cerevisiae* using microarrays (Wang et al. 2002; Grigull et al. 2004; Duttagupta et al. 2005; Puig et al. 2005). In this study, we have used microarrays to investigate global changes in transcript stability in response to mild (0.4 M) NaCl stress in yeast. We find that the changes in mRNA stability during the adaptation and recovery phases are of a magnitude that could account for a major fraction of the overall changes in mRNA steady-state levels. The stability changes are most pronounced in the latest (recovery) phase of the stress response. During this phase, previously induced mRNAs encoding proteins involved in stress survival undergo a distinct loss of stability. Conversely, mRNAs that are transcriptionally repressed in the early adaptation phase, principally encoding growth-related functions, e.g., protein translation, are markedly stabilized in the recovery phase. Thus, stability changes precede changes in steady-state mRNA levels. We also find that most of these changes are dependent on Hog1 and, in some cases, on Rck2.

RESULTS

Several gene categories are affected at the stability level upon osmotic stress

Transcript steady-state levels are determined not only by transcription rate but also by mRNA stability. The importance of transcriptional regulation in response to salt stress has been well studied. In order to investigate the importance of a regulation of mRNA stability in response to stress, we collected global data on mRNA stability and steady-state levels from unstressed cells and after mild hyperosmotic (0.4 M NaCl) shock (Fig. 1A). 1,10-Phenanthroline (Phen) blocks de novo synthesis of transcripts by inhibiting RNA polymerase II (Brown 1994). Hence, changes in transcript levels in Phen-treated cells may be assumed to be due exclusively to transcript degradation (Rodriguez-Gabriel et al. 2003; Grigull et al. 2004; Lackner et al. 2007). On the basis of mRNA levels measured at different time points after Phen treatment, relative slope coefficients representing stability were calculated (stability indices, k_S) (see Materials and Methods). Positive and negative k_S values indicate transcripts more stable and unstable than the average transcript. Positive differences in stability indices (Δk_S) indicate stabilization (with an unknown scale since absolute half-lives were not measured), while negative differences indicate destabilization. Genome-wide stability indices (k_S) were calculated in unstressed cells and after 6 and 30 min after NaCl exposure. The amplitudes of the Δk_S indicate that the overall contribution of stress-induced changes in mRNA stability on steady-state levels is considerable (see Materials and Methods, section Modeling mRNA Stability).

Figure 1B shows the average stability indices (k_S) for the 38 Gene Ontology (GO) Slim categories of biological processes in wild-type (wt) and *hog1Δ* cells before and during stress. In wt cells, the mean mRNA stability of

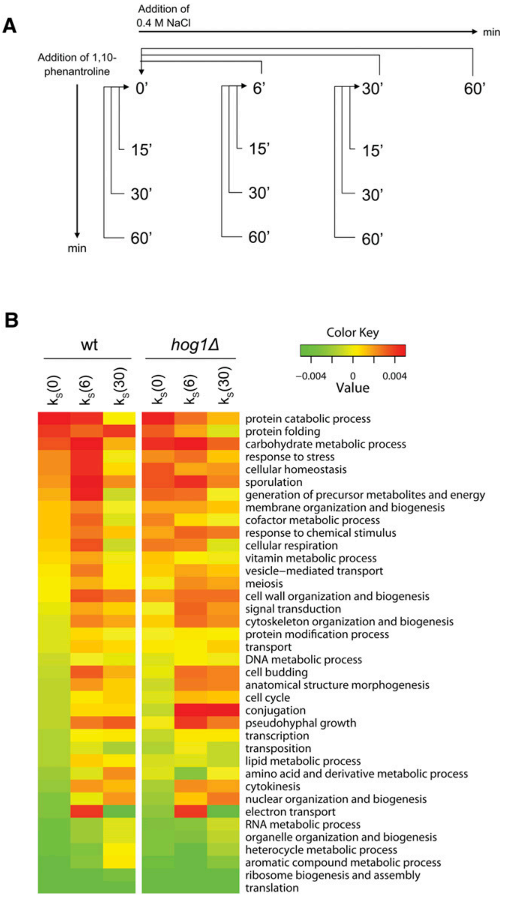


FIGURE 1. Several functional transcript categories are regulated at the stability level after salt stress. (*A*) Study design. Transcript steady-state levels were investigated at 6, 30, and 60 min after 0.4 M NaCl shock. Stability was investigated at the time points 0, 6, and 30 min by the addition of Phen. Aliquots were harvested at 15, 30, and 60 min after Phen addition to monitor mRNA decay and stability indices (k_S) were calculated. (*B*) mRNA stability in functional categories before and after stress. All 38 GO Slim broad functional categories (biological process) were ranked from most (*top*) to least stable (*bottom*) based on the mean k_S of unstressed wt cells (*left-most* column). Unstressed wt and *hog1Δ* cells have similar mean k_S for most categories. After 6 min of stress, categories that are initially stable tend to get more stabilized in the wild type, while initially unstable categories are further destabilized. After 30 min, the reverse is true, as stabilized categories are destabilized and vice versa. In the *hog1Δ* mutant, both steps in this response at the stability level are less pronounced.

several GO Slim categories increases after 6 min of stress, whereas after 30 min of stress almost all of these stabilized categories are destabilized to a level similar to or even below the original. These stabilized categories already tend to have a relatively high mean stability before stress. This includes processes such as "response to stress," the "carbohydrate metabolic process," and "cellular homeostasis." On the other hand, transcripts belonging to categories that already are unstable in unstressed conditions, such as "ribosome biogenesis and assembly" (RiBi) (Jorgensen et al. 2004), and the "RNA metabolic process," increase in stability in the later phase after stress. Before stress, wt and *hog1Δ* cells have very similar stability profiles (Fig. 1B). After stress, however, the mRNA stability response of the *hog1Δ* mutant clearly is weaker than that of the wt. Stability in the category "conjugation" stands out as misregulated in the *hog1Δ* mutant, indicating that the improper activation of the mating pheromone MAPK pathway observed in this mutant (O'Rourke and Herskowitz 1998) also extends to the mRNA stability regulation level.

Changes in mRNA stability precede changes in steady-state levels after hyperosmotic shock

Since there was a sharp decrease in mean stability after 30 min of stress of virtually all initially stabilized categories, indicating a narrow time frame of regulation, the temporal relationship between salt-induced changes in transcript stability (Δk_S) and changes in steady-state transcript levels (Δt_{TOT}) were investigated (Fig. 2A–D). We marked the 100 most up-regulated genes at 30 min in red in Figure 2, A–D. This group is dominated by genes known to be functionally important in the salt response (GO overrepresentation, Fisher's exact test: stress response [$P = 3 \times 10^{-16}$], carbohydrate metabolic process [$P = 6 \times 10^{-6}$], and sodium ion transport [$P = 0.0027$]).

Interestingly, the relative contribution of changes in transcript stability to the overall changes in transcript levels showed drastic temporal variations following salt addition. After 6 min of salt adaptation, a strong positive global correlation between Δt_{TOT} and Δk_S was observed (Fig. 2A), suggesting that changes in transcript stability account for a large fraction of the total salt-induced changes in transcript levels. Stability changes after 6 min also correlated positively with steady-state changes after 30 min, indicating that early stabilization influences later steady-state level changes (Fig. 2B), which demonstrates an expected lag in the effects of stabilizing mRNAs.

In contrast to the findings above, no correlation between Δk_S and Δt_{TOT} was seen after 30 min of salt adaptation (Fig. 2C), nor when comparing Δk_S (30) to Δt_{TOT} (60) min (Fig. 2D). Interestingly, however, at 30 min a group of genes corresponding to the red-colored salt-induced genes (including well-known salt targets such as the glycerol dehydrogenase *GPD1*, the glycerol phosphatase *HOR2*, and the aldose reductase *GRE3*) are still induced on a steady-state level, but now display a strongly reduced stability. This suggests that regulation of steady-state levels has a temporal dependency to mRNA stability through the

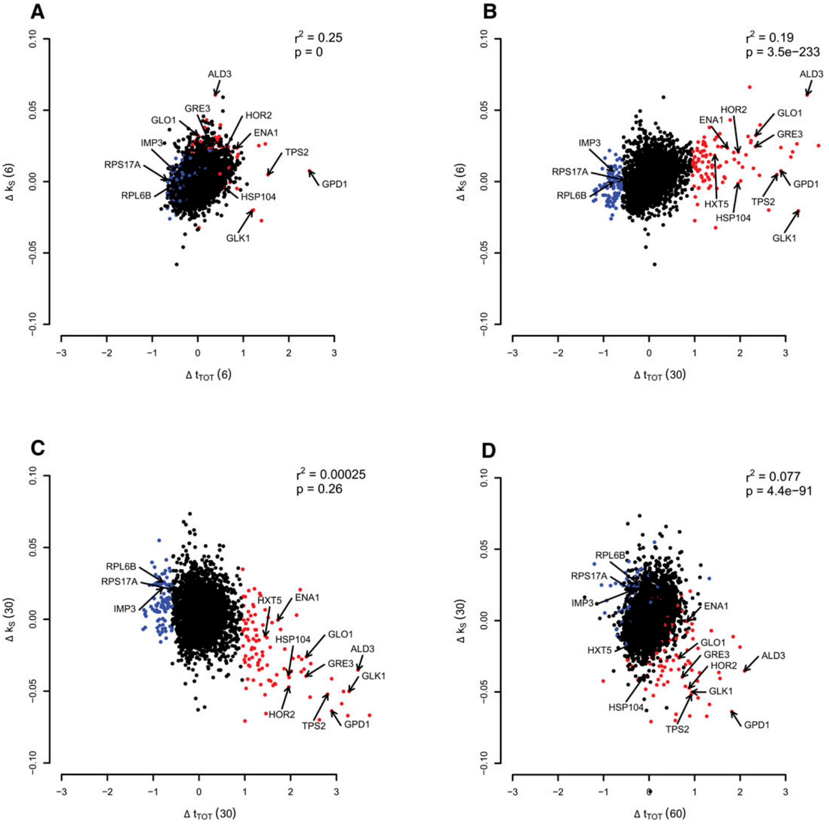


FIGURE 2. Transcript regulation at the stability level precedes the regulation at the steady-state level. Scatter plots of changes in mRNA stability (Δk_S) versus changes in steady-state levels (Δt_{TOT}) at various times after stress. The 100 most highly induced genes after 30 min at the steady-state level [Δt_{TOT} (30)] are colored in red, and the 100 most down-regulated genes at the same time point are colored blue. Ten well-known stress-responsive genes and three down-regulated genes required for protein synthesis are indicated by arrows (including the six genes verified by qPCR: *HOR2*, *GRE3*, *GPD1*, *IMP3*, *RPS17A*, and *RPL6B*). The alterations in steady-state levels and mRNA stability correlate positively when comparing stability after 6 min [Δk_S (6)] with steady-state levels after 6 min [Δt_{TOT} (6)] (*A*), as well as with steady-state levels after 30 min [Δt_{TOT} (30)] (*B*). The change in stability after 30 min [Δk_S (30)] was not globally correlated with the difference in steady-state levels after 30 min [Δt_{TOT} (30)] (*C*), but a group of outliers consisting of the most induced genes at the steady-state level (marked in red) show an inverse relationship, indicating that these salt-responsive genes are now destabilized. After 60 min of stress, the steady-state levels are lower for the salt-responsive genes (*D*). For information about the correlation test, see Materials and Methods.

phases of shock, adaptation, and recovery, and that increased mRNA degradation underlies the decline from the peak of steady-state levels in the late recovery phase. The indicated genes would correspond to transcripts that are required at high levels for salt adaptation, but not for recovery from salt stress. Out of the 100 most steady-state-induced genes at 30 min, 81 have significantly (moderated *t*-test: $P < 0.05$) reduced transcript stability at the same time point, indicating that a reduction in stability is the most common fate among this group of mRNAs. Also, the changes in stability between 6 and 30 min correlate with the changes in steady-state levels between 30 and 60 min (Spearman rank correlation $r^2 = 0.149$, Spearman rank test $P = 4 \times 10^{-176}$) (data not shown), strongly indicating that late in the salt response, mRNA turnover is a major factor driving changes in mRNA levels.

The 100 most down-regulated genes at Δt_{TOT} (30) are colored blue in Figure 2, A–D. This group includes ribosomal protein genes and the RiBi regulon encoding nucleolar proteins involved in ribosome biogenesis (GO overrepresentation, Fisher's exact test: translation [$P = 3 \times 10^{-3}$], ribosome biogenesis and assembly [$P = 3 \times 10^{-14}$]). These genes show an inverted, but less pronounced behavior compared to the up-regulated transcripts (Fig. 2, in red). Among these are the genes encoding the ribosomal proteins Rps17A and Rpl6B and the rRNA processor Imp3, for which the expression profiles were confirmed with quantitative PCR (qPCR) (Fig. 6C, see below).

Hence, at 6 min, transcription initiation and transcript turnover are in phase, stabilizing and transcribing transcripts critical for adaptation, whereas at 30 min the same transcripts are no longer stabilized, while still at a high steady-state level. Transcript turnover is at this time already preparing the entry into the recovery phase, suggesting a turning point of the response between 6 and 30 min.

Hog1 influences both early and late changes in mRNA stability

It is well established that the major part of the transcriptional initiation response following osmostress is mediated via the HOG pathway and its key component Hog1 (Posas et al. 2000; Proft et al. 2001; de Nadal et al. 2004). To investigate to what extent Hog1 also controls the regulation of transcript degradation during osmoadaptation, we compared the pattern of changes in transcript stability following 0.4 M NaCl stress in wt and *hog1Δ* mutant cells.

Confirming the function of Hog1 in transcriptional adaptation, we find that the average Δt_{TOT} (30) of the 100 most induced salt-responsive genes is significantly (Student's *t*-test, $P = 3 \times 10^{-27}$) lower in *hog1Δ* than in wt cells (Fig. 3A, upper panel). Correspondingly, for the 100 most repressed genes, the average Δt_{TOT} (30) is significantly higher in the *hog1Δ* mutant (Student's *t*-test, $P = 6.9 \times 10^{-11}$) (Fig. 3B, upper panel). Interestingly however, at 60 min the levels of these down-regulated genes have recuperated in the wt, while in the *hog1Δ* mutant, the levels keep declining until significantly lower than in the wt (Student's *t*-test, $P = 5 \times 10^{-25}$). Hog1 also has a clear effect on mRNA stability, especially with regard to the up-regulated genes. While similar in stability in unstressed cells, after 6 min of salt stress the top 100 salt-induced genes (Fig. 3A, lower panel) are stabilized in the wt but significantly less so in the *hog1Δ* mutant (Student's *t*-test, $P = 2 \times 10^{-4}$). Even more apparent is the difference in destabilization of these genes taking place between 6 and 30 min (Student's *t*-test, $P = 1.6 \times 10^{-7}$). As for the top 100 salt repressed genes, *hog1Δ* displays only a marginally enhanced destabilization after 6 min (Student's *t*-test, $P = 0.037$; Fig. 3B, lower panel), while the later stabilization is somewhat more affected (Student's *t*-test, $P = 5.5 \times 10^{-5}$). Hence, it is clear that Hog1 plays a critical role not only in regulating transcription initiation, but also transcript degradation rate in response to salt exposure, and most specifically so with regard to transcripts that are in high demand during the salt adaptation phase.

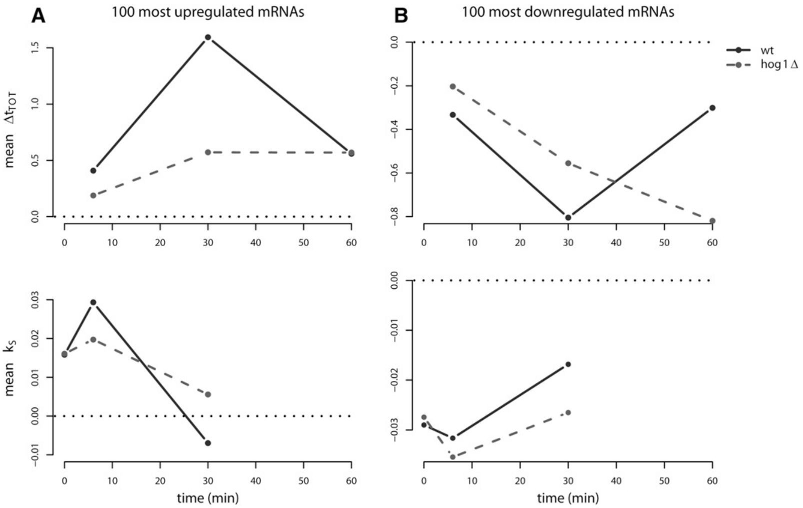


FIGURE 3. Hog1 affects both steady-state levels and stability. Average stability index (k_S) and steady-state log-fold changes (Δt_{TOT}) of the 100 most up-regulated (*A*) and down-regulated (*B*) genes at the steady-state level at 30 min in the wild type against time in wt and *hog1Δ* cells. Zero on the *y*-axis is indicated by a reference line (dotted black). These groups of genes display a distinct temporal pattern where changes in stability precede changes at the steady-state level. Hog1 influences both the changes in stability and at the steady-state level for both groups. Two-sample *t*-tests were used to calculate *P*-values for the average differences between the strains in all panels (see Materials and Methods).

The functional roles of mRNAs constrain their temporal profiles with regard to stability and steady-state levels

The temporal profile of stability and steady-state levels led us to define the intervals between 0 and 6 min (the "shock" phase), 6 and 30 min (the "adaptation" phase), and 30 and 60 min (the "recovery" phase). In order to investigate the impact of changing mRNA turnover rates in different functional categories in more detail, we utilized the highly resolved GO tree of the MIPS functional catalog (http://mips.gsf.de/genre/proj/yeast/), and examined the 200 most affected genes in the different phases with respect to Δk_S and Δt_{TOT} (Fig. 4). All categories exhibiting a significant enrichment in any of the intervals were included to get an overview of how the different categories behaved during the stress response. The categories that display the turning-point behavior seen in Figure 3A include "C (carbon)-compound and carbohydrate metabolism," "glycolysis and gluconeogenesis," and "stress response" (Fig. 4, blue dots). The transcriptional induction of genes in these categories during salt stress is well established. Carbohydrate metabolism, glycolysis and gluconeogenesis are important both for the production of the osmolytes glycerol and trehalose, as well as for glycogen, and for regulation of the energy needs of the cell. Stress response includes genes involved in sodium transport, redox metabolism, and heat shock proteins, all of which are known to be up-regulated after salt stress, but also several genes included in the carbohydrate metabolism processes mentioned above. Protein degradation (Fig. 4, violet dot) and related categories display a turning-point behavior at the steady-state level, while there is an immediate decrease in stability during the shock phase, which is continued into the adaptation phase. A plausible explanation is that the initial increase in steady-state levels is accomplished through transcription without an accompanying increase in stability. As can be seen in Figure 1B, the initial mean stability for the corresponding GO Slim category ("protein catabolic process") is very high, possibly negating a need for further stabilization in response to stress. An inverse behavior is seen for the categories involved in ribosome biogenesis (see below). Categories involved in amino acid metabolism (Fig. 4, yellow dot) show a tendency to be up-regulated in the recovery phase. It has previously been shown that salt stress causes starvation for amino acids through inhibition of their uptake, which in turn induces genes involved in their biosynthesis (Norbeck and Blomberg 1998; Pandey et al. 2007). The categories "DNA processing" and "cell cycle" are underrepresented throughout the response, indicating that those gene products are not in increased demand during the response. Categories involved in protein

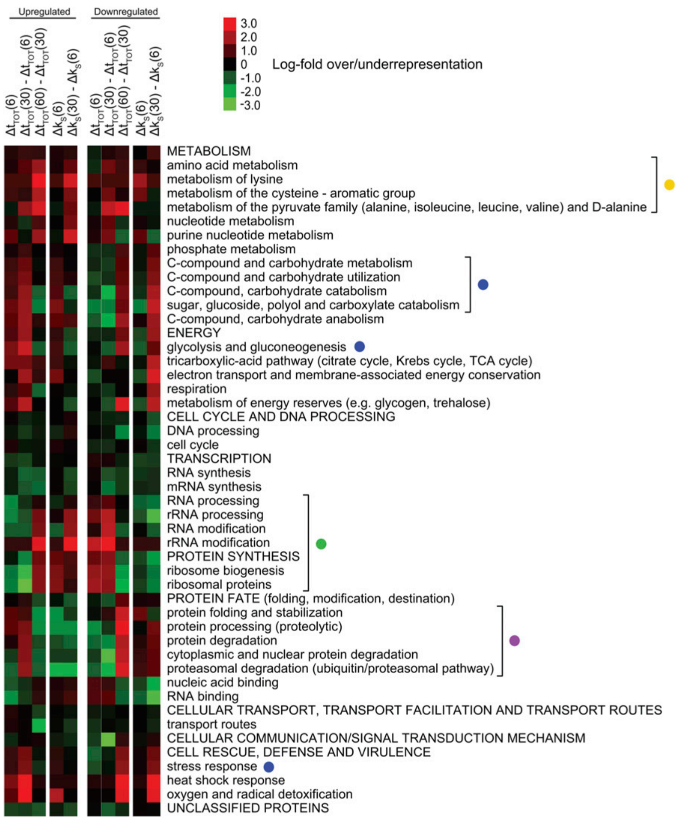


FIGURE 4. Functional categories have different temporal regulation profiles at the steady-state and the stability levels. Enrichment analysis (using hypergeometrical distribution) of functional categories on the 200 most up-regulated (*left* panel) and down-regulated (*right* panel) transcripts in the different phases of the response: the shock phase (0–6 min), the adaptation phase (6–30 min), and the recovery phase (30–60 min) in the wt. In each panel, the first three lanes show changes in steady-state levels, and the last two lanes show changes in stability. Only the categories significantly over- or underrepresented in any of the conditions (P-value < 0.01, Bonferroni corrected) were included in the plot. Red indicates over-representation and green indicates underrepresentation. Functional categories emphasized in the text are bracketed and marked with colored dots.

synthesis, such as rRNA processing, RiBi, and ribosomal proteins (Fig. 4, green dot) mirror the profiles seen in Figure 3B, although an early destabilization in the shock phase is not so apparent. Together, these observations imply that the temporal dependency of the steady-state levels to the stability regulation is a part of the response of several functional categories, first and foremost those involved in carbohydrate metabolism, response to stress, amino acid synthesis, protein degradation, and protein synthesis.

A subgroup of stress response genes is responsible for the distinct turning-point behavior

Hierarchical clustering of all 5551 genes reveals a subgroup of 338 genes with a turning-point behavior similar to the one seen in Figure 3A (data not shown). This is the most striking subgroup, with the biggest response magnitude both with respect to the steady-state and stability levels. GO enrichments include most categories mentioned above, including the "trehalose metabolic process" ($P = 9.6 \times 10^{-7}$), catabolic process ($P = 4.8 \times 10^{-5}$), and response to stress ($P = 0.0008$).

To provide a higher resolution of the stress-responsive transcripts, the 430 genes of the GO Slim category response to stress, encompassing a wide diversity of functions, were hierarchically clustered according to their profiles during the different phases at the steady-state as well as the stability levels in the wt and shown in Figure 5A. A group of 113 genes are responsible for the "turning-point" profile, while the rest of the genes show only a small or no response. This cluster consists of two subgroups that mainly differ in that one "early stabilized" cluster is stabilized in the shock phase, while the other, "destabilized" cluster is destabilized in the shock phase. The early-stabilized cluster contains well-known osmotic stress response genes such as the methylglyoxal reductase *GRE2*, the glycerol phosphatase *RHR2*, the aldehyde dehydrogenase *ALD3*, the dihydroxyacetone kinase *DAK1*, and the trehalose phosphate phosphatase *TPS2*. The destabilized cluster contains genes less well known in this context, but includes several chaperone-coding genes, such as *SSA1, HSP104, HSP78*, and *HSC82* (protein folding; GO enrichment, Fisher's exact test, $P = 0.007$).

The mean steady-state and stability changes in all phases were calculated for the two clusters for wt and *hog1Δ*, as well as for *rck2Δ* (Fig. 5B). In the wt, the early stabilized cluster has a stronger response at all time points, especially on the stability level where it displays a very clear turning-point behavior between 6 and 30 min. A similar turning-point behavior was observed on the steady-state level between 30 and 60 min, again reflecting the temporal difference between stability and steady-state level regulation. Hog1 affects both levels of regulation for this cluster, while in *rck2Δ* mutants, effects are seen mainly on the steady-state level. Contrary to expectations, this links Rck2 to transcription rather than stability regulation. This is especially true regarding the shock and adaptation phases, as the destabilization during the recovery phase is only mildly affected in the *rck2Δ* mutant. The stability indices and steady-state levels of three genes involved in the stress response (*HOR2, GRE3*, and *GPD1*) were confirmed with qPCR (Fig. 5C). Agreement between the qPCR data and the microarray data is very good, and the turning-point behavior, which is severely hampered in the mutants for these three genes, is clearly seen. qPCR confirmation of the steady-state level changes is also well in agreement with microarray data (Supplemental Fig. S3), although the microarray data give consistently lower responses ("signal compression," a well-known phenomenon) (see Materials and Methods).

In addition, the GO Slim categories "carbohydrate metabolism," "amino acid metabolism," and "proteasomal catabolism" were clustered (Supplemental Fig. S4A–C). All three categories contain groups of stress-responsive genes with an increase in steady-state levels along with a distinct destabilization during the adaptation phase (a turning-point behavior). The turning-point cluster from carbohydrate metabolism contains 28 genes, which are involved mainly in trehalose metabolism (trehalose metabolic process, Fisher's exact test, $P = 1.35 \times 10^{-16}$) and glucose transport ($P = 3.55 \times 10^{-5}$). This cluster shares 11 genes with the two clusters from response to stress of which nine are from the early stabilized cluster. The category amino acid metabolism mainly contains transcripts, which are weakly stabilized in the adaptation phase and up-regulated in the recovery phase (which is the behavior seen in Fig. 4), but a cluster of 33 genes instead has a turning-point profile. GO enrichment analysis of this group as compared to the rest of the category reveals that several of the genes in this group are involved in catabolic processes (catabolic process, $P = 1.44 \times 10^{-10}$) including the amino acid derivative catabolic process ($P = 1.66 \times 10^{-6}$), proline catabolic process ($P = 4.91 \times 10^{-5}$), and acetyl-CoA catabolic process ($P = 6.8 \times 10^{-5}$). A large cluster from protein catabolism (67 genes) is homogenous but does not have a prominent early stabilization. See Supplemental Table S5 for cluster gene lists.

Increased transcript stability contributes to restoration of ribosomal components in the recovery phase after NaCl exposure

Genes with ribosomal functions display quick and strong variations in transcript levels during the salt stress response phases (Gasch et al. 2000; Yale and Bohnert 2001). Our data show heterogeneity within this functional group with respect to changes in transcript stability (Δk_S) and transcript steady-state levels (Δt_{TOT}) concerning timing and strength of the shifts. A hierarchical clustering was performed

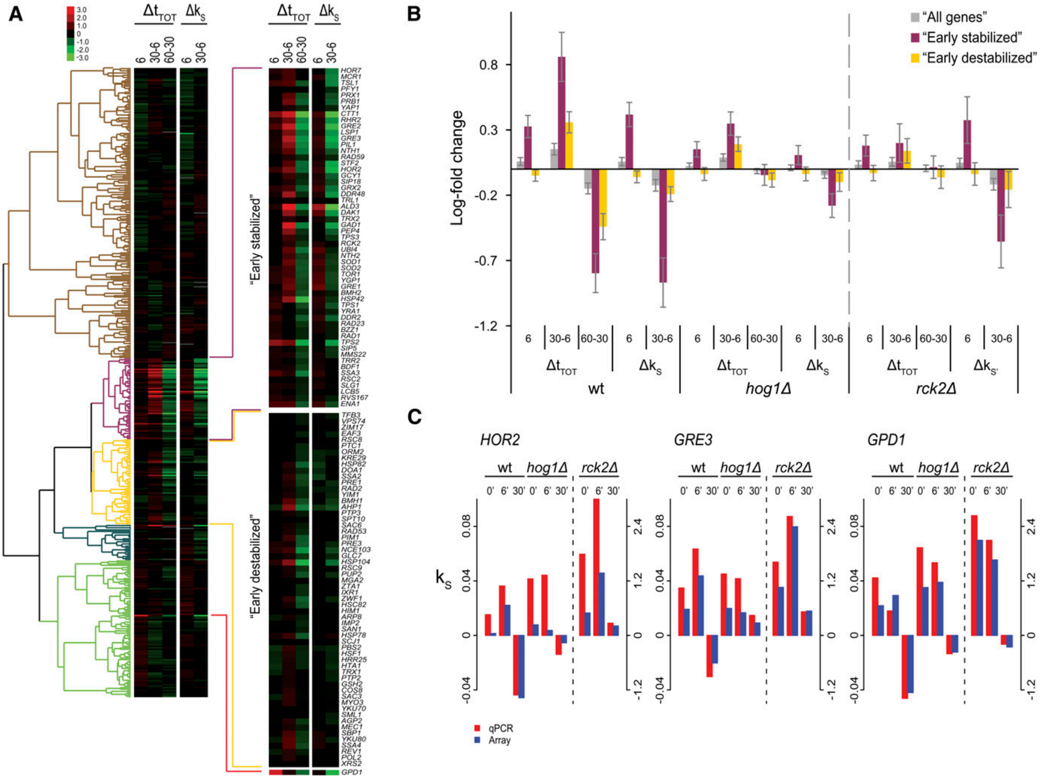


FIGURE 5. Cluster analysis of the GO Slim category *Response to stress*. (*A*) The genes in the GO Slim category response to stress were hierarchically clustered (uncentered Pearson correlation metric) with respect to the behavior of the transcripts in the wt during the phases of salt stress response. Genes with more than one missing value were omitted. To obtain value ranges comparable between steady state and stability, the stability indices were multiplied by 30 to approximate the log-fold differences 30 min after transcription inhibition (see Materials and Methods). The two clusters ("Early stabilized" and "Early destabilized," shown magnified to the *right*) include genes responsive to 0.4 M NaCl stress. *GPD1* fell outside the two clusters because of its fast up-regulation at the steady-state level. (*B*) Mean changes in steady-state and stability levels were calculated for the two clusters defined in (*A*) across the strains (wt, *hog1Δ*, and *rck2Δ*) shows that Hog1 influences both steady-state levels and stability, especially for the early stabilized cluster, while Rck2 mainly influences steady-state levels without major effects on stability. Error bars denote 95% confidence intervals. (*C*) Stability indices were confirmed by qPCR for three genes (*HOR2*, *GRE3*, and *GPD1*). *HOR2* and *GRE3* are part of the early-stabilized cluster, while *GPD1* fell outside the clusters. Spearman rank correlation between array and qPCR data: 0.9 (see Materials and Methods).

of the GO Slim category "translation," which contains 137 cytoplasmic (cRP) and 77 mitochondrial (mRP) ribosomal proteins along with slightly more than a hundred different translation factors and other proteins involved in the process of translation (all denoted TFs). The clustering resolved the translation genes into three large and two smaller groups with different expression profiles during the three phases of stress (Fig. 6A; see Supplemental Table S5 for lists of genes within each cluster). The mean Δt_{TOT} and Δk_S of the three large clusters across the different phases in wt, *hog1Δ*, and *Δrck2* cells were calculated (Fig. 6B). The first cluster ("destabilized cRPs") contained mainly cRPs and TFs and displayed an initial destabilization in the shock phase with a later increase in stability in the adaption phase. At the steady-state level, this cluster is repressed in the shock as well as in the adaption phase in the wt, while it is clearly up-regulated between 30 and 60 min, presumably reflecting the increase in stabilization. The cluster "stabilized cRPs" contains most of the remaining cRPs, along with a smaller proportion of TFs. This cluster is immediately stabilized in the wt, and even more so between 6 and 30 min, while the changes at the steady-state levels are less pronounced than for the destabilized cRPs cluster. Presumably this strong stabilization cluster corresponds to

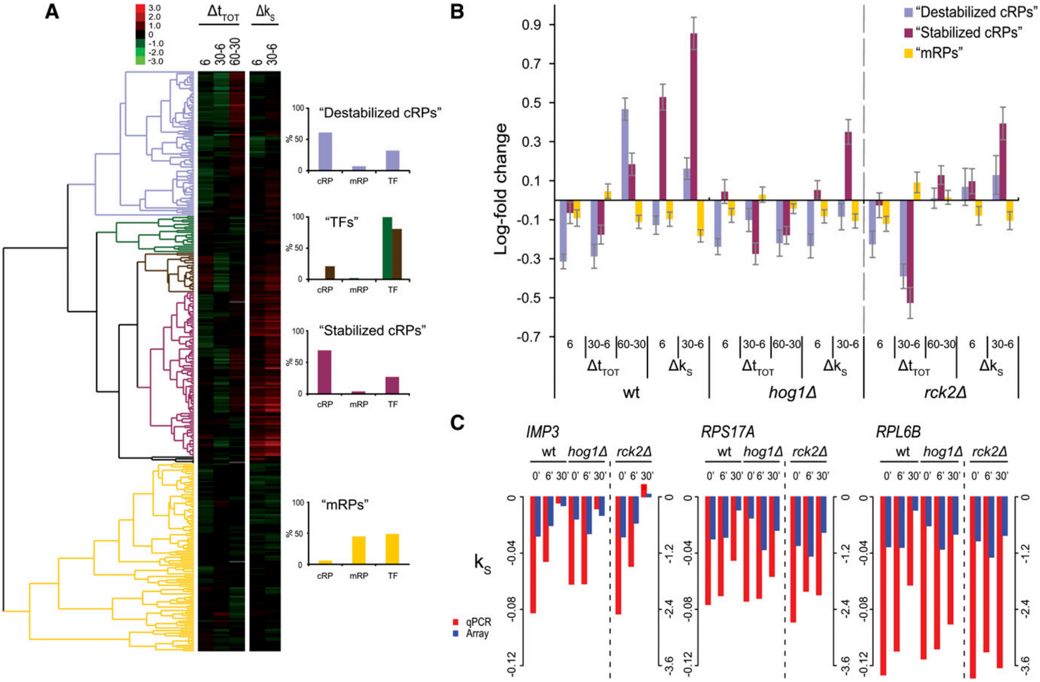


FIGURE 6. Cluster analysis of the GO Slim functional category translation. (*A*) The genes in the GO Slim category *Translation* were hierarchically clustered according to transcript behavior in the wild type, and the Δk_S values were multiplied by 30 as in Fig. 5 (genes with >1 missing value omitted). Five clusters were defined, highly enriched for cytoplasmic ribosomal proteins (cRPs), mitochondrial ribosomal proteins (mRPs), and translation factors (TFs), respectively, as shown in bar diagrams to the *right*. Note that the composition of clusters 2 and 3 are shown in the same diagram. (*B*) The mean log fold change at the steady-state and stability levels of the three RP-enriched clusters are plotted. The mean changes across the conditions were calculated for wt, *hog1Δ*, and *rck2Δ* cells. The mRPs cluster show only small changes during stress, while the cRPs are divided into two clusters, which differ in their regulation at the steady-state as well as the stability level. Especially the recuperation in the recovery phase is affected in both mutants to a similar extent. Error bars denote 95% confidence intervals. (*C*) The stability indices of two cRPs (*RPS17A* and *RPL6B*) were confirmed by qPCR, along with a gene in the RiBi group (*IMP3*). Spearman rank correlation between array and qPCR data: 0.9 (see Materials and Methods).

cRPs that are critical for the very early recovery from salt stress. The cluster "mRPs" contains about equal amounts mRPs and TFs but almost no cRPs. This cluster displays only marginal changes in stability and steady-state levels during the stress response, which is in line with earlier observations of mRP mRNA level responses to other stresses (Ihmels et al. 2002; Ihmels et al. 2005; Swaminathan et al. 2006).

The behavior of both cRP clusters was partially dependent on a functional HOG pathway. This is most apparent in the recovery phase where both the stabilization and the steady-state level induction are affected in both *hog1Δ* and *rck2Δ*, although the failure to up-regulate the steady-state level in the recovery phase is more apparent in the *hog1Δ* mutant. Hence, at least a part of the cRP mRNA regulation following NaCl exposure is dependent on a functional HOG pathway. To verify the expression profiles of cRPs following salt exposure, we quantified the decay rates of two cRP transcripts, *RPS17A* and *RPL6B*, independently and individually using qPCR. The relative trends were essentially confirmed, showing a marked stabilization of both transcripts at 30 min after salt exposure (Fig. 6C). Although these two genes belong to the destabilized cRP cluster, neither appeared destabilized in the shock phase either by arrays or qPCR. Instead, both displayed a very small stabilization. The qPCR data show that both these genes appear to be slightly more stable in the mutants in unstressed cells, but are destabilized during the shock phase in both *hog1Δ* and *rck2Δ* mutants. During the adaptation phase, however, they are not stabilized to the extent of the wt. *IMP3*, a gene belonging to the functional category ribosomal biogenesis and assembly (RiBi) was also included in the qPCR analysis. Genes belonging to RiBi behaved in the wt essentially like translation (named ribosomal proteins in the MIPS functional catalog) (Figs. 3, 4), with a strong stabilization in the adaptation phase, but *IMP3* does not seem to be affected in a *hog1Δ* mutant, except that the initial stability appears to be higher, as is the case for the cRPs.

DISCUSSION

Stability regulation precedes changes in steady-state levels

We have found that almost all genes known to be induced by salt shock at the steady-state level are also regulated at the stability level, suggesting that regulation of mRNA stability is a general and integral part of the regulation of transcript levels, and that most stress-responsive mRNAs have a dynamic range of half-lives that can be modulated to fit the situation, in line with earlier studies of the response to diauxic shift (Garcia-Martinez et al. 2004). The changes in stability can be fast and dramatic, exemplified by such well-known salt stress induced genes as *GPD1*, *GRE3*, and *HOR2*, which are stabilized early on after stress (0–6 min, the shock phase) and then greatly destabilized between 6 and 30 min (the adaptation phase). In fact, the relative stability of many salt-induced genes drop from being among the most stable in the shock phase to being among the least stable in the entire transcriptome during the adaptation phase. The steady-state levels of these genes only start to fall after 30 min (the recovery phase), indicating that the regulation at the stability level precedes the changes at the steady-state level. Genes that are repressed at the steady-state level display an inverse pattern with a similar temporal difference in regulation. It might be argued that within the population of induced transcripts, those species transcribed early would age and become susceptible to later degradation. However, this would not provide a satisfactory explanation for the behavior of repressed transcripts, which instead become stabilized late in the response. A recent study (Shalem et al. 2008) where fast destabilization of induced mRNAs was also observed, led to the seemingly paradoxical conclusion that changes in transcript abundance and stability in yeast cells were counterdirectional at a particular time point in two stress conditions. We believe that this paradox is resolved by considering the delay that occurs between a change in stability and resulting change in steady-state level for an mRNA: A stability change at an earlier time point is homodirectional with a change in abundance at a later time, as seen in our work.

It is important to emphasize, however, that although the phenomenon of a contribution from stability changes to the stress-induced increases or decreases in steady-state level is wide-spread, it does not apply to all stress-regulated mRNAs. This is noticeable for certain functional groups: among the stress-induced genes, there is no stability increase in the early destabilized cluster (Fig. 5). Similarly, for the cluster enriched for gene products with a role in protein catabolism, there is no mRNA stabilization despite a marked increase in steady-state mRNA levels (Supplemental Fig. S4).

Functional characteristics of the mRNA turnover response

In response to salt stress, the cell up-regulates the expression of genes involved in processes such as glycerol and trehalose metabolism, glycolysis and ion homeostasis, in order to relieve the stress. mRNAs encoding proteins involved in all these processes display the behavior seen for the above-mentioned genes, *GPD1*, *HOR2*, and *GRE3*, suggesting that the processes that are induced have an as well-defined and common regulation at the mRNA stability level as they do at the level of transcription initiation. Additionally, groups of genes with a role in amino acid metabolism and protein catabolism show a similar behavior. The pattern of an increase in steady-state levels and a decrease in stability during the adaptation phase followed by a decrease in steady-state levels in the recovery phase is common to all these genes, but there are differences in the stability regulation during the shock phase. This could reflect a difference in the timing or the intensity of transcriptional activity during the shock phase.

The production of osmolytes potentially involves several parts of the cellular metabolism; aside from the direct production of glycerol and trehalose, sugar transport and glycolysis, and thus, energy metabolism is affected since the metabolites involved are in constant flux. This should also affect amino acid synthesis, which shares much of the same metabolites, and this may be the explanation why amino acid catabolic processes are transiently up-regulated in response to stress. The slow up-regulation and stabilization of the majority of the mRNAs annotated to amino acid synthesis might reflect a recovery process, or it could reflect the amino acid starvation induced by hyperosmotic stress caused by an impaired ability to import amino acids (Norbeck and Blomberg 1998; Pandey et al. 2007). A group of mRNAs involved in protein catabolism is transiently up-regulated, but not stabilized in response to stress. This could be an effort to remove damaged or unwanted proteins, or a response to amino acid starvation, elevating recycling of protein. The early destabilization might indicate that changes in transcription rate are more important than changes in stability for this group. It should be noted that this group of genes have a very high initial stability (before stress), which might mean that transcripts can accumulate even though the stability declines.

In response to osmotic stress, the cell down-regulates the steady-state levels of transcripts involved in protein synthesis, cRP, and RiBi mRNAs. These transcripts are stabilized during the adaptation phase and their steady-state levels increase during the recovery phase. The cRPs follow a common trend, but differ considerably with respect to the amount and timing of stabilization. The fact that heterogeneity exists among the cRPs indicates that functional differences exist within this group, where a specific subgroup could be required early in recovery.

Generally, categories that are up-regulated during stress such as response to stress and carbohydrate metabolism tend to have a high mean stability already before stress. We find that categories with a high relative stability in the unstressed condition (most "stress genes") tend to get further stabilized in the shock phase, while categories with a low stability before stress (most growth-related genes including cRP and RiBi) tend to get destabilized.

A recent paper (Molina-Navarro et al. 2008) estimated the regulation of transcriptional rate during oxidative stress in *S. cerevisiae* ("genomic run-on"), and from this inferred the regulation at the stability level. The authors demonstrated that regulation of mRNA stability was a wide-spread feature of the response also to that stress condition. The results concerning genes involved in stress response and ribosome biogenesis were well in agreement with our direct measurements, although in our case the switch from adaptation to recovery (the turning point) is more dramatic. This could be because of different stress conditions, of time frames, or as a consequence of the methodology used. Block of RNA polymerase II, as in this paper, provides direct estimates of mRNA degradation rates, while genomic run-on only provides an indirect measure of transcript turnover. The authors express concern that the combination of Phen with stress would not be feasible. However, the apparent specificity of the regulation of stability in response to stress in our experiments, and the elimination from our analysis of genes previously shown to be Phen dependent (Grigull et al. 2004), give confidence that the use of Phen as a transcription inhibitor allows for a sensitive and convenient method of measuring mRNA stability, also in the presence of stress.

Hog1 dependency

Previously it has been shown that the osmostress-induced MAPK Hog1 affects mRNA levels through regulation of transcription. We have shown that Hog1 also regulates the stability of almost all transcripts that are up-regulated at the steady-state level after 30 min, both during the shock and adaptation phases. Hog1 also influences stability of the down-regulated transcripts, although to a lesser degree. The MAPKAP kinase Rck2 was suspected to be involved in stability regulation since the mammalian homolog MAPKAPK-2 is involved in the regulation of interleukin mRNA stability. Surprisingly however, Rck2 seems to influence steady-state levels of stress-responsive transcripts, with only minor effects on stability. This suggests a role for Rck2 in transcriptional regulation rather that stability regulation. It is noteworthy that the transcriptional response was abolished in *rck2Δ* mutants after hyperosmotic shock (this work), but not in oxidative stress (Swaminathan et al. 2006), indicating that the requirement for Rck2 is specific for the hyperosmotic transcriptional response.

Future perspectives

Degradation of mRNA is accomplished through several different mechanisms and regulated through mRNA binding proteins such as the poly A-binding proteins Pab1 and Pub1 and the RNA-binding proteins Puf1-5. mRNA degradation is also tightly coupled to translation, acting as a regulatory mechanism for mRNAs that are not properly translated. Since mammalian p38 acts on mRNA stability through RNA binding proteins like TTP and Hur, it is conceivable that this applies also to Hog1. Transcripts binding to both Pub1 and Puf1-5 have been investigated (Gerber et al. 2004; Duttagupta et al. 2005), but we have not found any obvious connection to our data, except that RiBi mRNAs are bound by both Puf4 and Pub1. If a primary RNA binding protein is the Hog1-targeted effector of mRNA stability regulation, it might be feasible to investigate sequences of, first and foremost, the untranslated regions of the mRNAs affected during stress. Another possibility is that Hog1 primarily affects the translation of the stress-responsive mRNAs and the stability only as a consequence of that. MAP kinases have been shown to regulate translation through phosphorylation of the initiation factor eIF2α and the elongation factor EF2, and might in this manner influence stability. A study of translational regulation of mRNAs following hyperosmotic shock in *S. cerevisiae* has recently been performed (Melamed et al. 2008). However, a preliminary comparison of that data set with ours shows a low degree of overlap in the response profile of steady-state total mRNA (not shown). This indicates that the conditions studied in their work (1 h after exposure to high salinity at 1 M NaCl) are too different from ours to allow a detailed analysis of this issue, especially in view of the strict time dependence revealed in this work.

Studies of mRNA stability provide insight into yet another layer of the complex regulatory network between gene and protein production. Together with investigations of transcription activity and translation, a picture can be shaped of the intricate and interconnected pattern of processes helping the cell to regulate its protein activity in its interaction with the environment.

CONCLUSIONS

Changes in mRNA stability are known to be an important regulatory mechanism in yeast, as well as in mammals. Here we performed global measurements of mRNA stability changes in the course of the response to hyperosmotic stress in *S. cerevisiae*. We show that regulation of stability makes general and substantial contributions to the changes in the steady-state level during transient stress. For certain stress-regulated functional categories, changes in mRNA stability are prominent, whereas for others, steady-state levels are determined solely by transcription rate. The changes at the stability level of stress-induced genes precede and anticipate the changes at the steady-state level, and display a distinct turning point before the peak of the steady-state levels. Finally, the MAP kinase Hog1 is required for most early as well as late changes in stability.

MATERIALS AND METHODS

Strains and growth conditions

S. cerevisiae strains used in the study were from the W303-1A background (wt genotypes: *MAT***a** *ura3-1*, *ade2-1*, *leu2-3,112*, *trp1-1*, *his3-11,15*; mutant genotypes: W303-1A, *hog1::KanMX*; W303-1A, *rck2::KanMX*), and stored in 20% glycerol at −80°C. Cells were precultivated for 24 h in rich medium (YPAD, 1.5% peptone, 1% yeast extract, 2% glucose), inoculated to $A_{595\ nm}$ = 0.05–0.1 in fresh medium (as above) and cultivated for 6–8 h to $A_{595\ nm}$ = 0.7, whereupon NaCl was added to a final concentration of 0.4 M. Ten microliter samples were harvested by centrifugation before (0 min) as well as after 6, 30, and 60 min of NaCl exposure. Samples were divided in two; one of these was washed (twice) and RNA was isolated as below, in the other, transcription was stopped by the addition of the transcriptional inhibitor Phen (Sigma-Aldrich) to a final concentration of 100 ng/mL. No transcriptional inhibition was performed on samples collected 60 min post-NaCl shock. From transcriptionally inhibited samples, 2.5 mL aliquots were harvested by centrifugation and washing (twice) at 5, 15, 30, and 60 min (the 5 min time point was not included in the subsequent analysis due to the high variability of the measurements). For *rck2Δ* mutants, aliquots were harvested only 30 min after Phen addition. Cells were disrupted and RNA extracted using the RiboPure-yeast kit and instructions from the supplier (Applied Biosystems/Ambion).

Microarray hybridization

Labeling of samples and array hybridization were performed essentially as previously described (Bilsland et al. 2007). Briefly, cDNA was synthesized from 15 to 20 μg of RNA using Superscript III reverse transcriptase and an 1:1 mixture of random primers and anchored oligo dT (all from Invitrogen). Samples were labeled with Cy5-dUTP (reference) or Cy3-dUTP (experiment sample) (GE Healthcare). Post-labeling, RNA was hydrolyzed with NaOH and labeled cDNA probes were purified using a CyScribe GFX purification kit (GE Healthcare). Probes were dissolved in DIG Easy hybridization buffer (Roche) and hybridized onto Yeast 6.4K microarray slides (Microarray center, University Health Network) overnight at 42°C. After washing, slides were scanned in a VersArray Chipreader (Bio-Rad).

For quantification of steady-state transcript levels, RNA from NaCl-stressed (experiment) samples were paired with RNA from 0 min, unstressed (reference) samples. For quantification of transcript stabilities, RNA from each Phen time point (experiment) sample was paired with a matching 0 min (before Phen addition) sample (reference).

Microarray data processing

Array spot intensities were quantified using ImaGene v. 6.0.1 (BioDiscovery). Spots flagged as empty, poor, or negative by the ImaGene software were removed. Data analysis was performed using the open source statistical software R, and the LIMMA array analysis package, which is available at the Bioconductor repository (http://www.bioconductor.org/). A loess smoother was applied to remove intensity dependent trends. For each gene on each array, the $\log_{(2)}$-fold change (*M*-value) was calculated and retained for downstream analysis. To ensure comparability within groups of arrays, scale normalization was applied to create the same median-absolute-deviation (MAD) across arrays within a group. The normalization groups were (1) all arrays measuring non-Phen treated pools (steady-state levels); (2) arrays measuring pools within the same Phen time point across strains and stress conditions (i.e., three groups comprising arrays hybridized with mRNA from cells treated with Phen after 15, 30, and 60 min, respectively). Annotations for all open reading frames (ORFs) were taken from the *Saccharomyces* Genome Database (www.yeastgenome.org/; February 2008). All dubious and deleted ORFs were excluded (Supplemental Table S1). In addition, a set of 31 Phen-induced genes previously identified by Grigull et al. (2004) were removed in order to avoid inhibitor specific effects (Supplemental Table S2). Five thousand five hundred and fifty-one genes were retained.

Modeling mRNA stability

To model transcript stability, a simple exponential decay model was adopted, implying that a regression line through the origin can be fit to the $\log_{(2)}$ scale *M*-values for individual mRNAs using the different time points (15, 30, and 60 min after Phen addition). Positive and negative values of the slope of the regression line (denoted k_S and referred to as the stability index) indicate transcripts more stable and unstable than an average transcript. The stability indices range from −0.135 to 0.138 for the least and most stable transcripts, respectively, with median negative and positive values of −0.0045 and 0.0047.

Comparisons of different transcriptional inhibitors in yeast using microarrays have been published previously (Grigull et al. 2004). From this data set, arrays hybridized with Phen-treated pools of mRNA after 12, 30, and 75 min were downloaded (experimental design similar to the design in this study). All arrays were normalized in the same way as described above, and stability indices (k_S) were calculated (only wild-type and unstressed conditions). Mean values for k_S within 38 GO Slim biological process categories (see Functional Annotations) were found to be similar in the two studies (Supplemental Fig. S1; $r^2 = 0.9$).

For the *rck2Δ* mutant, samples were only collected after 30 min of Phen treatment, and a reduced decay model was therefore adapted. With the reduced model, stability is indicated by the $\log_{(2)}$-fold changes (*M*-values) as compared to the corresponding zero time point in all stress conditions. The reduced model was also adapted for the wild-type and *hog1Δ* strains for validation purposes. The full and reduced models in both strains were in good agreement (Supplemental Fig. S2).

Assessing the impact of a stability change after, for example, 30 min can be achieved by multiplying Δk_S with 30. The magnitude of this entity represents a $\log_{(2)}$-fold change contributable to the alteration in stability for a given gene. For example, in Figure 3A (upper and lower panels), the change in stability from 6 to 30 min of stress in the wt contributes to a depletion of half of the transcripts for an average gene at the 60 min time point, as observed in the steady-state levels. This example shows that that the magnitude of fold changes induced by mRNA stability regulation is approximately at par with the changes at the steady-state level. Although the aim of this study was to compare relative changes and not to calculate absolute half-lives, the spread of our stability indices suggests that the average half-life would be less than 13.3 min in the wild type in the unstressed condition [max $k_S(0)$ for wt: 0.074], which is lower than the 23 min reported by Wang et al. (2002). We confirmed the expression levels and calculated stability indices for eight genes using qPCR. The trends were essentially confirmed although changes in k_S values calculated from qPCR data were of higher amplitude than from the array data. Such signal compression is commonly seen when comparing DNA arrays to quantitative methods such as qPCR (Canales et al. 2006; Shi et al. 2006; Arikawa et al. 2008). The signal compression was also apparent at the steady-state level. This complicates the estimation of half-lives, indicating that the spread is, in fact, larger than suggested from the microarrays. Therefore, we use comparisons between stability indices for our investigations of the effect of NaCl stress on mRNA stability instead of absolute half-lives.

Statistical analysis

Differences between estimated coefficients were modeled and ranked using the moderated t-statistic (Smyth 2004) in both decay models and between estimated coefficients in the steady-state level data. Gene-specific variances were estimated using all available experiments.

Correlations between data sets were modeled and tested with the Spearman rank correlation test. The r^2-value, the coefficient of determination, is a measure of explanatory power of the chosen predictive variable.

Differences between mean values for groups of genes were tested with the Welch two-sample *t*-test.

Gene clustering

Clustering was performed using Cluster 3.0 (Stanford University) and an uncentered Pearson correlation metric (average linkage mapping) and visualized using JavaTreeview (University of Tokyo).

Comparisons with previously published data at the steady-state level

Comparisons were made with previously published data on steady-state level changes of mRNA. Lists of osmoregulated genes from five papers were compared (Posas et al. 2000; Rep et al. 2000; Causton et al. 2001; Yale and Bohnert 2001; Krantz et al. 2004), and the 73 genes that were found to be up-regulated in at least three of the papers were selected as "common osmoregulated genes." Sixty-six of these were found on our arrays and compared to our data (Supplemental Table S3).

Quantitative RT-PCR data generation, processing, and analysis

Quantitative RT-PCR was performed to confirm results from the microarray data both on steady-state levels and stability for eight genes: *RIB5* (*YBR256C*), *ECM31* (*YBR176W*), *IMP3* (*YHR148W*), *RPS17A* (*YML024W*), *RPL6B* (*YLR448W*), *HOR2* (*YER062C*), *GRE3* (*YHR104W*), and *GPD1* (*YDL022W*). Steady-state levels (t_{TOT}) for all genes were examined 0, 6, 30, and 60 min after the addition of 0.4 M NaCl in wild type and the *hog1Δ* and *rck2Δ* mutants (Supplemental Fig. S3). All genes were also examined on the stability level with the same design as in the microarray experiment (Figs. 5C, 6C). cDNA synthesis was performed using 1.1–4.5 μg of total RNA. Random primers and Superscript III reverse transcriptase (Invitrogen) were used. Primers were designed using the Primer express 2.0 software (Applied Biosystems) using standard settings. RT-PCR reaction was performed using SYBR Green detection in the Göteborg genomics core facility (Swegene) in an ABI PRISM 7900HT Sequence Detection system (Applied Biosystems). Two biological and three technical replicates were used for each sample.

Both *RIB5* and *ECM31* were included as potential reference genes, and since *RIB5* proved to be stable over all conditions with the smallest variance, all signals were normalized against this gene. For each Phen time series, ratios were computed to the corresponding Phen untreated pool (e.g., unstressed wt cells harvested after 15 min of Phen against unstressed wt cells without Phen treatment). Both the reduced and full decay models were adapted to the data as in the microarray experiment. For the steady-state level data, expression data were modeled similarly as in the microarray experiment in order to make direct comparisons. Spearman rank correlation between array and qPCR data: 0.9 using data for *rck2Δ* and transformed stability index data for wt and *hog1Δ*. Fitting a regression line to the measurements with the qPCR data as the independent variable will produce a small intercept, indicating comparable normalization procedures between the two sets.

Data availability at ArrayExpress

The microarray data from this study are available at the ArrayExpress repository with accession number XY-123 (E-TABM-622).

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

This work was supported by a grant from the Swedish Research Council (2007-5460) to P.S. C.M. was supported by the National Graduate School of Genomics and Bioinformatics. A.J. and O.N. received support from the Swedish Research Council, the Gothenburg Stochastic Centre, Gothenburg Mathematical Modeling Centre, and the Swedish Foundation for Strategic Research.

Received October 6, 2008; accepted December 19, 2008.

REFERENCES

Albertyn, J., Hohmann, S., Thevelein, J.M., and Prior, B.A. 1994. *GPD1*, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in *Saccharomyces cerevisiae*, and its expression is regulated by the high-osmolarity glycerol response pathway. *Mol. Cell. Biol.* **14:** 4135–4144.

Arikawa, E., Sun, Y., Wang, J., Zhou, Q., Ning, B., Dial, S.L., Guo, L., and Yang, J. 2008. Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC Genomics* **9:** 328. doi: 10.1186/1471-2164-9-328.

Asp, E., Nilsson, D., and Sunnerhagen, P. 2008. Fission yeast mitogen-activated protein kinase Sty1 interacts with translation factors. *Eukaryot. Cell* **7:** 328–338.

Bilsland, E., Molin, C., Swaminathan, S., Ramne, A., and Sunnerhagen, P. 2004. Rck1 and Rck2 MAPKAP kinases and the HOG pathway are required for oxidative stress resistance. *Mol. Microbiol.* **53:** 1743–1756.

Bilsland, E., Hult, M., Bell, S.D., Sunnerhagen, P., and Downs, J.A. 2007. The Bre5/Ubp3 ubiquitin protease complex from budding yeast contributes to the cellular response to DNA damage. *DNA Repair (Amst.)* **6:** 1471–1484.

Bilsland-Marchesan, E., Ariño, J., Saito, H., Sunnerhagen, P., and Posas, F. 2000. Rck2 kinase is a substrate for the osmotic stress activated mitogen-activated protein kinase Hog1. *Mol. Cell. Biol.* **20:** 3887–3895.

Brewster, J.L., de Valoir, T., Dwyer, N.D., Winter, E., and Gustin, M.C. 1993. An osmosensing signal transduction pathway in yeast. *Science* **259:** 1760–1763.

Brown, A.J.P. 1994. Measurement of mRNA stability. In *Molecular genetics of yeast: A practical approach* (ed. J.R. Johnston), pp. 147–160. Oxford University Press, Oxford, UK.

Canales, R.D., Luo, Y., Willey, J.C., Austermiller, B., Barbacioru, C.C., Boysen, C., Hunkapiller, K., Jensen, R.V., Knight, C.R., Lee, K.Y., et al. 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24:** 1115–1122.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. 2001. Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12:** 323–337.

Chen, C.Y., Gherzi, R., Andersen, J.S., Gaietta, G., Jurchott, K., Royer, H.D., Mann, M., and Karin, M. 2000. Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. *Genes & Dev.* **14:** 1236–1248.

Clotet, J., Escote, X., Adrover, M.A., Yaakov, G., Garí, E., Aldea, M., de Nadal, E., and Posas, F. 2006. Phosphorylation of Hsl1 by Hog1 leads to a G_2 arrest essential for cell survival at high osmolarity. *EMBO J.* **25:** 2338–2346.

de Nadal, E., Casadome, L., and Posas, F. 2003. Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. *Mol. Cell. Biol.* **23:** 229–237.

de Nadal, E., Zapater, M., Alepuz, P.M., Sumoy, L., Mas, G., and Posas, F. 2004. The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. *Nature* **427:** 370–374.

Dunand-Sauthier, I., Walker, C.A., Narasimhan, J., Pearce, A.K., Wek, R.C., and Humphrey, T.C. 2005. Stress-activated protein kinase pathway functions to support protein synthesis and translational adaptation in response to environmental stress in fission yeast. *Eukaryot. Cell* **4:** 1785–1793.

Duttagupta, R., Tian, B., Wilusz, C.J., Khounh, D.T., Soteropoulos, P., Ouyang, M., Dougherty, J.P., and Peltz, S.W. 2005. Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. *Mol. Cell. Biol.* **25:** 5499–5513.

Escote, X., Zapater, M., Clotet, J., and Posas, F. 2004. Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nat. Cell Biol.* **6:** 997–1002.

Garcia-Martinez, J., Aranda, A., and Pérez-Ortín, J.E. 2004. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol. Cell* **15:** 303–313.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11:** 4241–4257.

Gerber, A.P., Herschlag, D., and Brown, P.O. 2004. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* **2:** E79. doi: 10.1371/journal.pbio.0020079.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M.D., and Hughes, T.R. 2004. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol. Cell. Biol.* **24:** 5534–5547.

Hitti, E., Iakovleva, T., Brook, M., Deppenmeier, S., Gruber, A.D., Radzioch, D., Clark, A.R., Blackshear, P.J., Kotlyarov, A., and Gaestel, M. 2006. Mitogen-activated protein kinase-activated protein kinase 2 regulates tumor necrosis factor mRNA stability and translation mainly by altering tristetraprolin expression, stability, and binding to adenine/uridine-rich element. *Mol. Cell. Biol.* **26:** 2399–2407.

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

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., and Barkai, N. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309:** 938–940.

Jorgensen, P., Rupes, I., Sharom, J.R., Schneper, L., Broach, J.R., and Tyers, M. 2004. A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes & Dev.* **18:** 2491–2505.

Kontoyiannis, D., Kotlyarov, A., Carballo, E., Alexopoulou, L., Blackshear, P.J., Gaestel, M., Davis, R., Flavell, R., and Kollias, G. 2001. Interleukin-10 targets p38 MAPK to modulate ARE-dependent TNF mRNA translation and limit intestinal pathology. *EMBO J.* **20:** 3760–3770.

Krantz, M., Nordlander, B., Valadi, H., Johansson, M., Gustafsson, L., and Hohmann, S. 2004. Anaerobicity prepares *Saccharomyces cerevisiae* cells for faster adaptation to osmotic shock. *Eukaryot. Cell* **3:** 1381–1390.

Lackner, D.H., Beilharz, T.H., Marguerat, S., Mata, J., Watt, S., Schubert, F., Preiss, T., and Bähler, J. 2007. A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol. Cell* **26:** 145–155.

MACQ Consortium. 2006. The MicroArray Quality Control (MAQC) Project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24:** 1151–1161.

Maeda, T., Wurgler-Murphy, S.M., and Saito, H. 1994. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* **369:** 242–245.

Mager, W.H. and Varela, J.C. 1993. Osmostress response of the yeast *Saccharomyces*. *Mol. Microbiol.* **10:** 253–258.

Marquez, J.A. and Serrano, R. 1996. Multiple transduction pathways regulate the sodium-extrusion gene *PMR2/ENA1* during salt stress in yeast. *FEBS Lett.* **382:** 89–92.

Melamed, D., Pnueli, L., and Arava, Y. 2008. Yeast translational response to high salinity: Global analysis reveals regulation at multiple levels. *RNA* **14:** 1337–1351.

Molina-Navarro, M.M., Castells-Roca, L., Bellí, G., García-Martínez, J., Marín-Navarro, J., Moreno, J., Pérez-Ortín, J.E., and Herrero, E. 2008. Comprehensive transcriptional analysis of the oxidative response in yeast. *J. Biol. Chem.* **283:** 17908–17918.

Norbeck, J. and Blomberg, A. 1998. Amino acid uptake is strongly affected during exponential growth of *Saccharomyces cerevisiae* in 0.7 M NaCl medium. *FEMS Microbiol. Lett.* **158:** 121–126.

O'Rourke, S.M. and Herskowitz, I. 1998. The Hog1 MAPK prevents cross talk between the HOG and pheromone response MAPK pathways in *Saccharomyces cerevisiae*. *Genes & Dev.* **12:** 2874–2886.

Pandey, G., Yoshikawa, K., Hirasawa, T., Nagahisa, K., Katakura, Y., Furusawa, C., Shimizu, H., and Shioya, S. 2007. Extracting the hidden features in saline osmotic tolerance in *Saccharomyces cerevisiae* from DNA microarray data using the self-organizing map: Biosynthesis of amino acids. *Appl. Microbiol. Biotechnol.* **75:** 415–426.

Posas, F., Chambers, J.R., Heyman, J.A., Hoeffler, J.P., de Nadal, E., and Ariño, J. 2000. The transcriptional response of yeast to saline stress. *J. Biol. Chem.* **275:** 17249–17255.

Proft, M. and Struhl, K. 2004. MAP kinase-mediated stress relief that precedes and regulates the timing of transcriptional induction. *Cell* **118:** 351–361.

Proft, M., Pascual-Ahuir, A., de Nadal, E., Ariño, J., Serrano, R., and Posas, F. 2001. Regulation of the Sko1 transcriptional repressor by the Hog1 MAP kinase in response to osmotic stress. *EMBO J.* **20:** 1123–1133.

Puig, S., Askeland, E., and Thiele, D.J. 2005. Coordinated remodeling of cellular metabolism during iron deficiency through targeted mRNA degradation. *Cell* **120:** 99–110.

Rep, M., Reiser, V., Gartner, U., Thevelein, J.M., Hohmann, S., Ammerer, G., and Ruis, H. 1999. Osmotic stress-induced gene expression in *Saccharomyces cerevisiae* requires Msn1p and the novel nuclear factor Hot1p. *Mol. Cell. Biol.* **19:** 5474–5485.

Rep, M., Krantz, M., Thevelein, J.M., and Hohmann, S. 2000. The transcriptional response of *Saccharomyces cerevisiae* to osmotic shock. Hot1p and Msn2p/Msn4p are required for the induction of subsets of high osmolarity glycerol pathway-dependent genes. *J. Biol. Chem.* **275:** 8290–8300.

Rodriguez-Gabriel, M.A., Burns, G., McDonald, W.H., Martin, V., Yates 3rd, J.R., Bähler, J., and Russell, P. 2003. RNA-binding protein Csx1 mediates global control of gene expression in response to oxidative stress. *EMBO J.* **22:** 6256–6266.

Schüller, C., Brewster, J.L., Alexander, M.R., Gustin, M.C., and Ruis, H. 1994. The HOG pathway controls osmotic regulation of transcription via the stress response element (STRE) of the *Saccharomyces cerevisiae* CTT1 gene. *EMBO J.* **13:** 4382–4389.

Shalem, O., Dahan, O., Levo, M., Martinez, M.R., Furman, I., Segal, E., and Pilpel, Y. 2008. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol. Syst. Biol.* **4:** 223. doi: 10.1037/msb.2008.59.

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3:** Article 3. http://www.bepress.com/sagmb/vol3/iss1/art3.

Swaminathan, S., Masek, T., Molin, C., Pospisek, M., and Sunnerhagen, P. 2006. Rck2 is required for reprogramming of ribosomes during oxidative stress. *Mol. Biol. Cell* **17:** 1472–1482.

Teige, M., Scheikl, E., Reiser, V., Ruis, H., and Ammerer, G. 2001. Rck2, a member of the calmodulin-protein kinase family, links protein synthesis to high osmolarity MAP kinase signaling in budding yeast. *Proc. Natl. Acad. Sci.* **98:** 5625–5630.

Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. 2002. Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci.* **99:** 5860–5865.

Warner, J.R. 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24:** 437–440.

Winzen, R., Kracht, M., Ritter, B., Wilhelm, A., Chen, C.Y., Shyu, A.B., Müller, M., Gaestel, M., Resch, K., and Holtmann, H. 1999. The p38 MAP kinase pathway signals for cytokine-induced mRNA stabilization via MAP kinase-activated protein kinase 2 and an AU-rich region-targeted mechanism. *EMBO J.* **18:** 4969–4980.

Yale, J. and Bohnert, H.J. 2001. Transcript expression in *Saccharomyces cerevisiae* at high salinity. *J. Biol. Chem.* **276:** 15996–16007.

Additional files

Additional file 1 – Supplementary figures.

Figure S1 - Comparison of stability data with previously published data

Stability data from untreated wt cells (this paper) were compared to previously published stability data (Grigull et al. 2004). The full mRNA decay model was applied to the raw data from Grigull *et al.* and means for all GO Slim-categories were compared to corresponding data from this study.

Figure S2 - Comparison between the full and the reduced mRNA decay models

The reduced decay model was compared to the full decay model for all stress conditions in the wild-type and *hog1Δ* mutant. In the reduced decay model, the log-fold change between 0 and 30 min. after Phen addition represents relative stability. The k-values from the full decay model were multiplied by 30 to emulate the relative difference after 30 min. of Phen treatment. As a reference, a line (red) is drawn through the origin with slope = 1. The correlation between the reduced and the full model was good overall, although extreme values tend to get more pronounced in the reduced model.

Figure S3 – RT-PCR verification of steady state level changes

RT-PCR was performed to verify expression profiles for the same six genes as shown in Figs. 5 and 6. Log-fold changes were calculated between each stress condition and the unstressed condition within each strain for comparison to array data. *GRE3*, *HOR2* and *GPD1* represent genes upregulated at the steady state level after 30 min. of stress. *RPS17A*, *RPL6B* and *IMP3* represent genes downregulated at this time point. One array data point is missing (*GPD1* in *hog1Δ* after 30 min. of stress). Spearman rank correlation between array and qPCR data: 0.85 (see Materials and Methods).

Figure S4 – Clusterings of the GO Slim categories *Carbohydrate metabolism*, *Amino acid metabolism* and *Protein catabolism*

The GO Slim functional categories *Carbohydrate metabolism* (**A**), *Amino acid metabolism* (**B**) and *Proteasomal catabolism* (**C**) were clustered (uncentered Pearson correlation metric) with respect to the wt. Genes with more than one missing value were omitted. Means of the changes during the phases of the stress response were calculated for the "turning point" clusters and compared to the rest of the genes in the category in wt, *hog1Δ* and *rck2Δ* cells. Error bars denote 95 % confidence intervals.

Additional file 2 – Supplementary tables.

Table S1 contains the dubious or deleted ORFs that were removed from the gene set analyzed.

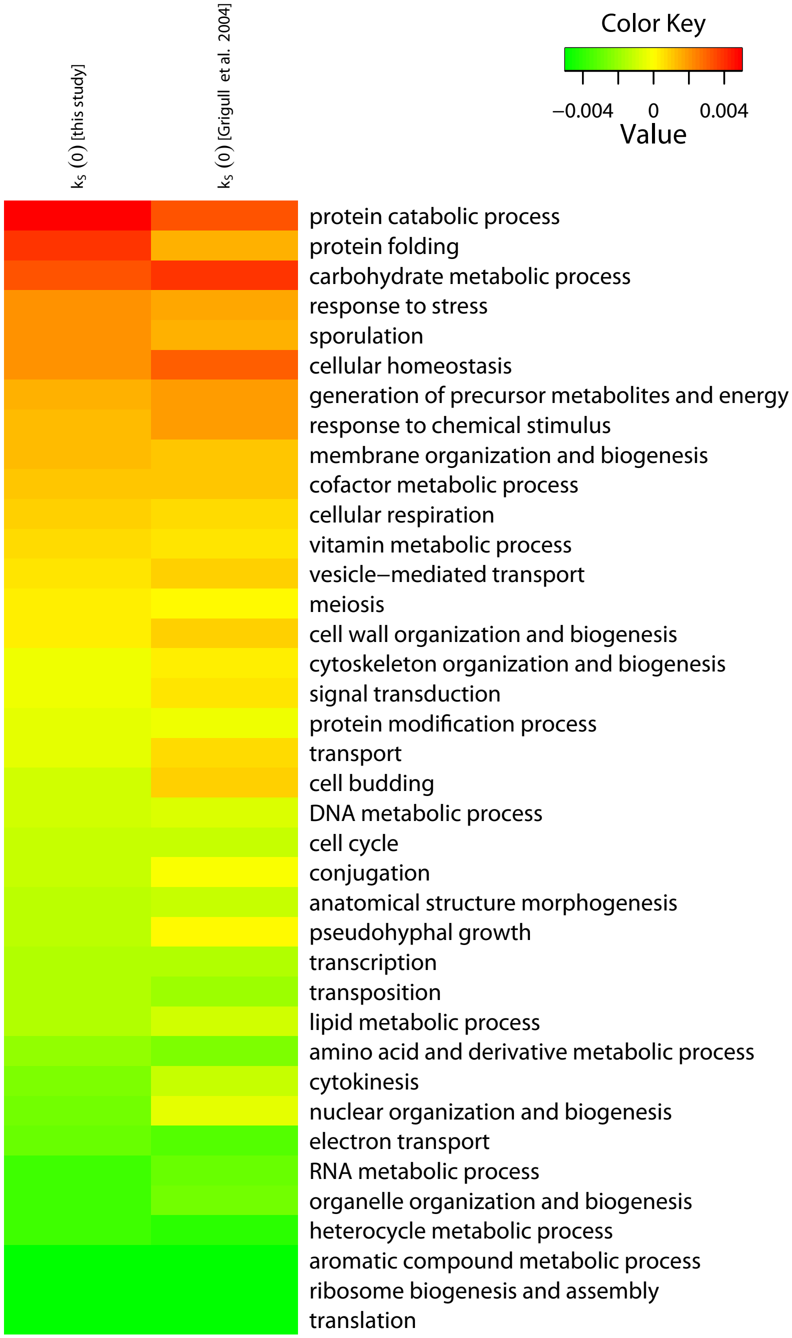
Table S2 contains the ORFs whose expression has been shown to be affected by Phen and hence were removed from the gene set.

Table S3 contains a list of 66 osmo-regulated genes identified from previous studies.

Table S4 contains lists of genes up- and downregulated at $\Delta t_{TOT}(30)$, respectively.

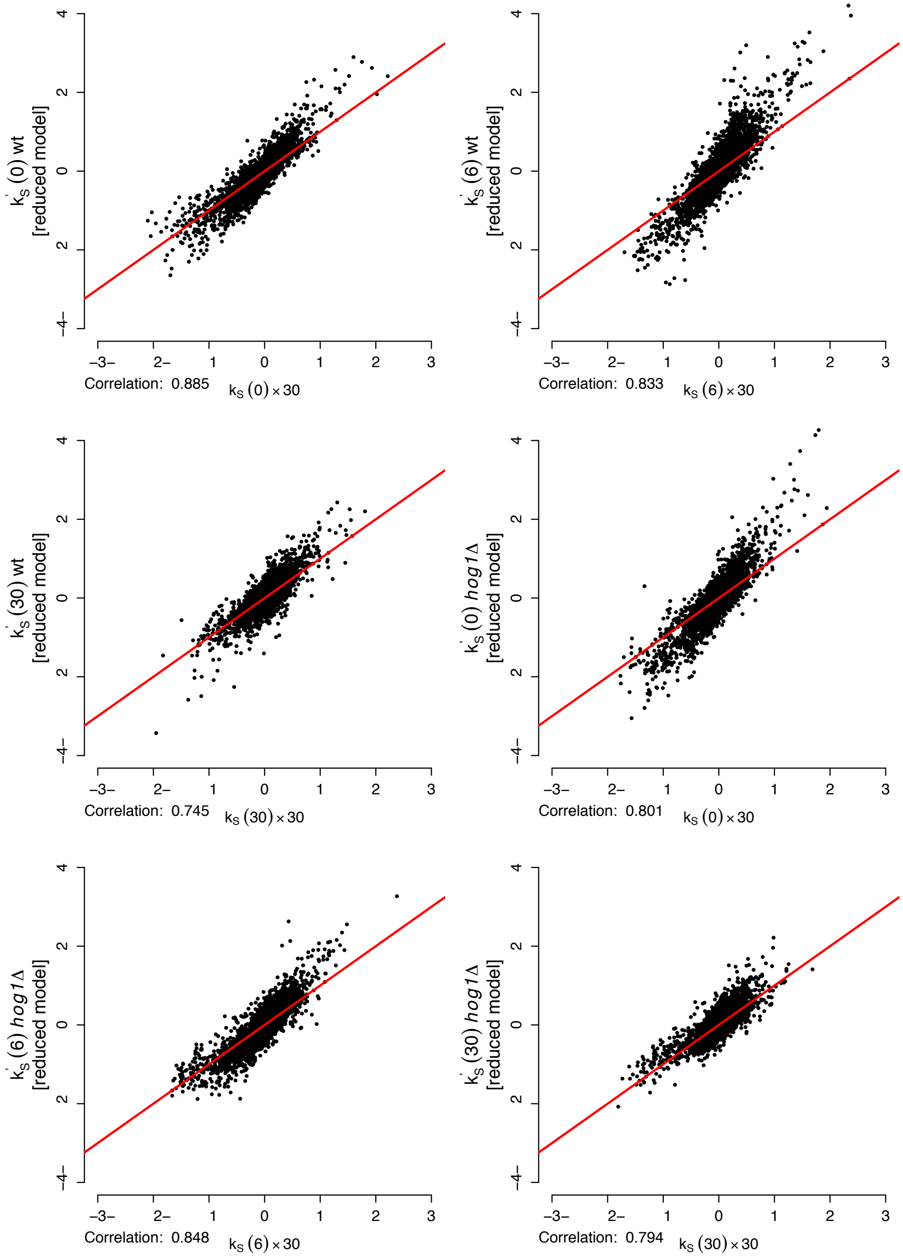
Table S5 contains lists of the genes belonging to the different clusters of the GO Slim categories from Figs. 5, 6 and S4, along with the "turning point" cluster from a clustering of all 5551 genes.

Figure S1



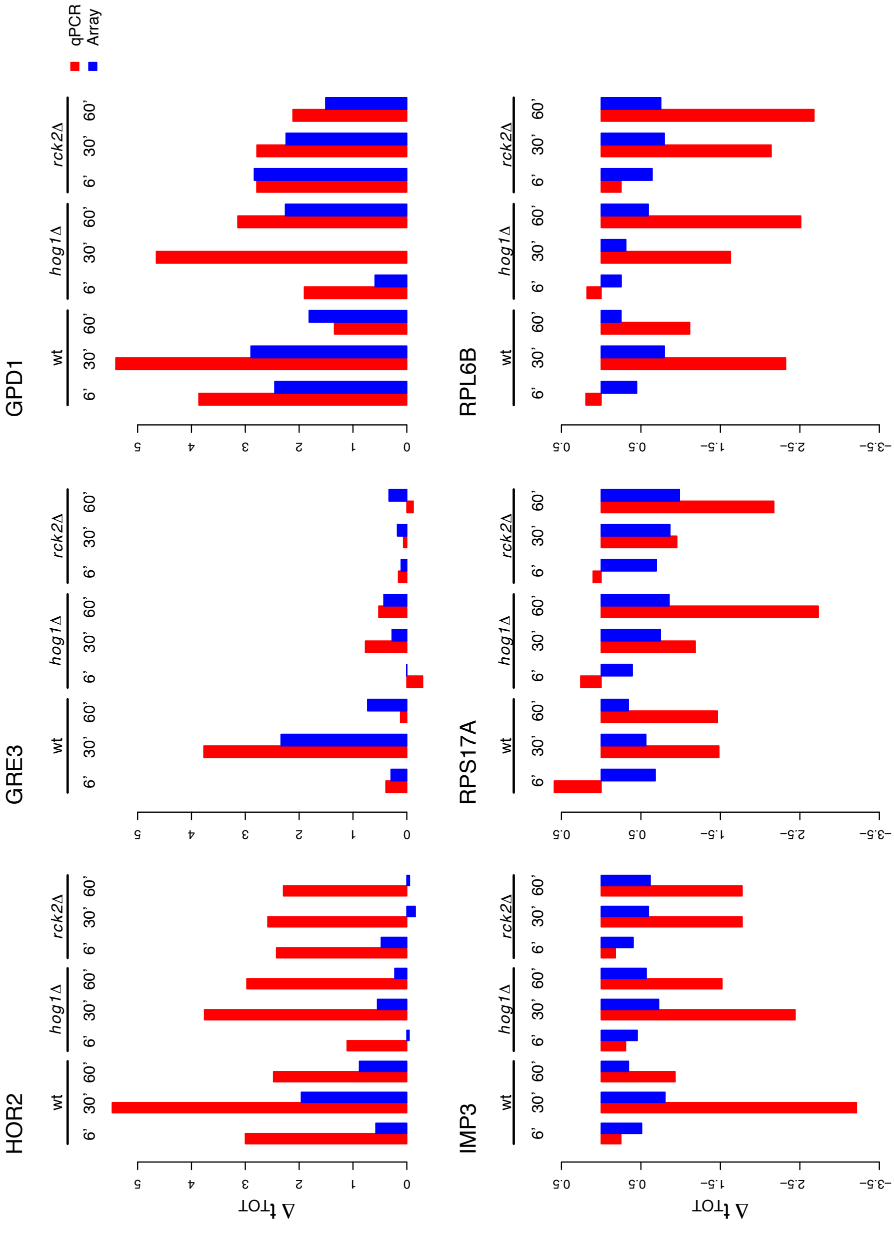
Color Key
−0.004
0
0.004
Value
k_S (0) [this study]
k_S (0) [Grigull et al. 2004]
protein catabolic process
protein folding
carbohydrate metabolic process
response to stress
sporulation
cellular homeostasis
generation of precursor metabolites and energy
response to chemical stimulus
membrane organization and biogenesis
cofactor metabolic process
cellular respiration
vitamin metabolic process
vesicle–mediated transport
meiosis
cell wall organization and biogenesis
cytoskeleton organization and biogenesis
signal transduction
protein modification process
transport
cell budding
DNA metabolic process
cell cycle
conjugation
anatomical structure morphogenesis
pseudohyphal growth
transcription
transposition
lipid metabolic process
amino acid and derivative metabolic process
cytokinesis
nuclear organization and biogenesis
electron transport
RNA metabolic process
organelle organization and biogenesis
heterocycle metabolic process
aromatic compound metabolic process
ribosome biogenesis and assembly
translation

Figure S2



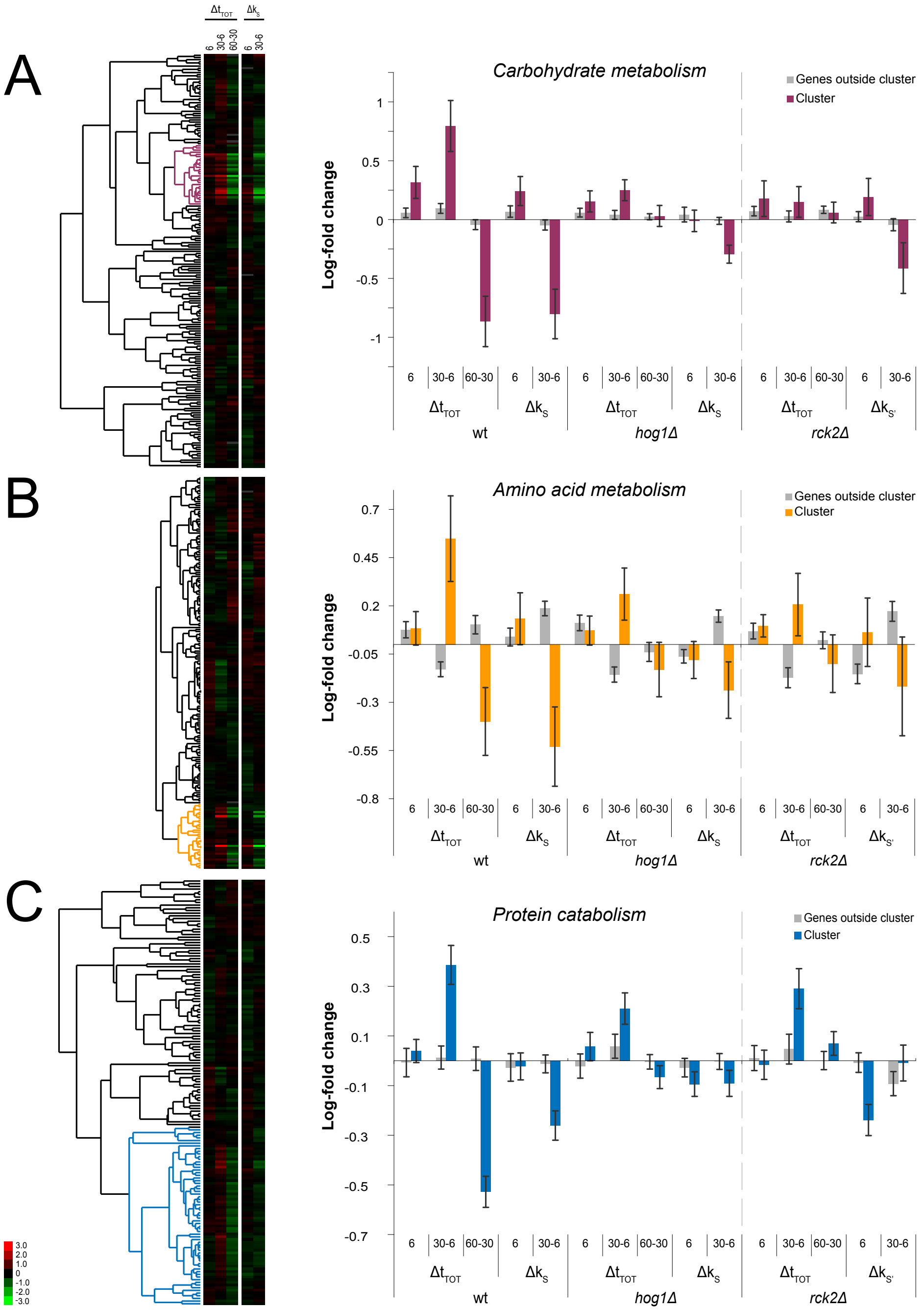
k'_S (0) wt
[reduced model]
4
2
0
2
−4
−3–
2–
1
0
1
2
3
Correlation: 0.885
k_S (0) × 30
k'_S (6) wt
[reduced model]
Correlation: 0.833
k_S (6) × 30
k'_S (30) wt
[reduced model]
Correlation: 0.745
k_S (30) × 30
k'_S (0) hog1Δ
[reduced model]
Correlation: 0.801
k_S (0) × 30
k'_S (6) hog1Δ
[reduced model]
Correlation: 0.848
k_S (6) × 30
k'_S (30) hog1Δ
[reduced model]
Correlation: 0.794
k_S (30) × 30

Figure S3



HOR2
GRE3
GPD1
IMP3
RPS17A
RPL6B
wt
hog1Δ
rck2Δ
6'
30'
60'
🟥 qPCR
🟦 Array
Δ tOT
0
1
2
3
4
5
0.5
-0.5
-1.5
-2.5
-3.5

Figure S4


Δt_{TOT}
Δk_S
6
30-6
60-30
6
30-6
A
Carbohydrate metabolism
Genes outside cluster
Cluster
1
0.5
0
-0.5
-1
Log-fold change
6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6
Δt_{TOT} Δk_S Δt_{TOT} Δk_S Δt_{TOT} $\Delta k_{S'}$
wt
hog1Δ
rck2Δ
B
Amino acid metabolism
Genes outside cluster
Cluster
0.7
0.45
0.2
-0.05
-0.3
-0.55
-0.8
Log-fold change
6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6
Δt_{TOT} Δk_S Δt_{TOT} Δk_S Δt_{TOT} $\Delta k_{S'}$
wt
hog1Δ
rck2Δ
C
Protein catabolism
Genes outside cluster
Cluster
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.7
Log-fold change
6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6 | 6 | 30-6 | 60-30 | 6 | 30-6
Δt_{TOT} Δk_S Δt_{TOT} Δk_S Δt_{TOT} $\Delta k_{S'}$
wt
hog1Δ
rck2Δ
3.0
2.0
1.0
0
-1.0
-2.0
-3.0

Paper III

Subcellular localization and effects of DDIT3/GADD153/CHOP

Running title: Effects of nuclear and cytoplasmic DDIT3

Alexandra Jauhiainen[1,2], Christer Thomsen[3], Linda Strömbom[3], Pernilla Grundevik[3], Carola Andersson[3], Anna Danielsson[4], Mattias K Andersson[3], Olle Nerman[1,2], Linda Rörkvist[3], Anders Ståhlberg[3] and Pierre Åman[3§].

[1] Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden.
[2] Department of Mathematical Statistics, University of Gothenburg, Göteborg, Sweden.
[3] Lundberg Laboratory for Cancer Research, Department of Pathology, Sahlgrenska Academy at the University of Gothenburg, Göteborg, Sweden.
[4] Department of Oncology, Institute of Clinical Sciences, University of Gothenburg, Göteborg, Sweden

[§] Corresponding author
Lundberg Laboratory for Cancer Research, Department of Pathology, Sahlgrenska Academy at the University of Gothenburg, SE- 413 45 Göteborg, Sweden
telephone: +46 31-342 2842
fax: +46 31-828733
e-mail: pierre.aman@llcr.med.gu.se

Abstract

DDIT3, also known as GADD153 or CHOP, encodes a key regulator of cellular stress response. DDIT3 is a basic leucine zipper transcription factor of the dimer forming C/EBP family. Originally described as a nuclear, dominant negative factor, it also binds DNA as heterodimers and induces transcription of target genes. The aim of the present investigation was to study the subcellular localization of DDIT3 and to identify target genes and functions regulated by cytoplasmic and nuclear DDIT3. Employing microarray, RT-qPCR, and immuno-techniques, we analyzed cultured normal human fibroblasts and sarcoma cells carrying amplified *DDIT3* or tamoxifen inducible DDIT3 expression constructs. We report that DDIT3 is localized both in the cytoplasmic and nuclear compartments. Cytoplasmic and nuclear DDIT3 regulated 94 and 84 genes respectively with only three genes overlapping. Functional annotation showed cell migration, proliferation and apoptosis/survival as the most affected categories. Cytoplasmic DDIT3 gave a migration inhibitory effect whereas nuclear DDIT3 induced a G1 arrest. No common DNA sequence motifs were identified in promoters of DDIT3 regulated genes. This may be explained by divergent sequence specificity of DDIT3 in heterodimers with different partners. We conclude that stress induced DDIT3 can be expressed both in the cytoplasm and the nucleus and that different genes are regulated by DDIT3 in these compartments.

Introduction

DDIT3 (DNA damage induced transcript 3) also known as GADD153 (G1 arrest and DNA damage 153) or CHOP (C/EBP homologous protein), encodes a key regulator of stress response. DNA damage, ER stress, hypoxia, and starvation induce *DDIT3* transcription and release translation of the *DDIT3* encoding reading frame, resulting in DDIT3 protein accumulation (1-8). Forced expression of *DDIT3* triggers cell cycle arrest and apoptosis, indicating a central role of *DDIT3* in these stress effects (9-10). *DDIT3* has recently been implicated in stress responses leading to death of pancreatic insulin producing β-cells and in neurodegenerative disorders (11-13). *DDIT3* is also involved in differentiation of specialized tissues and cells (14-15).

The *DDIT3* encoded protein is a basic leucine zipper transcription factor of the dimer forming C/EBP family (16-17). Unlike other C/EBP transcription factors, DDIT3 does not form homodimers but acts as a dominant negative factor that blocks the activities of its C/EBP dimer partners. However, it may also bind DNA as a heterodimer with other basic zipper transcription factors and induce transcription of down-stream target genes (18).

DDIT3 has been considered to be a nuclear transcription factor but recently it was reported to be expressed as a cytoplasmic protein in erythroid leukemia and kidney proximal tubular epithelial cells (19-20). The cytoplasmic localization and the intrinsically disordered domain that mediates binding and interactions with proteins other than leucine zippers suggest that DDIT3 may have additional functions (21).These observations raise the questions whether different types of stress could induce controlled localization to the cytoplasmic or nuclear compartments and if nuclear and cytoplasmic DDIT3 trigger distinct responses.

To address these questions we studied the effects of ER and genotoxic stress in cultured normal human fibroblasts and in a liposarcoma cell line carrying a genetically amplified and constitutively expressed *DDIT3* gene. The effects of cytoplasmic and nuclear DDIT3 protein was further studied in stably transfected cell lines that express high levels of DDIT3 fused to the ligand binding parts of the mouse estrogen receptor (mor) and tagged with the enhanced green fluorescence protein (EGFP). The mor part retains the recombinant protein in the cytoplasm and addition of tamoxifen induces a rapid translocation of the recombinant DDIT3 from cytoplasm to the nuclei of stably transfected cells. Employing microarray, RT-qPCR, immunofluorescence, and western blot methods we identified DDIT3 regulated downstream target genes. Our analysis shows that the cellular localization of DDIT3 is determined by the inducing stress conditions and that cytoplasmic and nuclear DDIT3 induce distinct gene expression patterns indicating separate roles in stress response.

Results

Cytoplasmic accumulation of DDIT3 in stressed human fibroblasts and GOT3 liposarcoma cells

Native DDIT3 was previously reported to be expressed as a nuclear transcription factor (17). Here we show that DDIT3 accumulates as a cytoplasmic protein in tunicamycin and etoposide stressed human cultured serum deprived fibroblasts and GOT3 liposarcoma cells (Figure 1a). In contrast, transfection of GOT3 cells or normal fibroblasts with a DDIT3-EGFP expression vector (not containing the mouse estrogen receptor part), resulted in a nuclear localization of the recombinant protein (data not shown). To evaluate if stress conditions caused by the transfection procedure lead to nuclear DDIT3 expression, we transfected GOT3 cells with plasmid DNA and stained for the endogenous DDIT3.

Tamoxifen induces nuclear translocation of DDIT3morEGFP

Routine cultured HT1080 cells containing the p*DDIT3*mor-EGFP or pmorEGFP constructs showed a cytoplasmic expression of the recombinant proteins (Figure 1c). The DDIT3morEGFP recombinant protein translocated within 30 minutes to the nuclei (Figure 1b and Supplementary Film Clip 1) in tamoxifen treated cells whereas morEGFP showed no changed subcellular distribution in this time span (Supplementary Film Clip 2). Four or eight hours pre-inhibition of P38 kinase and casein kinase 2, known to phosphorylate DDIT3 at four different positions, gave no effects on the nuclear translocation of DDIT3morEGFP (data not shown).

Genes, functions, and networks regulated downstream of cytoplasmic DDIT3morEGFP

Microarray experiments were made in two biological replicates and in two p*DDIT3*morEGFP-transfected and two morEGFP-transfected clones (Supplementary Figure 3). Analysis of the microarray data from cells expressing cytoplasmic DDIT3morEGFP revealed 94 genes that were regulated at least three-fold compared to their expression in morEGFP expressing cells (Supplementary Table 1a). Of the 94 genes, 33 genes were up regulated and 61 down regulated.

Functional analysis showed that the functional categories *cellular movement, cell death, cellular development,* and *cellular growth and proliferation* were the most significantly affected (Figure 2), with some overlap between the genes annotated to the different categories (Supplementary Figure 1a).

The 94 regulated genes were also matched to the gene/protein network constructed by Ingenuity Pathway Knowledge Base in attempts to identify affected networks and network hubs. Hubs are genes/proteins connecting several paths of the network. The most significant network associated with cytoplasmic DDIT3 did not depict any of the regulated genes as major hubs (Supplementary Figure 2a).

Cytoplasmic DDIT3 inhibits migration of HT1080 cells

In total 20 of 94 regulated genes were annotated to the functional category *cellular movement* (Supplementary Table 2a). For example, *DSTN*, an actin-depolymerizing factor important for remodeling of the cytoskeleton was down regulated and *ATF3*, which has been shown to block migration, was up regulated. Several extracellular matrix related genes such as *FN1,* fibronectin, *HAS2*, hyaluronan synthetase, and *CSPG4*, chondroitin sulfate proteoglycan 4, were downregulated. The adhesion molecule cadherin 11was also down regulated. Taken together, the cytoplasmic DDIT3 regulation of movement-associated genes suggested a negative effect on cellular movement/migration.

This suggestion prompted an experimental analysis. Using a modified scratch wound migration assay we confirm that DDIT3morEGFP and DDIT3-EGFP expressing HT1080 cells had an impaired migration capacity compared to wild type and EGFP expressing HT1080 cells (Figure 3).

Genes, functions, and networks regulated by nuclear DDIT3morEGFP

To study direct effects of nuclear DDIT3, microarray analysis was performed two and eight hours after tamoxifen activation and cytoplasmic release of DDIT3 (for experiment overview, see Supplementary Figure 3). As only two or eight hours were allowed for accumulation or degradation of DDIT3 regulated mRNAs, a lower differential expression threshold was chosen (compared to the threshold used for genes regulated by cytoplasmic DDIT3). The analysis showed that 45 genes were up or down regulated at least two-fold compared to expression in cells with cytoplasmic DDIT3 (Figure 4a, Supplementary Table 1b). At eight hours after DDIT3 activation, several initially regulated genes were repressed back to their initial expression levels, but replaced by other response genes. In total 52 genes were regulated at eight hours after DDIT3 activation but only 13 of these genes were regulated after two hours (Supplementary Table 1b). Most of the regulated genes were repressed compared to the control, supporting the hypothesis that DDIT3 acts as a dominant negative factor (17).

To confirm the accuracy of the microarray experiments, 32 genes were selected for RT-qPCR analysis. The results confirmed small changes, in most cases down regulation of selected mRNAs at two and eight hours after tamoxifen activation of DDIT3 (Supplementary Table 1c).

Functional analysis among the two-hour response genes for nuclear DDIT3 showed that the top enriched categories were *cell death*, *cellular development*, *cellular growth and proliferation*, and *cell cycle* (Figure 2), with large overlaps of genes between the categories (Supplementary Figure 1b). Within each category, affected subgroups were for example *apoptosis* ($p = 1 \times 10^{-6}$), *development of cells* ($p = 9 \times 10^{-4}$), *growth of cells* ($p = 3 \times 10^{-5}$), and *interphase* ($p = 8 \times 10^{-6}$) (Supplementary Table 2b).

Cell death, *gene expression*, *cellular development*, and *cell cycle* were the most significant categories enriched among the eight-hour response genes of nuclear DDIT3, however all subgroups within *cellular development* contained only a small number of genes (Figure 2, Supplementary Table 2c). Genes of all categories overlapped, but not to same extent as for the categories in the two-hour time point (Supplementary Figure 1c). Within the *cell death, gene expression*, and *cell cycle* categories, enriched groups were for example *apoptosis* ($p = 1 \times 10^{-5}$), *transcription* ($p = 2 \times 10^{-5}$), and *M phase of eukaryotic* cells ($p = 4 \times 10^{-4}$), (Supplementary Table 2c).

Network analysis based on the regulated genes from two and eight hours after nuclear transition of DDIT3 (Supplementary Figure 2b-c) show that *EGR1* may work as a hub connecting several paths in the network. No overrepresentation of the EGR1 binding site could be found among the group of regulated genes, either at two or eight hours, suggesting that the downstream effects of EGR1 may appear at a time point later than eight hours after DDIT3 nuclear translocation.

Cytoplasmic and nuclear DDIT3 regulate different genes and functions

Most genes affected by cytoplasmic DDIT3 remained at their initial levels after the nuclear translocation of DDIT3 (Supplementary Figure 4). Exceptions were *ATF3* and *HSPA1A*, which were up regulated, and *TPO*, which was down regulated in cells with cytoplasmic DDIT3. Thus, tamoxifen induced nuclear translocation of DDIT3 partially or completely reduced the cytoplasmic DDIT3 effect on the expression of these three genes (Figure 4b). In addition to the genes and functions regulated by cytoplasmic DDIT3 (Figure 2, open bars) additional genes became regulated at two and eight hours after nuclear translocation (Figure 2, striped and solid bars, respectively). Thus, nuclear DDIT3 regulated new genes belonging to functional categories that were already significantly enriched among the genes regulated by cytoplasmic DDIT3. However, for the *cellular movement* category, the enrichment is notably more significant, suggesting that mainly cytoplasmic DDIT3 is involved in regulation of migration.

Tamoxifen activation of DDIT3morEGFP leads to a transient G1 arrest

Functional analysis of our microarray results suggested that DDIT3 regulated genes were involved in cell cycle control and apoptosis. Tamoxifen activation of DDIT3morEGFP lead to a transient growth arrest of HT1080 cells, accumulation of cells in G1 and depletion of cells in the S- and G2-phases (Figure 5) whereas no growth effects were observed in the morEGFP control transfected cells (data not shown). No increased level of apoptotic cells were observed in DDIT3 transfected cells before or after tamoxifen treatment (data not shown).

Transcription factor binding site enrichment for nuclear DDIT3 regulated genes

Genes regulated by nuclear DDIT3 at two hours were considered direct targets for DDIT3. Two different scoring methods for enrichment of predicted transcription factor binding sites (TFBS) using 652 position-scoring weight matrices (PWMs) were applied. Ideally, hits to a PWM should be present in many of the regulated genes, while preferably not present in an overly large proportion of all genes.

A few binding sites (for example V$AP2_Q6, V$SP1_Q2_01, V$CNOT3_01, and V $SRF_01) show moderately low p-values for enrichment among the regulated genes with both scoring methods (Supplementary Table 3). However, several of the top scoring PWMs have putative binding sites in promoter regions of many genes along the genome, while a few top scoring PWMs instead are very rare. Follow up analyses including clustering and functional annotation failed to identify a common pattern of binding site occurrence in the promoters for the regulated genes.

DDIT3 has been shown to form dimers with other C/EBP leucine zipper factors and act as a dominant negative factor or bind as heterodimers to specific sites (18). However, none of the C/EBP sites were scored as enriched among the regulated genes. Neither of the reported CHOP binding sites (18) were found to be enriched.

Twenty of the genes regulated by nuclear DDIT3 contained a cAMP responsive element (CRE) site in their promoter region. Despite difficulties in showing significant enrichment for these CRE sites, we decided to further investigate a possible interaction with DDIT3 since several CRE binding proteins belong to the basic leucine zipper family of transcription factors. HT1080 cells expressing EGFP or nuclear DDIT3-EGFP were treated with the cAMP inducing agent forskolin and expression of responsive genes was analyzed by RT-qPCR. Six genes reported to be cAMP regulated were analyzed. Although the DDIT3 effects on forskolin induction varied between the genes, there was no general inhibition of the forskolin induced effect (Supplementary Figure 5). This indicates that there is no dominant negative effect of DDIT3 in cAMP regulated TF-complexes of these six genes.

Discussion

Mammalian cells recognize and process stress signals and respond with specific gene expression and protein activation programs evolved to minimize damage or induce apoptosis. DDIT3 is a key regulator in stress response and may be triggered by several stress induced signal pathways (1-7, 22). DDIT3 accumulation is regulated both at transcriptional and translational levels (23) and its biological activity is reported to be further modulated by phosphorylation by P38 kinase and CK2 in at least four different sites (24-25).

The GOT3 liposarcoma cell line carries a strongly amplified region of chromosome 12 that harbors the DDIT3 gene (26). This leads to a constitutive expression of cytoplasmic DDIT3, which is further up regulated in stress-exposed cells. In this study we show that cultured human fibroblasts and GOT3 cells accumulate cytoplasmic DDIT3 under tunicamycin and etoposide induced stress conditions (Figure 1a). Tunicamycin is an inhibitor of protein glycosylation and thus a potent inducer of ER stress without immediate genotoxic effects, whereas etoposide is a topoisomerase inhibitor causing multiple double strand breaks. The accumulation of cytoplasmic DDIT3 may thus be a specific response to ER stress and genotoxic double strand breaks in these cell types.

DDIT3 was originally described as a nuclear transcription factor but our data supported by reports from other investigations show that it is often expressed as a cytoplasmic protein (19-20). Most studies reporting nuclear localization of DDIT3 are made with cells transfected or transduced with DDIT3 expression plasmid or virus vectors. Transfection of fibroblasts and GOT3 cells with a DDIT3-EGFP expression vector showed, in agreement with results from other groups, a nuclear expression of the recombinant protein, indicating that the cells are capable of nuclear DDIT3 localization. With these results we hypothesized that the stress induced by the transfection procedure could induce a nuclear DDIT3 expression. To test this possibility we made transient transfections of GOT3 cells with a plasmid. We conclude that stress conditions induced by the transfection procedure lead to increased nuclear accumulation of DDIT3 whereas tunicamycin or etoposide induced stress results in cytoplasmic DDIT3 expression.

The mechanism behind cytoplasmic versus nuclear localization of DDIT3 remains unexplained. Specific inhibition of the P38 and CK2 kinases that are known to phosphorylate DDIT3 showed no effects on tamoxifen induced nuclear translocation of DDIT3morEGFP protein, ruling out that phosphoprylation of these sites are controlling cytoplasmic/nuclear localization of DDIT3. Furthermore, inspection of our western blot results showed no detectable size difference between cytoplasmic and nuclear DDIT3, thus ruling out major protein modifications such as ubiquitinylation or sumoylation as a mechanism for the selective localization.

The cytoplasmic DDIT3 has to be accounted for when biological functions of this protein is studied. We addressed the effects of cytoplasmic DDIT3 and nuclear

DDIT3 by analysis of two stably transfected clones carrying recombinant DDIT3 protein fused to the ligand binding parts of a mutated mouse estrogen receptor and EGFP. Under standard culture conditions, the recombinant DDIT3 is retained in the cytoplasmic compartment and 94 genes were down or up regulated three times or more compared to two morEGFP control clones. These downstream response genes are most likely regulated by several mechanisms and intermediary steps but not by DNA binding and transcription factor activity of DDIT3. DDIT3 may also bind other leucine zipper containing proteins in the cytoplasm and the recently discovered intrinsically disordered domain that mediates binding and interactions with proteins other than leucine zippers (21).

ATF3 is a basic leucine zipper transcription factor that may form heterodimers with DDIT3 (27). ATF3 and DDIT3 are both up regulated by stress-induced ATF4, another basic leucine zipper transcription factor upstream in ER stress signaling. In the present work we show that cytoplasmic DDIT3 induce ATF3 transcription whereas nuclear DDIT3 reduces this effect suggesting a negative feedback loop on the production of ATF3. These differential effects support our hypothesis that cytoplasmic/nuclear localization of DDIT3 has different roles in stress response.

Ontogeny analysis of the cytoplasmic DDIT3 regulated genes indicated that functions related to cellular movement and migration were affected and a migration assay showed an impaired migration thus confirming the ontogeny analysis. The assay showed an impaired migration also in HT1080 cells stably expressing nuclear DDIT3, indicating that the migration may be inhibited both by cytoplasmic and nuclear DDIT3. DDIT3 regulation of several migration/movement-associated genes may provide mechanistic explanations for the impaired migration (Table 1). DSTN, an actin depolymerising protein was, for example, down regulated in DDIT3 expressing cells. We have previously shown that DDIT3 binds cyclin dependent kinase 2 (CDK2) and that CDK2 also binds cytoskeletal proteins such as myosin 9, myosin 10 and plectin in DDIT3 expressing cells (28). This binding of CDK2 to some cytoplasmic cytoskeleton proteins may be a part of mechanisms that affect migration capacity.

It is also interesting to note that well differentiated liposarcomas, regularly expressing DDIT3 due to gene amplifications, are slow growing non-aggressive tumors that rarely metastasize into surrounding tissues (26). These features may be an effect of the DDIT3 impaired migration. DDIT3 is also expressed in terminal adipocyte differentiation and inhibition of migration may be a part of this process.

Many of the genes regulated within the first hours after DDIT3 nuclear transition, are probably direct targets of DDIT3. Our results and ontogeny analysis confirm earlier notions that DDIT3 controls growth and apoptosis (13). Forced expression of DDIT3 was reported to induce a G1 cell cycle arrest (10) and our experimental system recapitulates this effect when the protein is translocated to the nuclei (Figure 5). A direct role in proliferation/cell cycle regulation was also pointed out as a significant functional category by our functional analysis. DDIT3 regulation of several genes/functions may execute the growth arrest but further investigations are needed to dissect the detailed mechanism. From our study of transfected HT1080 cells it is clear that nuclear but not cytoplasmic DDIT3 causes a G1 arrest. The stably transfected cell lines are, however the result of an extremely strong artificial selection for cells that can tolerate and grow in the presence of cytoplasmic DDIT3. The experiments in this study may therefore not give information on this issue and we cannot rule out that cytoplasmic DDIT3 could cause an anti-proliferative effect in other cells and conditions.

Several studies report DDIT3 induced apoptosis. We detected, however, no DDIT3 induced apoptosis cells in our experiments although several apoptosis controlling genes were regulated by DDIT3. The apoptotic effect of DDIT3 is probably cell type dependent since the divergent reports are based on experiments with different cell types and stress agents. Our results showing that cytoplasmic accumulation of DDIT3 is a normal response to tunicamycin induced ER stress may add insight to these conflicting reports. Thus, these differences in regulated genes/apoptotic functions may indicate distinct responses to cytoplasmic and nuclear DDIT3. Upon nuclear localization, apoptosis protective genes *PAX2, PHLDA1, SGK1, SPRY2*, and *SYVN1* were all down-regulated supporting a DDIT3 induction of apoptosis functions. But these effects may be balanced by the simultaneous down-regulation of the pro-apoptotic genes *KLF6, PLK2, RND3* and *TXNIP* in our HT1080 based experimental system.

The most up regulated DDIT3 induced genes was *EGR1* (Supplementary Table 1b), a zinc finger type transcription factor involved in a variety of biological responses and effects (29). *EGR1* has been linked to growth and apoptosis control (30). In some tumor types *EGR1* is recurrently deleted or down regulated and reported to act as a tumor suppressor gene, while it is over expressed and considered an oncogene in others (29-31). We conclude from our study that *EGR1* may be one of the most important immediate target genes for nuclear *DDIT3.* EGR1 forms a DNA binding complex with C/EBPB, which also is an important dimerization partner with DDIT3 (32).

Many of the genes regulated within the first hours after DDIT3 nuclear transition, are probably direct targets of DDIT3 and by investigation of the promoter regions it would theoretically be possible to identify DDIT3 DNA binding sites. Our attempts to identify recurrent DDIT3 binding sites failed however. A possible explanation is that DDIT3 forms dimers with several alternative partners and the alternative heterodimers could bind to different sites.

We conclude that cytoplasmic DDIT3 has specific effects on migration. Except for migration, cytoplasmic and nuclear DDIT3 regulate the same functional categories of genes but nuclear DDIT3 adds more regulated genes to each category. Nuclear translocation thus leads to a step up of functions initiated by cytoplasmic DDIT3.

Materials and methods

Expression vectors and transfections

The full length coding regions of *DDIT3* was cloned into the pEGFP-N1 vector (Clontech Laboratories, Inc.) in frame with the EGFP as previously described. morGFP and morEGFP vectors were constructed by an in-frame ligation of the morLBD construct immediately upstream of the gene encoding GFP. All constructs were confirmed by sequencing.

The mouse estrogen receptor ligand binding domain (MOR-LBD) construct was made by mutating the wild-type mouse oestrogen receptor (a kind gift from Dr. M. Parker). The ligand binding domain (DNA encoding amino acids 290-599) of the receptor was cloned using the primer set: MORLBD BamHI-U (5'TATGGATCCAGGAGACATGAGGGCTGCCAACCTTTG3') and MORLBD BamHI-L (5'TATGGATCCATCGTGTTGGGGAAGCCCTCT3'). The G525R point mutation was introduced by PCR mutagenesis and amplification of circular DNA in vitro (33) using the primer set (5'-3'): (GGCACATGAGTAACAAACGCATGG) and MORLBD mut-L (ATGTTGTAGAGATGCTCCATGCGTTTGTT). The MOR-LBD G525R mutant is unable to bind estrogen, yet it retains affinity for a synthetic ligand, 4-hydroxy-tamoxifen. For nuclear translocation of mutant MOR-LBD fused DDIT3-GFP, 4-hydroxy-tamoxifen was added to the medium at a final concentration of 100 nM (34).

Cell culture and growth conditions

The human fibrosarcoma cell lines HT1080, HT1080-EGFP and HT1080 DDIT3-EGFP, were kept frozen in liquid nitrogen or cultured at 37° C and 5% CO_2 in RPMI 1640 medium with HEPES buffer supplemented with 2 mM L-glutamine, 50 U/ml penicillin, 50 µg/ml streptomycin and 8% FCS (Invitrogen). G418 (200 µg/ml, Invitrogen) was constantly added to cell lines HT1080 DDIT3-EGFP and HT1080 EGFP to ensure stable expression of EGFP constructs in the cell population.

RNA was extracted using RNAeasy extraction kit (Qiagen) from the cells at zero, two, and eight hours after addition of 4-hydroxy-tamoxifen and stored at -140^{o}C. RNA from the cell samples and a common human universal reference RNA (Stratagene 740000) was used as templates for cDNA synthesis with Cy3 and Cy5 labeled nucleotides according to the instructions for the Pronto Plus 6 labeling kit (Corning).

For experiments with forskolin treatment, cells were seeded in Petri dishes at a density of 180 000 cells / plate and a total volume of 4 ml/plate. One Petri dishes plate for each cell line and each test were done in triplets. Forskolin was added in a concentration of $1{,}8*10^{-5}$ M. The experiments were performed in three independent biological replicates. For kinase inhibition experiments, P38 kinase inhibitor

SB-203580 Promega and Casein kinase II inhibitor I (Calbiochem) were added at 10 µM and 50 µM respectively four hours before tamoxifen treatment.

Cell migration assay

Wild type and stably transfected HT1080 cells were seeded to petri dishes (35 mm in diameter) at a 1:1 ratio to a total of 100 000 cells/well. At 80% confluence a scratch wound was made in the monolayer. After two days of incubation the cultures were fixed in 4% formaldehyde and stained with ethidium bromide at a final concentration of 20 µg/ml. Wounded areas with cell densities suitable for counting were photographed on a fluorescence microscope and the number of ethidium bromide stained cells and EGFP stained cells were counted. The ratio of cells stained with ethidium bromide (all cells) and cells stained with EGFP were counted in several non-wounded control areas. The experiment was repeated six times for each of the HT1080 cell lines transfected with EGFP, DDIT3-EGFP, and DDIT3morEGFP.

A detailed description of migration assay modeling is provided in the Supplementary Information. Briefly, the ratios of migration rates for all EGFP stained cell types and migration rate for wild type cells can be deduced by using Bayes theorem. To compare migration rates for EGFP stained cells to the migration rate for wild type cells we used a sign test. A Wilcoxon test was employed to test the differences in migration rates between different EGFP-stained cell lines.

Microarray hybridization and feature extraction

Equal quantities of labeled cDNA and reference cDNA were hybridized to Agilent G4112F microarrays and the arrays were scanned using an Agilent G2565CA microarray scanner (Agilent Technologies, Palo Alto, CA). Feature extraction was performed with Agilent's Feature Extraction 10.4 Image Analysis Software.

Microarray preprocessing and analysis

Data analysis was performed with the open source statistical software R using the LIMMA package available within the Bioconductor suite (35). A loess smoother was applied to each array to remove intensity dependent trends and the arrays were quantile normalized for comparability. All spots not corresponding to human genes were removed before further analysis. The expression levels of duplicated probes were averaged.

Since the biological replicates were made with different clones, the within replicate similarities were assessed using Pearson correlations. The replicates exhibited a consistent high correlation (all $>$ 0.85) and no systematic deviations were found within or between clones (data not shown).

For each probe on the arrays, the normalized $\log_2$-fold changes (M-values) were calculated and retained for downstream analysis. To assess the changes in gene expression induced by cytoplasmic DDIT3, the M-values were compared for the zero time point in the DDIT3morEGFP cell line with the zero time point in the cell line with the morEGFP construct alone. Similarly, the differentially expressed genes induced by nuclear DDIT3 for the two and eight-hour time points were compared to the zero time point of the DDIT3morEGFP cell line (creating two so-called contrasts). The same procedure was employed in the control cell line, and all genes with a fold-change of 1.5 or higher in either contrast were discarded before further analysis, in order to remove any effects induced by the morEGFP construct alone. *EGR1*, with a regulation slightly larger than 1.5 fold in the morEGFP cell line, was also included because of a very large fold-change at the eight hour time point. Genes regulated by the morEGFP construct alone were not removed from the analysis of cytoplasmic DDIT3 induced genes.

Raw and normalized microarray data is available at the ArrayExpress repository with accession number E-MEXP-2709.

Functional annotation and network analysis

Probes on the microarrays were ranked according M-value ($\log_2$ fold-change). Since the responses induced by cytoplasmic DDIT3 were more pronounced, genes with a fold change of three or more were considered differentially expressed. For the genes induced by nuclear DDIT3, the cutoff was a two-fold regulation or more. The functional analysis of the regulated genes was generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com). Genes from the different dataset that met the above described expression criteria and were associated with biological functions and/or diseases in the Ingenuity Pathways Knowledge Base were considered for the analysis.

Enriched functional categories and subgroups within each category among the regulated genes were identified for all gene sets, and the four most significant categories representing fundamental cellular functions in each set were chosen for further study. The "cancer" category was omitted since we do not consider cancer as a cellular functional category.

Fischer's exact test (36) was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that data set is due to chance alone.

To generate functional networks, the differentially expressed genes were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus genes were then algorithmically generated based on their connectivity.

Transcription factor binding site prediction

Transcription factor binding sites (TFBS) were predicted using the MATCH program (37) and a collection of 652 vertebrate positions-scoring weight matrices (PWMs) from the TRANSFAC database (38). Promoter sequences were extracted from the TRANSPro database, where 500 base pairs upstream and 100 base pairs downstream of the predicted transcription start site were selected for each gene. Matches to a PWM were considered a hit if either the core similarity score was 1.0 or the matrix similarity score exceeded 0.95 (conservative choice).

CRE sites present in the promoters of the nuclear DDIT3 regulated genes at two hours were predicted by counting hits for the following PWMs; V$CREBATF_Q6, V $CREBP1CJUN_01, V$CREBP1_01, V$CREBP1_Q2, V$CREB_01, V $CREB_02, V$CREB_Q2, V$CREB_Q2_01, V$CREB_Q3, V$CREB_Q4, V $CREB_Q4_01, V$TAXCREB_01, V$TAXCREB_02.

TFBS enrichment

The enrichment of all TFBS among the regulated genes at two hours was tested in two different ways. A hypergeometric test (or equivalently Fisher's exact test) was used to compare the proportion of hits of TFBS among the regulated genes and the non-regulated genes. This method is standard procedure, but is highly dependent on the choice of cut-off value for regulation, and may also be sensitive to the fact that the number of regulated genes is relatively small. As a complement, a permutation test was applied to all the genes present on the arrays (i.e. not based explicitly on the regulation cut-off). Details concerning the design of the permutation test are given in the Supplementary Information. The permutation test works as a complement to the hypergeometric test, and ideally low p-values with both methods should indicate significant enrichment of TFBS.

Fluorescence microscopy and western blot analysis

Fluorescence microscopy and western blot analysis was preformed as previously described (39). A Leica DMI 600B microscope with a Leica DFC 360 FX camera was used for life imaging. The recording was made during a one-hour time span with an image taken every 20 seconds (tamoxifen was added after 5 minutes). The software used for recording was the Leica Application Suite AF.

Reverse transcription quantitative real-time PCR (RT-qPCR)

For microarray validation the following protocol was used: Reverse transcription was performed on approximately 1µg total RNA using SuperScript III (Invitrogen) according to the manufacturers instructions, using a mixture of 5 µM oligo(dT) and 5 µM random hexamers (both Invitrogen) as primers. Real-time PCR measurements were performed on a LightCycler 480 (Roche) using the iQ SYBR Green Supermix (Bio-Rad) with 400nM of each PCR primer. Primer sequences are available in Supplementary Table 4.

Formation of correctly sized PCR products was confirmed by agarose gel electrophoresis for all assays and melting curve analysis for all samples. Gene expression data was normalized against PPIA and HPRT by geometric averaging (40). The reference genes were selected using the Human Endogenous Control Gene Panel (TATAA Biocenter) and GenEx software (MultiD Analyses).

For the Forskolin experiment QuantiTect Reverse Transcription Kit and QuantiTect SYBR Green Kit (both QIAGEN) were used for reverse transcription and real-time PCR respectively. Data was normalized against GAPDH.

Acknowledgements

This work was supported by grants from the Swedish Cancer Society, Assar Gabrielssons Research Foundation, Johan Jansson Foundation for Cancer Research, Swedish Children's Cancer Society, LUA/ALF Västra Götaland, Svenska Sällskapet för Medicinsk Forskning, Wilhelm and Martina Lundgren Foundation for Scientific Research, the Swedish Research Council, the Gothenburg Stochastic Centre, Gothenburg Mathematical Modeling Centre, and the Swedish Foundation for Strategic Research.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Information accompanies the paper on Cell Death and Differentiation website.

References

1. Liu T, Laurell C, Selivanova G, Lundeberg J, Nilsson P, Wiman KG. Hypoxia induces p53-dependent transactivation and Fas/CD95-dependent apoptosis. *Cell death and differentiation* 2007 Mar; **14**(3)**:** 411-421.

2. Tang JR, Nakamura M, Okura T, Takata Y, Watanabe S, Yang ZH*, et al.* Mechanism of oxidative stress-induced GADD153 gene expression in vascular smooth muscle cells. *Biochem Biophys Res Commun* 2002; **290**(4)**:** 1255-1259.

3. Ma Y, Brewer JW, Diehl JA, Hendershot LM. Two distinct stress signaling pathways converge upon the CHOP promoter during the mammalian unfolded protein response. *J Mol Biol* 2002; **318**(5)**:** 1351-1365.

4. Jousse C, Bruhat A, Harding HP, Ferrara M, Ron D, Fafournoux P. Amino acid limitation regulates CHOP expression through a specific pathway independent of the unfolded protein response. *FEBS letters* 1999 Apr 9; **448** (2-3)**:** 211-216.

5. Jackman J, Alamo I, Jr., Fornace AJ, Jr. Genotoxic stress confers preferential and coordinate messenger RNA stability on the five gadd genes. *Cancer Res* 1994; **54**(21)**:** 5656-5662 Issn: 0008-5472.

6. Gately DP, Jones JA, Christen R, Barton RM, Los G, Howell SB. Induction of the growth arrest and DNA damage-inducible gene GADD153 by cisplatin in vitro and in vivo. *Br J Cancer* 1994; **70**(6)**:** 1102-1106.

7. Luethy JD, Holbrook NJ. Activation of the gadd153 promoter by genotoxic agents: a rapid and specific response to DNA damage. *Cancer Res* 1992; **52**(1)**:** 5-10 Issn: 0008-5472.

8. Oyadomari S, Mori M. Roles of CHOP/GADD153 in endoplasmic reticulum stress. *Cell death and differentiation* 2004 Apr; **11**(4)**:** 381-389.

9. Zinszner H, Kuroda M, Wang X, Batchvarova N, Lightfoot RT, Remotti H*, et al.* CHOP is implicated in programmed cell death in response to impaired function of the endoplasmic reticulum. *Genes Dev* 1998; **12**(7)**:** 982-995.

10. Barone MV, Crozat A, Tabaee A, Philipson L, Ron D. CHOP (GADD153) and its oncogenic variant, TLS-CHOP, have opposing effects on the induction of G1/S arrest. *Genes Dev* 1994 Feb 15; **8**(4)**:** 453-464.

11. Diakogiannaki E, Dhayal S, Childs CE, Calder PC, Welters HJ, Morgan NG. Mechanisms involved in the cytotoxic and cytoprotective actions of saturated versus monounsaturated long-chain fatty acids in pancreatic beta-cells. *The Journal of endocrinology* 2007 Aug; **194**(2)**:** 283-291.

12. Araki E, Oyadomari S, Mori M. Impact of endoplasmic reticulum stress pathway on pancreatic beta-cells and diabetes mellitus. *Experimental biology and medicine (Maywood, NJ* 2003 Nov; **228**(10)**:** 1213-1217.

13. Mori N, Stein R, Sigmund O, Anderson D. A cell-type specific silencer element that controls the neural-specific expression of the SCG10 gene. *Neuron* 1990; **4:** 583-594.

14. Engstrom K, Willen H, Kabjorn-Gustafsson C, Andersson C, Olsson M, Goransson M*, et al.* The myxoid/round cell liposarcoma fusion oncogene FUS-DDIT3 and the normal DDIT3 induce a liposarcoma phenotype in transfected human fibrosarcoma cells. *Am J Pathol* 2006 May; **168**(5)**:** 1642-1653.

15. Thorp E, Li G, Seimon TA, Kuriakose G, Ron D, Tabas I. Reduced apoptosis and plaque necrosis in advanced atherosclerotic lesions of Apoe-/- and Ldlr-/- mice lacking CHOP. *Cell metabolism* 2009 May; **9**(5)**:** 474-481.

16. Wedel A, Ziegler-Heitbrock HWL. The C/EBP family of transcription factors. *Immunobiol* 1995; **193:** 171-185.

17. Ron D, Habener JF. CHOP, a novel developmentally regulated nuclear protein that dimerizes with transcription factors C/EBP and LAP and functions as a dominant-negative inhibitor of gene transcription. *Genes Dev* 1992 Mar; **6**(3)**:** 439-453.

18. Ubeda M, Wang X-Z, Zinszer H, Wu I, Habener J, Ron D. Stress-induced binding of the transcription factor CHOP to a novel DNA control element. *Mol Cell Biol* 1996; **16**(1)**:** 1479-1489.

19. Cui K, Coutts M, Stahl J, Sytkowski AJ. Novel interaction between the transcription factor CHOP (GADD153) and the ribosomal protein FTE/S3a modulates erythropoiesis. *J Biol Chem* 2000 Mar 17; **275**(11)**:** 7591-7596.

20. Lorz C, Justo P, Sanz A, Subira D, Egido J, Ortiz A. Paracetamol-induced renal tubular injury: a role for ER stress. *J Am Soc Nephrol* 2004 Feb; **15**(2)**:** 380-389.

21. Singh VK, Pacheco I, Uversky VN, Smith SP, MacLeod RJ, Jia Z. Intrinsically disordered human C/EBP homologous protein regulates biological activity of colon cancer cells during calcium stress. *J Mol Biol* 2008 Jul 4; **380**(2)**:** 313-326.

22. Bruhat A, Jousse C, Wang XZ, Ron D, Ferrara M, Fafournoux P. Amino acid limitation induces expression of CHOP, a CCAAT/enhancer binding protein-related gene, at both transcriptional and post-transcriptional levels. *J Biol Chem* 1997 Jul 11; **272**(28)**:** 17588-17593.

23. Jousse C, Bruhat A, Carraro V, Urano F, Ferrara M, Ron D*, et al.* Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5'UTR. *Nucleic Acids Res* 2001 Nov 1; **29**(21)**:** 4341-4351.

24. Wang XZ, Ron D. Stress-induced phosphorylation and activation of the transcription factor CHOP (GADD153) by p38 MAP Kinase. *Science* 1996 May 31; **272**(5266)**:** 1347-1349.

25. Ubeda M, Habener JF. CHOP transcription factor phosphorylation by casein kinase 2 inhibits transcriptional activation. *J Biol Chem* 2003 Oct 17; **278**(42)**:** 40514-40520.

26. Persson F, Olofsson A, Sjogren H, Chebbo N, Nilsson B, Stenman G*, et al.* Characterization of the 12q amplicons by high-resolution, oligonucleotide array CGH and expression analyses of a novel liposarcoma cell line. *Cancer Lett* 2008 Feb 18; **260**(1-2)**:** 37-47.

27. Chen BP, Wolfgang CD, Hai T. Analysis of ATF3, a transcription factor induced by physiological stresses and modulated by gadd153/Chop10. *Mol Cell Biol* 1996; **16**(3)**:** 1157-1168.

28. Bento C, Andersson MK, Aman P. DDIT3/CHOP and the sarcoma fusion oncoprotein FUS-DDIT3/TLS-CHOP bind cyclin-dependent kinase 2. *BMC cell biology* 2009; **10:** 89.

29. Adamson ED, Mercola D. Egr1 transcription factor: multiple roles in prostate tumor cell growth and survival. *Tumour Biol* 2002 Mar-Apr; **23**(2)**:** 93-102.

30. Yu J, Baron V, Mercola D, Mustelin T, Adamson ED. A network of p73, p53 and Egr1 is required for efficient apoptosis in tumor cells. *Cell death and differentiation* 2007 Mar; **14**(3)**:** 436-446.

31. Krones-Herzig A, Mittal S, Yule K, Liang H, English C, Urcis R*, et al.* Early growth response 1 acts as a tumor suppressor in vivo and in vitro via regulation of p53. *Cancer Res* 2005 Jun 15; **65**(12)**:** 5133-5143.

32. Zhang F, Lin M, Abidi P, Thiel G, Liu J. Specific interaction of Egr1 and c/EBPbeta leads to the transcriptional activation of the human low density lipoprotein receptor gene. *J Biol Chem* 2003 Nov 7; **278**(45)**:** 44246-44254.

33. Chen Z, Ruffner DE. Amplification of closed circular DNA in vitro. *Nucleic Acids Res* 1998 Dec 1; **26**(23)**:** 1126-1127.

34. Littlewood TD, Hancock DC, Danielian PS, Parker MG, Evan GI. A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. *Nucleic Acids Res* 1995 May 25; **23**(10)**:** 1686-1690.

35. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S*, et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**(10)**:** R80.

36. Agresti A. *Categorical data analysis*, 2nd ed. edn. Wiley-Interscience: New York ; Chichester, 2002, xv, 710 p.pp.

37. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 2003 Jul 1; **31**(13)**:** 3576-3579.

38. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006 Jan 1; **34**(Database issue)**:** D108-110.

39. Andersson MK, Stahlberg A, Arvidsson Y, Olofsson A, Semb H, Stenman G, *et al.* The multifunctional FUS, EWS and TAF15 proto-oncoproteins show cell type-specific expression patterns and involvement in cell spreading and stress response. *BMC Cell Biol* 2008; **9:** 37.

40. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002 Jun 18; **3**(7)**:** RESEARCH0034.

Titles and legends to figures

Figure 1. Subcellular localization of DDIT3. (a) Immunoblot analysis of nuclear (Nu) and cytoplasmic (Cy) extracts of human liposarcoma cell line GOT3 and normal human fibroblasts F470 following tunicamycin treatment. Cytoplasmic accumulation of DDIT3 is seen in both cell lines compared to untreated cells. GAPDH and Lamin A are cytoplasmic and nuclear markers, respectively. (b) Confocal microscopy (upper panel) of HT1080 cells containing DDIT3morEGFP before and after addition of tamoxifen. A translocation of the EGFP tagged recombinant DDIT3 protein from the cytoplasm to the nucleus can be seen after the addition of tamoxifen (see also Supplementary Film Clip). Immunoblot analysis (lower panel) of nuclear and cytoplasmic extracts. GAPDH and H1A were used as cytoplasmic and nuclear markers, respectively.

Figure 2. Enrichment of functional categories among genes regulated by DDIT3. Enrichment of functional categories among genes regulated by cytoplasmic DDIT3 (white bars), nuclear DDIT3 at two hours (striped bars), and nuclear DDIT3 at eight hours (black bars). The categories are ordered by the significance among the cytoplasmic DDIT3 regulated genes. The y-axis shows -log10-transformation of enrichment p-values.

Figure 3. DDIT3 affects cell migration. (a) Ratios of migration rates for EGFP (black dots), DDIT3-EGFP (grey triangles), and DDIT3morEGFP (diamonds) compared to wild type cells for the six replicates within each of the three experiments. EGFP expressing cells migrate faster than wild type cells, while DDIT3 expressing cells migrate slower (sign test, most extreme outcome of the statistic, p-value ~ 0.03). (b) Pair wise ratios (of all replicates) of migration rates for DDIT3-EGFP expressing cells compared to EGFP expressing cells (grey triangles) and DDIT3morEGFP expressing cells compared to EGFP expressing cells (black squares). The pair wise ratios are used to form the Mann-Whitney U statistic, which with the most extreme outcome gives a p-value ~ 0.002 for differences in migration rates of the two DDIT3 expressing cell types compared to EGFP expressing cells.

Figure 4. Genes regulated by cytoplasmic and nuclear DDIT3. (a) Nuclear DDIT3 up and down regulated genes at two and eight hours after the addition of tamoxifen (at least two-fold regulation compared). (b) Genes regulated both by cytoplasmic DDIT3 and nuclear DDIT3 (excluding genes regulated in the morEGFP cell line). White bars represent regulation by cytoplasmic DDIT3. Grey and black bars represent the regulation of nuclear DDIT3 compared to levels in cells with cytoplasmic DDIT3 at two and eight hours after tamoxifen addition, respectively.

Figure 5. Flow cytometry analysis of DDIT3 effects. pDDIT3morEGFP transfected cells cultured with and without tamoxifen were analyzed with flow cytometry. A transient growth arrest and accumulation in G1 and depletion of cells in S- and G2-phases was observed for tamoxifen treated cells.

Figure 1

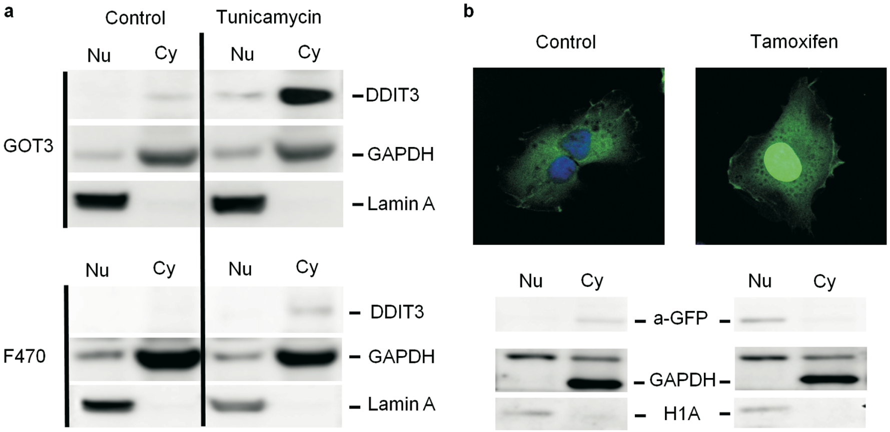


Figure 2

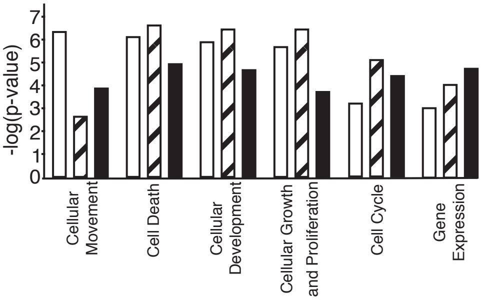


Figure 3

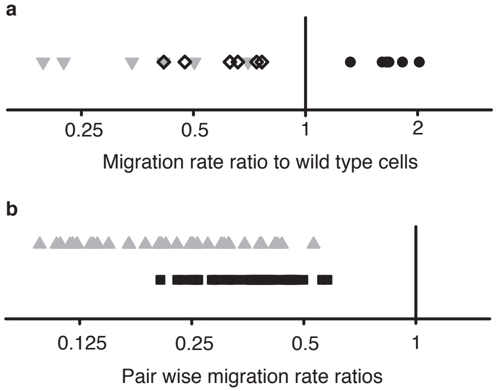


Figure 4

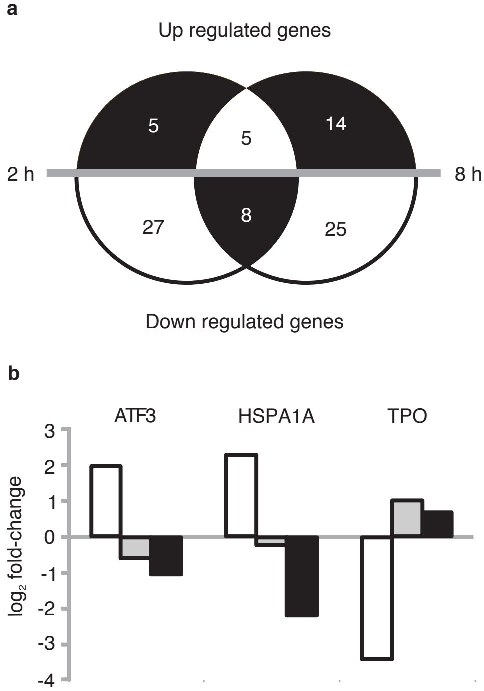
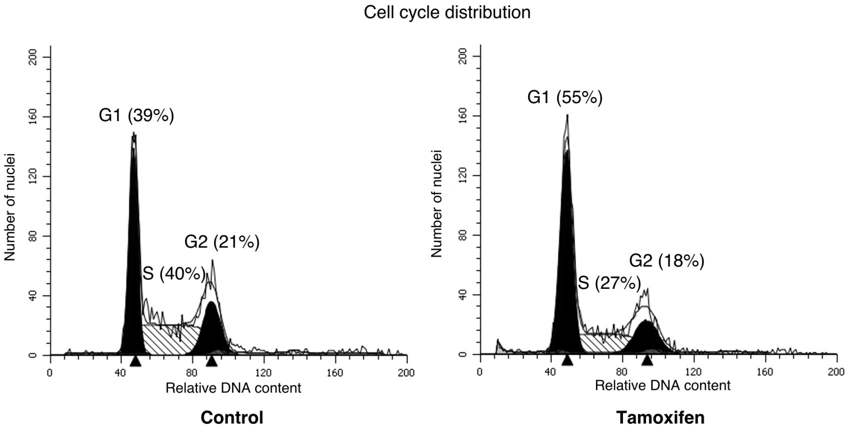


Figure 5


Cell cycle distribution
Number of nuclei
0
40
80
120
160
200
G1 (39%)
S (40%)
G2 (21%)
0
40
80
120
160
200
Relative DNA content
Control
Number of nuclei
0
40
80
120
160
200
G1 (55%)
S (27%)
G2 (18%)
0
40
80
120
160
200
Relative DNA content
Tamoxifen

Supplementary Figure Legends

Supplementary Figure 1: Overlap of gene content in functional categories. (a) Cytoplasmic DDIT3 induced genes. (b) Nuclear DDIT3 induced genes after two hours of tamoxifen addition. (c) Nuclear DDIT3 induced genes after eight hours of tamoxifen addition.

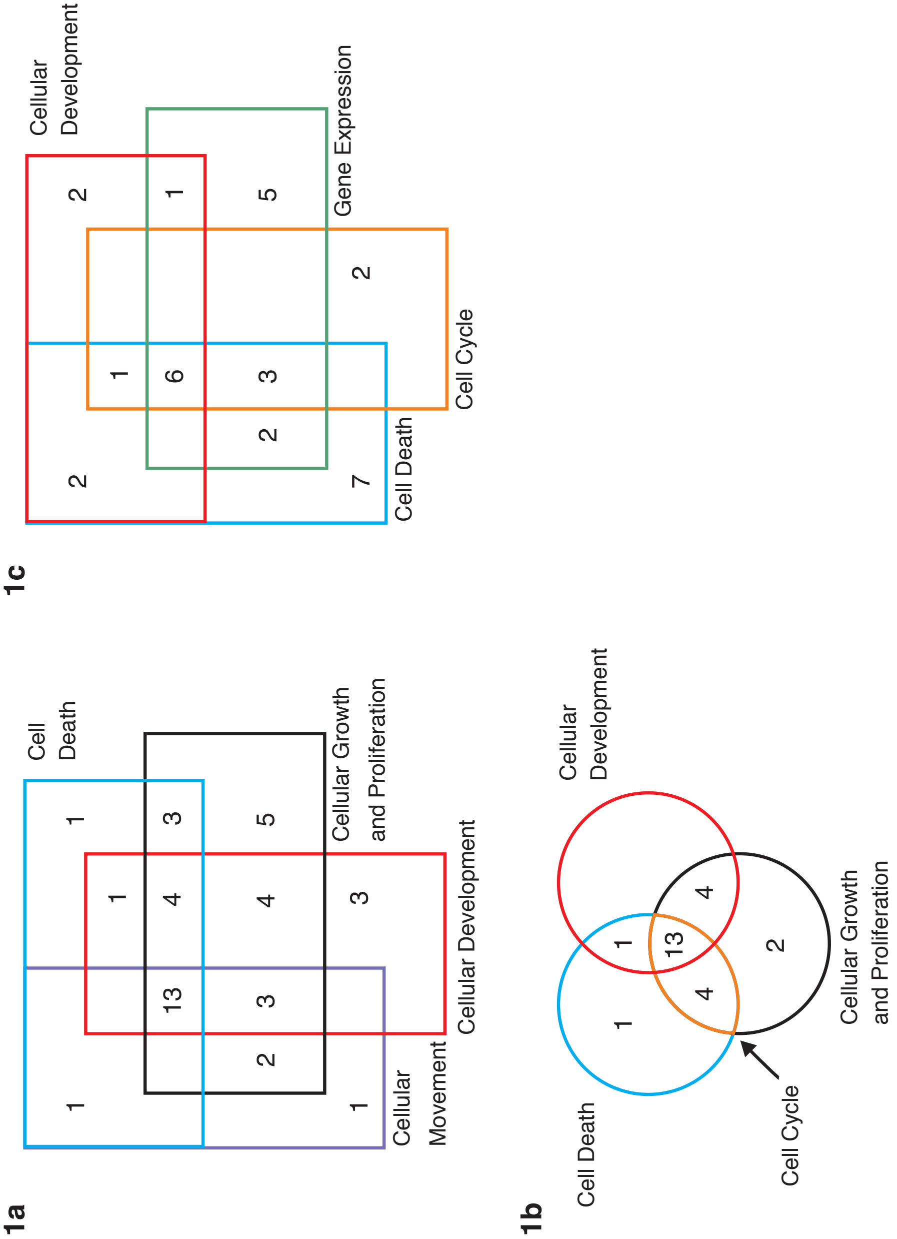
Supplementary Figure 2: Interaction networks formed by the regulated genes. (a) Cytoplasmic DDIT3. (b) Nuclear DDIT3 after two hours of tamoxifen addition. (c) Nuclear DDIT3 after eight hours of tamoxifen addition. Panels (b) and (c) indicate EGR1 as a possible network hub (node connecting several paths in the network). Red hues correspond to up regulation, while green hues correspond to down regulation (no coloring means no differential expression). Vertical diamonds, horizontal diamonds, vertical ovals, and horizontal ovals represent enzymes, peptidases, transmembrane receptors, and transcription factors, respectively. Squares, rectangles, up-pointing triangles, and down-pointing triangles denote cytokines, G-protein coupled receptors, phosphatases, and kinases, in turn. Double circles, trapezoids, and single circles symbolize complexes, transporters, and finally, gene products with other functions. A dashed line corresponds to an indirect relationship, while a solid line represents a direct relationship.

Supplementary Figure 3. Overview of design and analysis of the microarray experiment. Genes regulated by cytoplasmic DDIT3 were extracted by making the comparison indicated with (1), comparing the expression between the morEGFP and DDIT3morEGFP cell lines before addition of tamoxifen. Genes regulated by nuclear DDIT3 were extracted with the comparisons indicated by (2) and (3) within the DDIT3morEGFP cell line, for two and eight hours after tamoxifen addition respectively. Genes regulated in the morEGFP cell line were removed before analysis of nuclear DDIT3 regulation. Each square corresponds to a microarray experiment replicate.

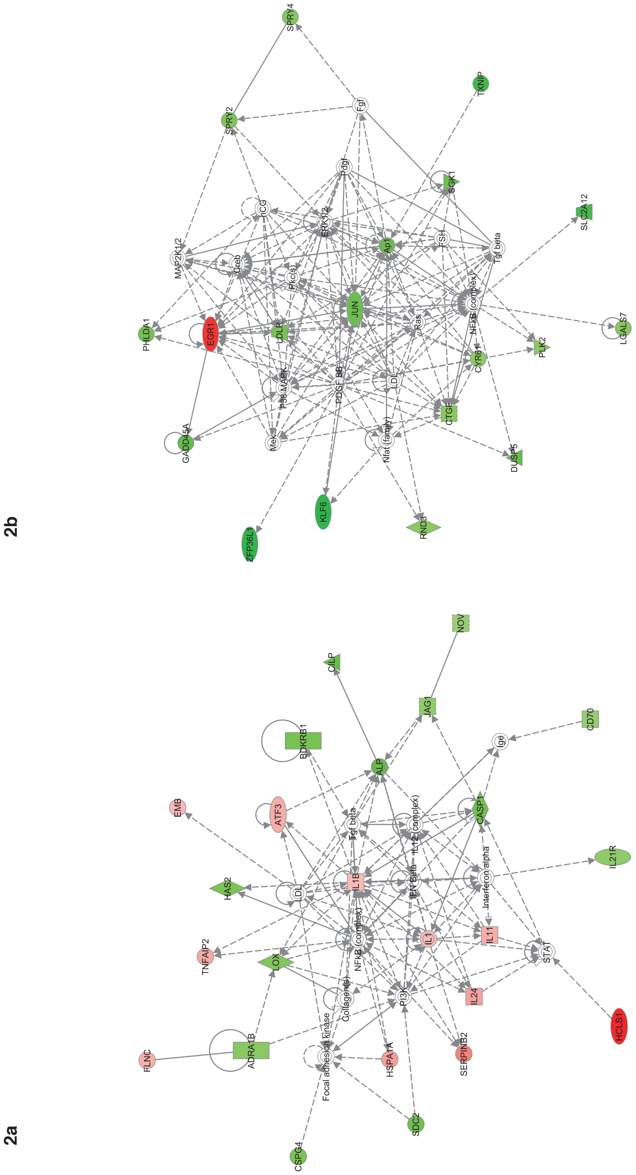
Supplementary Figure 4. Gene expression in morEGFP and DDIT3morEGFP cell lines – comparison of regulated genes. Expression within the morEGFP cell line, within the DDIT3morEGFP cell line, and between both cell lines for genes regulated by cytoplasmic DDIT3. Please note that genes showing differential expression in the morEGFP cell line were excluded in the analysis of nuclear DDIT3 regulation.

Supplementary Figure 5. No DDIT3 inhibition of cre-binding factors in regulation of gene expression. Log$_2$ fold-changes for six genes following treatment with the cAMP inducing agent forskolin in cells expressing EGFP or nuclear DDIT3-EGFP. The experiment was made with three biological replicates, and differences in induction between the two cell types were assessed with a two-sample t-test (* P-value <0.05). Error bars indicate the standard error of the mean values.

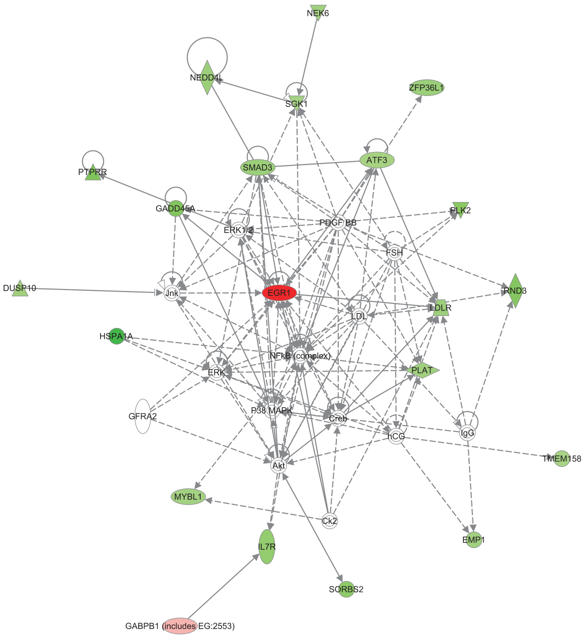
Supplementary Figure 1



Supplementary Figure 2

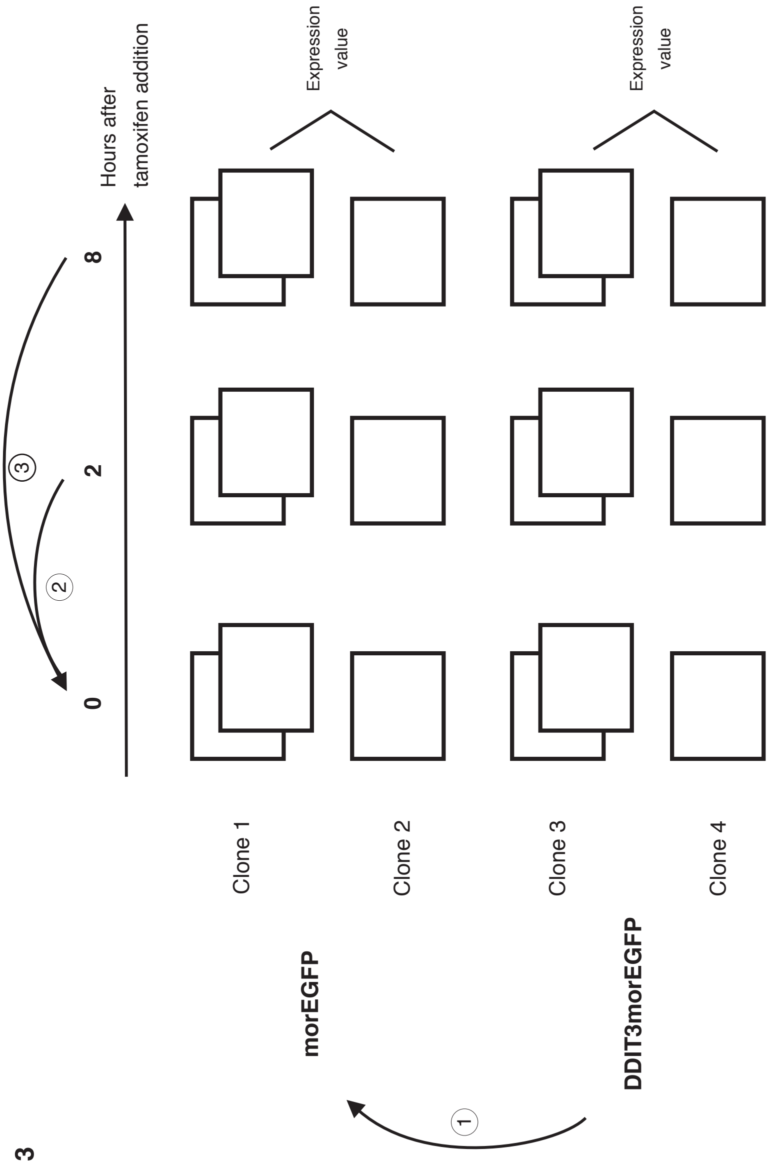


2c



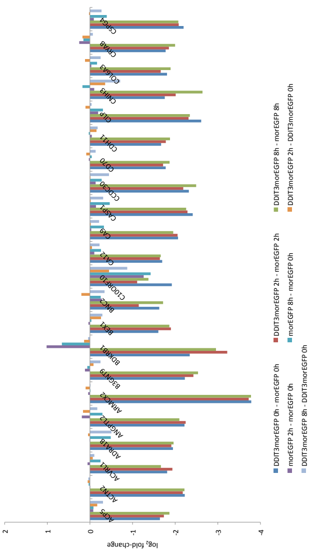
NEK6
NEDD4L
SGK1
ZFP36L1
SMAD3
ATF3
PTPRR
GADD45A
PLK2
ERK1/2
PDGF BB
FSH
RND3
DUSP10
Jnk
EGR1
LDLR
LDL
HSPA1A
NFkB (complex)
ERK
PLAT
GFRA2
P38 MAPK
Creb
hCG
IgG
TMEM158
Akt
MYBL1
Ck2
EMP1
IL7R
SORBS2
GABPB1 (includes EG:2553)

Supplementary Figure 3

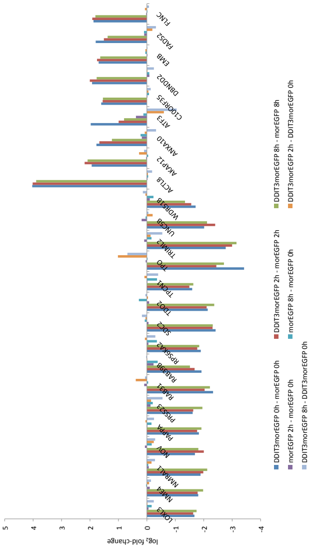


Supplementary Figure 4

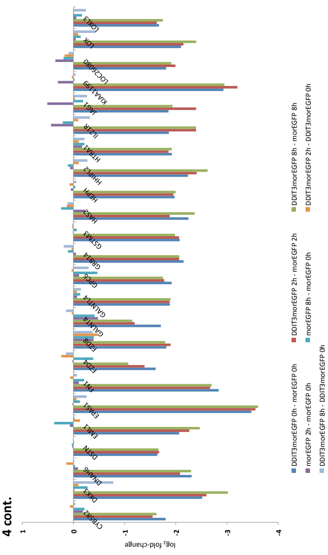
4



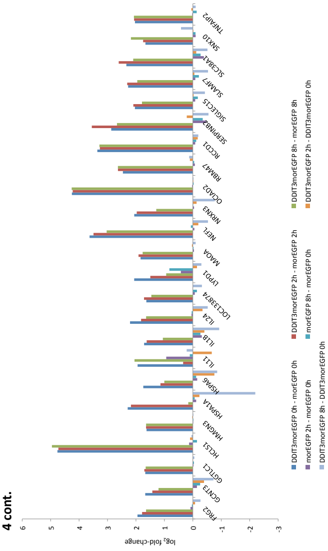
4 cont.



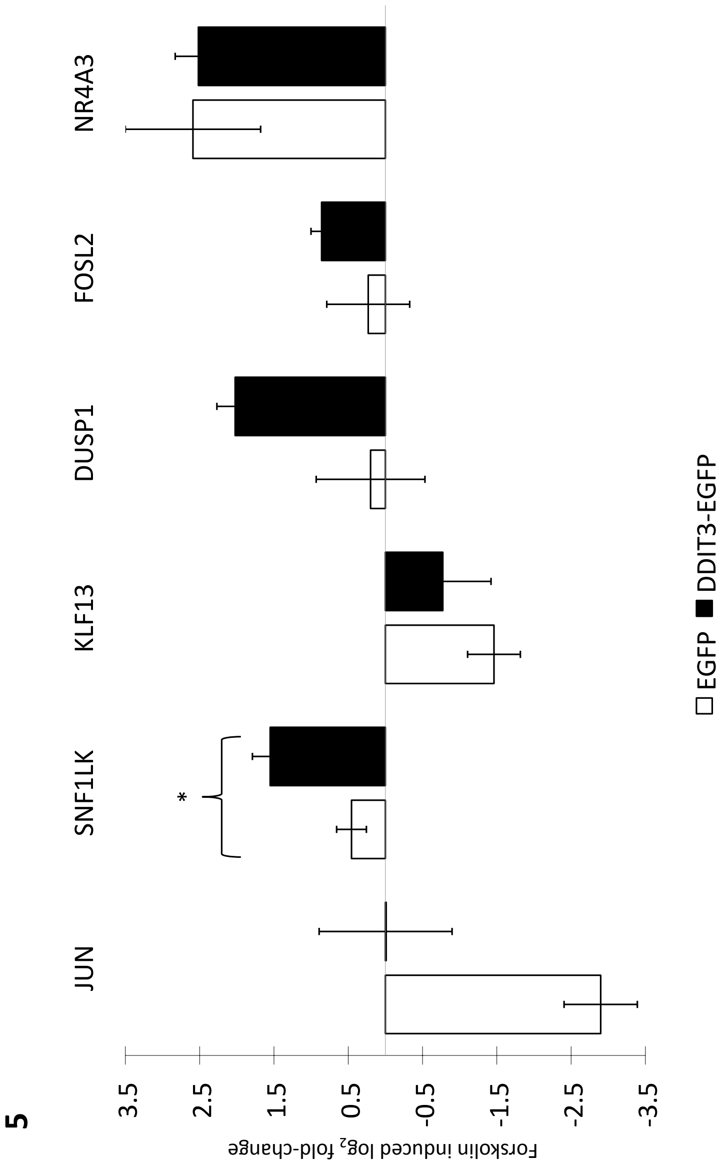
4 cont.



4 cont.



Supplementary Figure 5



Supplementary Table and Film Clip Legends

Supplementary Table 1. DDIT3 regulated genes. Differentially expressed genes induced by (a) cytoplasmic DDIT3, and (b) nuclear DDIT3 after two and eight hours of tamoxifen addition. (c) Validation of microarray expression with qPCR for selected genes (raw values and $\log_2$-fold changes compared with corresponding values in array data).

Supplementary Table 2. Functional annotations for differentially expressed genes. Functional annotations (Ingenuity Pathway Analysis) for differentially expressed genes induced by (a) cytoplasmic DDIT3, (b) nuclear DDIT3 after two hours of tamoxifen addition, and (c) nuclear DDIT3 after eight hours of tamoxifen addition. The functional categories are divided into subgroups for which p-values for enrichment are given.

Supplementary Table 3. Enrichment of transcription factor binding sites. Enrichment of predicted transcription factor binding sites (TFBS) in the promoters of genes regulated by nuclear DDIT3 after two hours of tamoxifen addition. P-values reported for both a Fisher test and a permutation test. A predicted TFBS in a promoter of a gene is referred to as a "hit".

Supplementary Table 4. Primer sequences used in RT-qPCR experiments.

Supplementary Film Clip 1. DDIT3morEGFP cells during 60 minutes incubation with tamoxifen (added after 5 minutes).

Supplementary Film Clip 2. morEGFP cells during 60 minutes incubation with tamoxifen (added after 5 minutes).

Supplementary Methods

Cell migration assay modeling

The experimental setup for the migration assay is described in the main text (six replicates for three separate experiments were performed). The assay differs from traditional scratch wound assays by the use of two cell types that are seeded and co-migrate together. The use of wild type cells as reference in the experiments decreases experimental variability and possible contributions of cell proliferation to the migration assay differences.

To model the migration rates (probabilities) for the different cell types (EGFP, DDIT3-EGFP, DDIT3morEGFP, and wild type), we make the following assumptions. Let the different proportions of EGFP-stained cells to wild type cells in the three experiments be denoted by $\Pi=(\pi_1, \pi_2, \pi_3)$ and let M denote the event that a cell has migrated into the scratch wound. From each experiment i we can estimate the conditional probability that a cell is green given that it has migrated: $p^{(i)}{}_{G|M}= P(G^{(i)}|M)$ simply as the proportion of green cells in the scratch(es). What we wish to investigate is the migration rate for the EGFP-stained cells in each experiment, i.e. the probabilities $P(M|G^{(i)})$. We denote the migration probability for wild type cells with p and the migration probability for transfected cells with $p+\delta^{(i)}$ in experiment i. We can now apply Bayes theorem.

$$\begin{aligned} p^{(i)}_{G|M} = P(G^{(i)}|M) &= \frac{P(M|G^{(i)})P(G^{(i)})}{P(M|G^{(i)})P(G^{(i)}) + P(M|W^{(i)})P(W^{(i)})} \\ &= \frac{(p+\delta^{(i)})\pi_i}{(p+\delta^{(i)})\pi_i + p(1-\pi_i)} \end{aligned}$$

where *W* denotes wild type cells and *G* transfected (EGFP-stained) cells. With some algebra we can by rearranging the terms in the above equality deduce

$$\frac{p}{p+\delta_i} = \frac{\pi_i(1-p^{(i)}_{G|M})}{(1-\pi_i)p^{(i)}_{G|M}}.$$

We estimate this quantity for all replications in each of the three experiments by plugging in the estimates for π_i and $p^{(i)}{}_{G|M}$.

If we wish to compare the migration probabilities of EGFP, DDIT3-EGFP, and DDIT3morEGFP to wild type cells we can use these ratios and assess significance with a sign test. By noting how many of the ratios are above one, we deduce that DDIT3morEGFP and DDIT3-EGFP cells migrate slower than wild type cells, while EGFP cells migrate faster (all p-values ~0.03) as indicated in Figure 3 in the main text.

To compare the migration rates of DDIT3-EGFP and DDIT3-MOR-EGFP with that of EGFP cells, we use all pair wise quotients of ratios deduced above for a given comparison. We estimate all 36 ratios for each of the comparisons as

$$\widehat{\frac{p+\delta^{(2)}}{p+\delta^{(1)}}} = \frac{\hat{\pi}_1(1-\hat{p}^{(1)}_{G|M})}{(1-\hat{\pi}_1)\hat{p}^{(1)}_{G|M}} \Bigg/ \frac{\hat{\pi}_2(1-\hat{p}^{(2)}_{G|M})}{(1-\hat{\pi}_2)\hat{p}^{(2)}_{G|M}}$$

By noting how many of the quotients are larger than zero, we get the Mann-Whitney U statistic and we can use the Mann-Whitney U test (equivalent to a Wilcoxon rank sum test). The hypotheses we test are whether observations from one population exceed the observations from another population or not, i.e. if the migration probability in one group exceeds the migration probability in another group.
All 36 quotients of ratios between EGFP and DDIT3-EFGP, likewise for EGFP and DDIT3morEGFP, deviate from one in the same direction. This corresponds to the most extreme outcome of the statistic and gives a p-value of approximately 0.002. For the ratio between DDIT3mor EFGP and DDIT3-EGFP migration probabilities, we observe a p-value of 0.065. We can hence deduce that the migration probabilities between DDIT3morEFGP and DDIT3-EGFP most likely differ from the migration probability of EGFP cells, but we cannot on the 0.05 level claim that the migration probabilities are different in the two DDIT3 groups.

Permutation test for TFBS enrichment

The test to detect enrichment for TFBS among the regulated genes was based on a weighted statistic and significance assessed with permutation. We assume that we have expression values for a set of genes in two conditions. The genes are ranked for differential expression using for example log-fold change, or the moderated t-statstic. These gene level statistics are denoted by d_g. We also have a set of scores for the occurrence of motifs in the promoter of each gene. The indicator I_{gj} equals 1 if gene g contains motif j in its promoter and 0 otherwise. We use the following test statistic

$$u = \sum_g w_{gj}(d_g) \cdot I_{gj}$$

with weights w_{gj}. The weights score the values of the gene level between 0 and 1. If a gene is highly differentially expressed, it receives a score close to 1, otherwise it should receive a score close to 0. We use a logistic curve for the weights, for which we can vary the location and scale parameters according to the gene expression data.

If a motif is present in the promoter of several differentially expressed genes, the weights will be closer to 1 for these genes, and the test statistic u should be “large”. Conversely, if a motif is rarely seen in the promoters of the differentially expressed genes, it results in a small value of u.

The significance of motif occurrence and high differential expression is tested with permutation on the indicators I_{gj}. The motif occurrence is permuted 1000 times and the value of the test statistic u is calculated for each permutation. The p-value for enrichment of motif j among the differentially expressed genes is calculated as

$$P = \sum_p I(u_p > u)$$

where u_p denotes the value of the statistic in permutation p.

The parameter values for the location and scale parameters in the weigh functions have to be chosen by the user, but we recommend setting the location parameter to roughly the 80%-quantile of the gene level statistics.

We compared our method with another common permutation procedure called Gene Set Enrichment Analysis (GSEA)[1] using a simulation study previously described[2]. Briefly, the expression for 600 genes in 20 samples was simulated using a multivariate normal distribution (all with variance 1). 520 genes constituted the background set, and were simulated with a mean $\mu = 0$ and correlation $\rho = 0$. The remaining 80 genes were simulated with different means and correlations mixed of values $\mu = (0.75, 1, -1)$ and $\rho=(0, 0.6, -0.6)$. Nine sets were used to test the enrichment methods, of which sets 1, 2, 6, and 7 should be detected by any well working method, and sets 4, 5, 8, 9 ideally also should be detected (although only half of the genes were differentially expressed in these sets). Set 3 should work as a negative control[2].

We simulated 100 data sets, ranked the genes in each data set by $\log_2$ fold-change (absolute values) as well as by the moderate-t statistic (also absolute values), and tested each method on these sets. Our method was tested with three different values on the location parameter, corresponding to the 75, 80, and 85 percentiles of the gene level statistics. The scale parameter was set to 0.1.

[1] Subramanian et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 102(43):15545-50.

[2] Ackerman and Strimmer (2009). A general modular framework for gene set enrichment ananlysis. *BMC Bioinformatics.* 10:47.

	log-fold change	moderated-t
set 1	0.62	0.63
set 2	0.93	0.93
set 3	0	0
set 4	0.47	0.47
set 5	0.45	0.43
set 6	0.89	0.89
set 7	1	1
set 8	0.71	0.73
set 9	0.85	0.84

Table 1: Enrichment results from GSEA. The values correspond to the proportion of p-values < 0.05 in the 100 data sets.

	log-fold change (1)	log-fold change (2)	log-fold change (3)	moderated-t (1)	moderated-t (2)	moderated-t (3)
set 1	0.7	0.67	0.64	0.7	0.66	0.63
set 2	0.99	0.99	0.99	0.99	0.99	0.98
set 3	0	0	0	0	0	0
set 4	0.56	0.55	0.51	0.57	0.52	0.5
set 5	0.52	0.56	0.55	0.52	0.52	0.49
set 6	0.92	0.92	0.88	0.92	0.9	0.86
set 7	1	1	1	1	1	1
set 8	0.78	0.78	0.76	0.77	0.75	0.75
set 9	0.89	0.96	0.95	0.86	0.93	0.9

Table 2: Enrichment results from our permutation test. The values correspond to the proportion of p-values < 0.05 in the 100 data sets. The location parameter was set to 0.6, 0.68, and 0.77 for the log-fold change ranked data and to 1.35, 1.52, and 1.73 for the data ranked with the moderated-t statistic. The scale parameter was set to 0.1.

We observe that the results for the permutation test performs slightly better than GSEA on all data sets. The results seem to be quite robust to the choice of the location parameter. The scale parameter can also be varied, an influences how sharply the logistic curve switches from values close to zero to values close to one. The results seem quite robust also to the choice of this parameter (data not shown), but we recommend values in the range 0.05 - 0.2. According to our simulations, a good choice for the location parameter is in the range given above (75-85th percentiles of the gene level statistics).

Our permutation method is very easy to implement and the better performance of our statistic u to the running sum in the GSEA is probably due to the fact that our statistic is not sensitive in the same way to the absolute gene ranking. Although there is a need to choose the extra location and scale parameters, our method offers more versatility in how the expression values are allowed to influence the results (we can choose to only use highly differentially expressed genes, or be more liberal and allow genes with moderate expression values to also influence the statistic). We can choose to apply the weights (the logistic curve) to absolute values of the gene level statistics, or to the original values.

For the motif enrichment p-values presented in the paper (Supplementary Table 4), we ranked the genes by absolute $\log_2$ fold-change and chose the values 0.75 for the location parameter and 0.1 for the scale parameter. We also tested the method on the down regulated genes, with similar negative results (data not shown).

Paper IV

TRANSCRIPTIONAL AND METABOLIC DATA INTEGRATION AND MODELING FOR PATHWAY IDENTIFICATION

WORKING PAPER

ALEXANDRA JAUHIAINEN[1,2], OLLE NERMAN[1,2], GEORGE MICHAILIDIS[3], AND REBECKA JÖRNSTEN[1,2]

Abstract. With the growing availability of 'omics' data generated to describe different cells and tissues, the modeling and interpretation of such data has become increasingly important. Pathways are sets of reactions involving genes, metabolites, and proteins highlighting functional modules in the cell. Therefore, to discover activated or perturbed pathways when comparing two conditions, for example two different tissues, it makes sense to use several types of 'omics' data. We present a model that integrates transcriptomic and metabolomic data in order to make an informed pathway level decision. We view the gene expression data as explanatory for the metabolite data model, since metabolites can be seen as end-points of perturbations happening at the gene level. With real data, we show that the transcript profiles can be used to explain the metabolite data, and with simulations that the proposed model offers a better performance in identifying active pathways than for example enrichment methods performed separately on the transcript and metabolite data.

1. Introduction

The development of different 'omics' technologies in molecular biology have resulted in many ways to characterize cells and tissues. In addition to the complete sequence of the genome, the overall collection of gene transcripts (the transcriptome), proteins (the proteome), and metabolites (the metabolome) can be investigated with various techniques. The end purpose of such a characterization is to find genes, metabolites, and proteins that constitute networks which are perturbed in certain tissues or cell states.

Taking cancer as one particular example, cancer tumors originate from alterations in the DNA sequence of cells which transform them into cancer cells. These alterations range from point mutations (alterations in a single nucleotide) to large chromosomal aberrations [1]. Whole-genome microarrays can be used to monitor changes in the transcriptome, e.g., when comparing samples from cancer patients and matched normals. Recently, more efforts have been focused on understanding the metabolome of cancer cells, in order to

[1] Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden
[2] Mathematical Statistics, University of Gothenburg, Göteborg, Sweden
[3] Department of Statistics, University of Michigan, Ann Arbor, MI, USA
E-mail addresses: alexandra.jauhiainen@chalmers.se, nerman@chalmers.se, gmichail@umich.edu, jornsten@chalmers.se.

gain understanding of the cellular physiology and biochemical activity of tumors. An example is given in [2] where the authors investigate the metabolome in prostate cancer progression.

A metabolic profile holds information on concentrations of different small molecules (metabolites) in the cell, for example sugars, amino acids, organic acids, and vitamins, in contrast to a transcriptional profile that characterizes mRNA transcript levels. Metabolic profiles are generated via techniques like GC-MS (gas chromatography, followed by mass spectrometry) or very commonly NMR (nuclear magnetic resonance). Identifying metabolites from NMR or GC-MS spectra is a difficult task, and the resolution is usually 50-150 unique molecules, significantly less than the number of transcripts identified in an array experiment. Since alterations on the genomic level often manifest themselves as downstream variations in metabolite concentrations, the information on the (relatively few) metabolites is still very important. The metabolites can be viewed as end-points of perturbed pathways, often referred to as altered or active pathways [3].

The most common analysis of transcriptional data involves some type of functional annotation of differentially expressed genes. To this end, genes are mapped to functional categories representing different cellular functions (or other groupings), and the functional categories are analyzed for enrichment among the differentially expressed genes. The Gene Ontology (GO) [4, 5] is a widely used resource for functional annotation and contains controlled and consistent vocabularies for annotation of gene products. The three main ontologies consist of a terminology describing the molecular function of gene products, their associated biological processes, and their cellular localization. However, as the name implies, GO contains annotation for gene products only, and not for metabolites.

Pathways, which are sets of reactions involving genes, metabolites, and proteins, can be viewed as functional groups with a more complicated structure than GO groups. The reactions connect different metabolites and genes in an intricate network. The KEGG database [6, 7] as well as the MetaCyc initiative [8, 9], are collections of pathway information for a large set of species. The pathway information can be used to perform enrichment analysis with transcript, proteome, or metabolite data in the same way as annotation to GO groups. As the availability of different transcriptomic data sets increases, the combination of such data sets can be used to improve the power of enrichment analysis in pathways [10].

Enrichment analysis of functional groups or pathways can be done in multiple ways. Generally, using the terminology in [11], performing enrichment with gene expression data consists of three steps. The first being individual gene scoring with e.g. fold-change or t-statistic, followed by transformations of the calculated gene levels statistics to e.g. p-values or ranks. Finally, an overall gene set statistic, which could be e.g. a sum, a median, or any suitable summary statistic of the transformed gene level statistics, is calculated and significance assessed. Over 250 methods for enrichment analysis are compared in a general framework in [11].

A standard and simple form of enrichment analysis is the hypergeometric test (Fisher test) in which the proportions of genes in a gene set within the groups of differentially and non-differentially expressed genes respectively are compared. Other common choices are the more sophisticated Gene Set Analysis (GSA) [12] and Gene Set Enrichment Analysis (GSEA) [13] methods that use the maxmean statistic and a running sum respectively, as the gene set statistic.

The availability of multiple 'omics' data sets in an experiment raises the question on how to combine the data sets in an enrichment analysis context. For pathways in particular, a combined analysis of transcript and metabolic data may be very informative, as the metabolites can be viewed as end-points of perturbations caused by changes in the expression of key genes in the same pathway predominantly.

Previous studies working with an integrative approach for different types of 'omics' data include several studies on plants, particularly *Arabidopsis thaliana* [14, 15, 16], tomato [17], and hybrid aspen [18], a study on the yeast *Saccharomyces cerevisiae* [19], and studies on rat or mouse [20, 21, 22].

The two main approaches for data integration in these studies are correlation analysis and/or unsupervised multivariate techniques like PCA and PLS (see e.g. [23]). In the correlation approach, significant correlations between transcripts and metabolites are extracted, followed by clustering and network visualization or functional annotation. The functional annotation is generally limited to association of genes to GO groups (not using the metabolites). Other studies rely on unsupervised techniques like PCA and PLS to extract connections between the metabolome (not always using detailed metabolite data) and the transcriptome. The connections can then be investigated by network visualization.

An example of a more mathematical approach for data integration includes the sparse PLS methodology which was adapted to select a subset of important genes to explain transcriptomic data [24].

In general, in studies published to date, the individual connections between transcripts and metabolites are integrated into a pathway/functional group decision. We propose a different approach, in which we adapt a global model with the direct purpose of making decisions on active or enriched pathways/functional groups as opposed to the two-step individual modeling. We model the metabolite and transcript data jointly to make an informed pathway decision, and hence the interpretation of the data becomes more straightforward since the need for post-analysis of identified metabolite/transcript connections is reduced.

The final goal of our analysis is to highlight pathways in which there is a considerable difference between treatment groups manifested both in mRNA transcription profiles and on the metabolite level.

2. Pathway Model of Transcript - Metabolite Connections

We propose a global model with the purpose of making decisions on active or enriched pathways by joint modeling of transcript and metabolite data. The experimental setup that we base our model on is the availability of transcript and metabolite data for two groups, which we for simplicity call treatment and control. Examples of this is cancer and matched normal, or mutant and wild-type. A generalization to the model is to e.g. include more treatment groups, which we touch briefly upon in Section 6, but in this paper we focus on two treatment groups.

We index the pathways used in the model by k=$1, \ldots, K$. The indicator variable c_k equals one if pathway k is active, and equals zero otherwise. The purpose of the modeling and estimation is to do sparse selection on the c_k indicators.

2.1. **Prior information.**

The global model contains the pathway information in the form of membership for genes and metabolites in the different pathways. We denote these memberships with indicators a_{ik} and b_{jk} such that $a_{ik} = 1$ if gene i is in pathway k and $b_{jk} = 1$ if metabolite j is in pathway k for genes indexed by $i = 1, \ldots, N_g$ and metabolites indexed by $j = 1, \ldots, N_m$.

2.2. **Data and Model Formulation.**

We let the control and the treatment groups in the data be indexed by $t = 1, 2$, respectively. The expression of metabolite j in condition t is denoted by f_{jt} and the expression of gene i in condition t by g_{it}. We distinguish the control and treatment groups by the indicator variable x_t which is equal to zero for the control group and equal to one for the treatment group; $x_1 = 0, x_2 = 1$. The gene level of our model specifies a dependence of the gene expressions on pathway membership:

$$g_{it} = \alpha_i + x_t\, \beta_i \left(1 - \prod_k (1 - a_{ik}\, c_k)\right) + \varepsilon_{it}$$

The gene model includes an intercept term denoted by α_i while the second term with the parameters β_i represents a direct effect from a potential pathway membership. If gene i is a member of one or more active pathways, the direct effect can be included in the model for the observations from the treatment group. The model selection procedure (see below) determines if $\beta_i \neq 0$ for gene i.

Similarly, the metabolite model includes an intercept α'_j for each metabolite.

$$f_{jt} = \alpha'_j + \sum_i \delta_{ij}\, g_{it} \mathbf{1}\{\beta_i \neq 0\} \left(1 - \prod_k (1 - a_{ik}\, b_{jk}\, c_k)\right) + \varepsilon'_{jt}$$

The second term in the metabolite model is included to account for a potential effect of gene expression on the metabolite expression. The expression of metabolite j can be affected by gene i if both are members of the same pathway, provided that the direct effect for gene i was included in the gene model. The parameters δ_{ij} estimate the relationship between gene i and metabolite j. Please note that the biological replicates in each treatment group are not indexed in the model equations stated above.

2.3. **Model Selection and Parameter Estimation.** Model selection occurs on two levels in the estimation procedure; within pathways and on the global pathway level. Within pathways (for pathway k), we firstly select genes comparing the null and non-null gene models. For a specific gene i this means that we compare the model using only the intercept parameter α_i with the model where the direct effects parameters β_i are included as well. For each gene within pathway k, BIC is calculated for both the null and non-null models. A rate-distortion criterion is used to select the genes for which we use the non-null models, see Appendix B. The rate-distortion slope minimizing the overall BIC (sum of individual gene BIC:s) is chosen and models selected based on that slope. In using BIC to select models, we make the assumption that the model errors are Gaussian.

For the model selection within pathways on the metabolite level, we allow genes within pathway k with an included treatment effect (i.e. $\beta_i \neq 0$) to work as predictors. The number of genes may be large (and the number of replicates small), and hence some regularization is needed to select which genes to include as predictors. An elastic net penalty [25] with a high level of sparsity imposed to select predictors, see Appendix A.3. For each metabolite we have the null model (only intercept α'_j) and a set of increasingly complex models indexed by inverse of the penalty parameter λ (large penalty results in small models). To select which model to use for each metabolite, we again use rate-distortion criterion, but with a cross-validation selection method instead of BIC. In effect, we wish to maximize the overall predictive likelihood (instead of minimizing BIC which seems to be unstable in this context.

The model selection on the global pathway level is made via a stepwise procedure. In each step, a pathway is chosen and added to the active set of pathways according to an R^2 criterion. For each pathway, the R^2 values for the gene and metabolite models are combined to an overall R^2 value and the pathway with the largest R^2 is added (for details, see Appendix A.2).

For subsequent steps in the pathway selection procedure, the residuals from the metabolite model from previous steps are used as responses, while genes are not allowed to be re-used once they have been included in the model. A detailed description of the whole estimation procedure is given in Appendix A.

3. Simulations and Application to the NCI-60 data

3.1. **The NCI-60 Data Set.** The NCI-60 is a set consisting of 59 human cancer cell lines derived from various tissues, and characterized into nine broad cancer categories; leukemia, non-small cell lung (NSCLC), colon, CNS, melanoma, ovarian, renal, prostate, and breast cancer. The cancer cell line set has been used for extensive screening of chemical compounds, and has also been characterized by gene expression profiling, with

CGH copy number arrays, by microRNA expression and metabolite profiling, just to mention a few. The gene expression levels have been analyzed multiple times using different platforms, for example Affymetrix, as in the most recently published (as of April 2010) Chiron data. Metabolite concentrations have been characterized only once so far with triplicate technical replications [26].

3.2. **Data Preprocessing.** The cell lines belonging to the NSCLC and Leukemia groupings were chosen for the simulation study. Since control cell lines are not included in the NCI-60 set, the six Leukemia cell lines were chosen to act as controls, while the nine cell lines in the NSCLC group were considered to be the treatment group, but any grouping is of course possible.

The gene expression values were averaged over multiple probe IDs in each cell line and features missing a Unigene ID were filtered out. Features with the same Unigene ID were averaged to produce a final set of 17118 gene expression values. The metabolite data contained characterization of levels for 352 metabolite features. The subset of features uniquely identified as a specific metabolite were selected. Information on all metabolites contained in the KEGG database [7, 6] were downloaded (January, 2010) via the Taverna tool [27]. Systematic metabolite names were matched via fuzzy matching and manual curation to unique KEGG compound IDs resulting in total 136 compounds to be retained.

3.3. **Pathway Information.** Pathway information for 199 human pathways were downloaded from KEGG (January, 2010) through Taverna. The collective metabolic pathway (with more than 1000 and 80 metabolites) was removed. All genes and metabolites taking part in the pathways were identified and mapped (creating the indicators a_{ik} and b_{jk} defined in Section 2). Pathways not containing any metabolites were excluded from the modeling (although it is possible to include them in the gene model only) since the integration of the different data sources was the main focus of the modeling. A total of 75 pathways were used in the simulations. After mapping of metabolites and genes to the pathways, a total of 110 metabolites and 4526 genes were kept, see Table 1.

Table 1. Summary of the filtered NCI-60 data and pathway information.

	Number of features	Example
Genes	4526	hsa:2561
Metabolites	110	cpd:C00249
Pathways	75	path:hsa00471

3.4. **Simulation Study.** The NCI-60 data set (Leukemia and NSCLC groups) was employed for simulation and evaluation of the methodology in the following way. First, the data in its original form was used to assess the overall performance of the model and to judge the predictive power in the gene expression data on the metabolite expression data. In the case of poor predictive power in the gene expression data, the overall rate-distortion model selection criterion should frequently pick the null models, or models with few parameters.

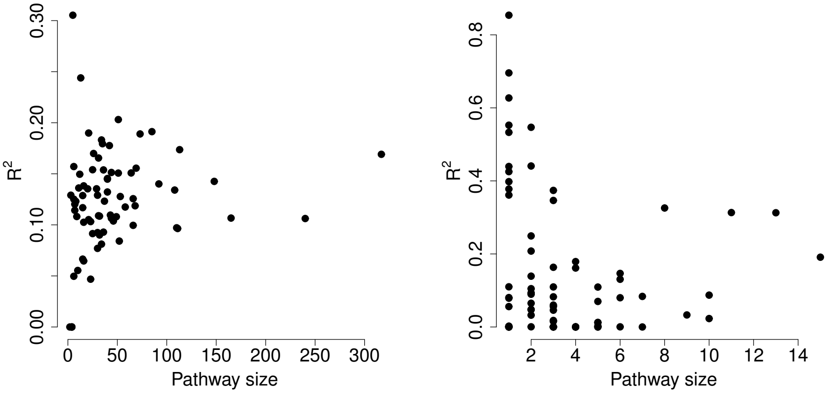
Second, three simulated data sets were created with the aim to mimic different scenarios of pathway activation; one scenario in which a majority of the genes and metabolites are perturbed between the treatment and control groups, one having a small set of genes with a strong signal, and similarly a strong signal on correlated metabolites, and one in which we observe differential expression between the groups on half of the genes and metabolites. We refer to the three data sets as "original", "all active", "a third active", and "half active".

In detail, five non-overlapping pathways with varying sizes were selected to work as active pathways in the three simulated data sets. For the first simulated data set ("all active"), the precision in the measurements for all member genes and metabolites was increased by a factor of four element-wise within treatment groups. Any existing differences between the two treatment groups were thus increased for the subset of genes and metabolites in the selected pathways. In the second simulated data set ("a third active"), the precision was increased (similarly as in the all active simulated set) within each selected pathway for a third of the genes (chosen at random), while the remaining genes had their precision decreased by a factor of two. Half of the metabolites within each pathway (which had the highest correlation to the genes with increased precision) were spiked in the same way. The precision of the remainder of the metabolites was decreased. For the third simulated set ("half active"), half of the genes and metabolites in each selected pathway were chosen at random and their precision was increased by a factor of four, while the remaining genes and metabolites levels were left unchanged.

The first round in the estimation procedure aims to identify the most important pathway according to the R^2 scoring scheme. The R^2 criterion is defined within each pathway, and since the number of parameters used within each pathway is optimized with rate-distortion, the R^2-criterion should not pick only large pathways.

In fact, we see in Figure 1 that small pathways generally are scored highly, and especially in the metabolite model. Small pathways can be easily explained in R^2-sense in the metabolite model if they contain just one or two metabolites, and we happen to have a strong signal on those genes in the data. Perhaps such small pathways can be excluded from the analysis, but we have chosen to keep them in the analysis, and since the gene model also influences the pathways selection, spurious signals in the metabolite model are down-weighed somewhat.

For the original data, we observe that the expression values of genes seem to have some predictive power on the metabolite expression, as non-null models are picked frequently, which motivates our model formulation. Similar plots for the simulated data sets for the first round of estimation is given in Appendix C. Figure C.1 and Figure C.2 show the R^2 scores for the gene and metabolite models separately with the spiked pathways marked with red squares. The original NCI-60 data is also depicted for comparison. The general trend is that the spiked pathways have a high rank in the simulated sets



(a) Gene model

(b) Metabolite model

FIGURE 1. Overall R^2 scores for each pathway versus pathway size in the gene and metabolite level models on the original NCI-60 data.

especially on the gene model. The exception is pathway `hsa:00400`, and after closer inspection we observed that this pathway contains genes exhibiting high variance and inconsistent expression values within the two treatment groups. Hence, the R^2 is also very low in metabolite model for this pathway. For all the other pathways, the increase in precision on the member genes also leads to higher R^2 values in the metabolite model.

Running the estimation method for the model several rounds gives a set of pathways with the highest rank. The choices of when to stop the stepwise procedure are several; after a given set of rounds, when the overall R^2 falls below some cut-off, or when the overall R^2 falls below a certain percentage of the R^2-value of maximum ranked pathway.

For the original NCI-60 data, none of the spiked pathways are ranked among the top ten pathways, indicating that higher ranks within the simulated data sets show that our spiking procedure makes sense. For the simulated data sets, the results are given in Table 2 where the ranks for the pathways are ordered according to `hsa:00410, hsa:00400, hsa:00061, hsa:00230, hsa:00562`.

From Table 2 we observe that the model performs well on the all active and half active sets, but slightly worse on the a third active set. Closer inspection shows that for example the top ranked pathway on set 2 is a small pathway which partly overlaps with one of the spiked pathways, indicating that the model still picks up the spiked signal.

The performance of the model on the different simulated sets, as well as the original data can be further investigated by depicting the selected pathways in a rank-rank plot. In Figure 2, the ranks for the different pathways in the first round of estimation for the gene

Table 2. Ranks (according to R^2) for the active pathways for the three simulated data sets in the NCI-60 data. The stopping criteria was chosen to a fixed set of 10 estimation rounds.

Data set	Spiked pathway ranks (combined)[a]
NCI-60 Original	-, -, -, -, -
NCI-60 all active (set 1)	6, -, 1, 4, 3
NCI-60 a third active (set 2)	-, -, 3, 9, -
NCI-60 half active (set 3)	5, -, -, 2, 3

[a] A dash indicates that the pathway was not present among the top ten ranked pathways.

and metabolite models are depicted on the x- and y-axes. The top ten selected pathways are highlighted with red squares.

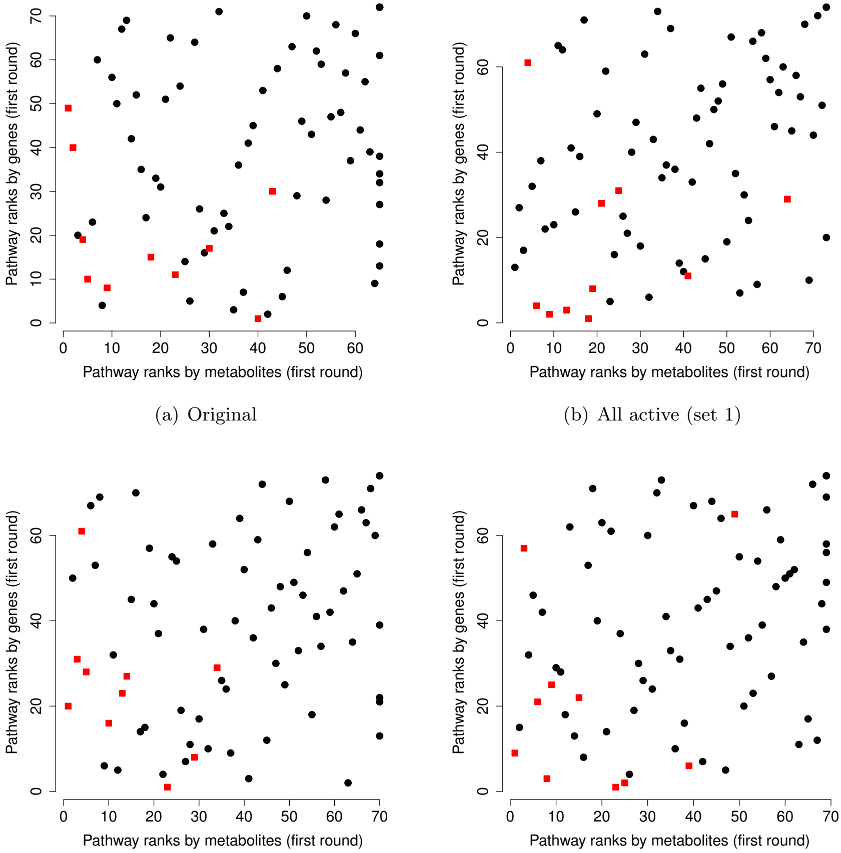
The rank-rank plots show that pathways with high ranks in both the metabolite and gene models preferably are picked. However, pathways with a strong signal in for example the gene model, and a weak signal in the metabolite model can also be picked, due to the R^2 selection criterion. One possibility is to exclude such pathways from the selection process, and select pathways under an additional criterion on the maximum of the ranks. However, such a solution is quite ad-hoc and we chose to keep the simple R^2 criterion for selecting pathways.

4. Comparisons with Other Methods

The purpose of the model we present in this paper is to integrate transcriptomic and metabolomic data, and analyze the data with respect to pathways. The current methods for analysis and integration of such data is usually (gene) set enrichment, or different versions of correlation analysis.

4.1. Gene Set Enrichment.

Gene set enrichment on the transcript and metabolite data is not an integrated approach per se, but we can analyze the different data types separately and try to combine the two analyses. How to combine enrichment p-values is not obvious though, and we don't attempt to do so here. Instead we report the p-values for the two analyses separately. The metabolites and genes in all four data sets were ranked with respect to differential expression using the moderated t-statistic [28]. Pathway enrichment analysis was done separately on the gene expression and metabolite data within each NCI-60 data set with the GSEA method [13] and the GSA methods [12]. P-values were calculated with a permutation test based on 1000 permutations of gene and metabolite pathway membership.



(a) Original

(b) All active (set 1)

(c) A third active (set 2)

(d) Half active (set 3)

FIGURE 2. Ranks in the first round of estimation for all the NCI-60 data sets. The top ten picked pathways are indicated with red squares. It should be noted that the ranks for the pathways change between the different rounds, as genes are sequentially removed, and the residuals for metabolites from previous rounds of estimation are used as responses.

In the tables below, the pathways are ranked according to p-value. When applicable, ties in the ranking have been resolved by giving pathways with equal p-values the same minimum rank. The ranks for the pathways are ordered according to `hsa00410, hsa00400, hsa00061, hsa00230, hsa00562`.

Table 3. Ranks (according to p-value) given by GSEA for the active pathways in the four data sets from the NCI-60 data. In comparing with Table 2, please note that for our model, we integrate the two data sources, and hence present an overall ranking.

Data set	Spiked pathway ranks [a] (genes)	Spiked pathway ranks [b] (metabolites)
NCI-60 Original	9, 2, 33, 40, 14	3, 35, 17, 20, 31
NCI-60 all active	1, 28, 15, 1, 1	1, 5, 10, 2, 12
NCI-60 a third active	1, 24, 18, 16, 2	1, 12, 4, 2, 8
NCI-60 half active	6, 24, 42, 1, 1	1, 17, 9, 3, 26

[a] Number of p-values < 0.01: 1, 12, 5, 11, respectively.

[b] Number of p-values < 0.05: 4, 4, 5, 4, respectively.

The simulated data sets were spiked in such a way to increase the differences in the gene and metabolite models between groups. If this difference is present among many of the genes/metabolites within one pathway, GSEA should perform well. We observe that GSEA does fairly well on the all active data set. The worst performance of GSEA on the gene model and is within the a third active set. Since just a third of the genes are spiked within each pathway, the signal is hard for the algorithm to pick up. GSEA fails in picking up the signal among the genes belonging to pathway `hsa00400`, as is natural due to the conflicting expression values. Enrichment of pathway `hsa00061` is neither detected well, probably due to the smaller size (in the gene expression data) of that pathway.

Table 4. Ranks (according to p-value) given by GSA and the maxmean statistic for the active pathways in the four data sets from the NCI-60 data.

Data set	Spiked pathway ranks [a] (genes)	Spiked pathway ranks [b] (metabolites)
NCI-60 Original	20, 62, 56, 74, 42	6, 32, 7, 8, 34
NCI-60 all active	1, 67, 44, 1, 1	1, 12, 5, 3, 13
NCI-60 a third active	1, 62, 45, 67, 25	1, 18, 8, 1, 11
NCI-60 half active	1, 66, 54, 1, 19	3, 28, 7, 2, 17

[a] Number of p-values < 0.01: 9, 15, 20, 21, respectively.

[b] Number of p-values < 0.05: 13, 10, 9, 9, respectively.

The maxmean score in GSA is very good at picking up one-tailed expression changes, i.e., when the genes or metabolites are perturbed between the two treatment groups mainly in the same direction. However, it performs less well in the two-tailed situation. Also, the maxmean statistic is designed to avoid situations where few strong signals could dominate the score [12]. Thus, this may be the reason that it performs poorly on the a third active data set, especially on the gene-level where only a third of the genes are spiked. GSA seems otherwise to perform similarly to GSEA.

4.2. **Correlation Analysis.** The correlation analysis for integration of the gene expression and metabolite data (partly adapted from [17]) was done as follows. Differential expression for the genes in all four data sets was assessed using the moderated t-statistic, and p-values were adjusted with the Bonferroni-Holm (BH) method [29] to account for multiple testing. Genes with a p-value below a given threshold were selected and the corresponding pairwise Pearson correlations were calculated with each metabolite (over both treatment and control groups simultaneously). Prior to correlation analysis, both the genes and metabolites were centered to have mean zero and scaled to have a variance of one.

Significance for correlations between all gene-metabolite pairs was assessed and p-values adjusted with the BH-criterion. All gene-metabolite pairs with a correlation p-value lower than the given threshold were selected and clustered with single-linkage clustering to create groups of co-correlated metabolites and genes. The clustering procedure was adopted to exclude spurious small clusters of co-correlated metabolite-gene pairs (only adding noise). The largest cluster (which either was the only one, or by far bigger than the other clusters) among the co-correlated genes and metabolites was therefore used for enrichment analysis. The enrichment of pathways within the genes and metabolites was assessed separately with a hypergeometric test (equivalent to a Fisher test). The p-value threshold was set to 0.01 for all data sets.

The results for the analysis on all four data sets are given in Table 5. We would expect the method to pick up at least some of the signal in the gene data set since genes with a differential expression between the treatment and control groups were chosen in the first step and hence are present in the clustered groups. Within the gene expression data, three of the active pathways received a relatively high rank for all the simulated data sets. However, the ranks, especially not for the a third active set, are not impressive. The method seems to perform very well on the metabolite data, although the fact that the gene-metabolite connections may be induced by between-pathway correlations, instead of correlations within pathways, may raise concerns (see discussion in Section 5).

Table 5. Ranks (according to p-value) given by correlation analysis for the active pathways for the three data sets in the NCI-60 data. The ranks for the pathways are ordered according to `hsa00410, hsa00400, hsa00061, hsa00230, hsa00562`.

Data set	Spiked pathway ranks [a] (genes)	Spiked pathway ranks [b] (metabolites)
NCI-60 Original	48, 47, 29, 6, 3	23, 45, 21, 1, 14
NCI-60 all active	5, 55, 42, 1, 2	4, 3, 2, 1, 9
NCI-60 a third active	1, 54, 40, 2, 5	3, 16, 4, 2, 7
NCI-60 half active	2, 55, 42, 1, 3	2, 1, 5, 4, 7

[a] Number of p-values < 0.01: 0, 9, 5, 10, respectively.

[b] Number of p-values < 0.05: 0, 4, 1, 3, respectively.

5. Discussion

We propose a model for the purpose of identifying pathways (sets of reactions involving genes, metabolites and proteins) with altered behavior when comparing two different experimental conditions. The model is intended to make informed pathway level decisions by integrating gene expression and metabolite data.

Model selection is necessary on two levels in the model; within pathways and on the global pathway level. First, for the within pathway model selection, the model complexity in both the gene and metabolite models is chosen according to a rate-distortion criterion. The rate-distortion criterion works well in a situation in which a small set of features, for example genes, show a strong signal, and hence chooses non-null models for the features with a strong signal. If a larger set of features instead exhibits moderately strong signals, the rate-distortion criterion tends to select a subset of the features, but generally not all of them. However, according to our simulations, sufficiently many genes and metabolites seem to be selected for the model to function satisfactory (data not shown). The three simulation scenarios used in this paper attempt to mimic different levels of pathway activity with both strong and moderate levels of gene and metabolite signals (see below).

For the within pathway model selection in the metabolite models, we choose to use linear regression with an elastic net penalty. The elastic net has some appealing properties, and works better in a $p > n$ setting than for example the lasso [25]. The elastic net penalty (called α) is not chosen via cross-validation for each metabolite, but instead a global α-value is fixed to largely mimic the behavior of the lasso, but still allow for highly correlated genes to function as predictors together in the model by shrinking their coefficients towards each other. Undesirable lasso effects like model saturation and the restriction of the maximum number of included predictors (n) in the model can be circumvented by using an elastic net penalty. However, we experienced problems in using a modified BIC for selection of the penalty in the elastic net (the same problems occurred with the lasso) and thus recommend that the model be run with cross-validation techniques within the metabolite model instead.

To make a global pathway level decision, we select pathways in a stepwise procedure according to an R^2-based criterion. Some limitations are inherent in a stepwise procedure to select predictors. The presence of correlated predictors, i.e., in our case overlapping pathways (common member genes and metabolites), may be a problem. If two highly correlated pathways are active when comparing two conditions, the stepwise procedure tends to choose one of the pathways, and leaves the other outside the active set. One solution is to redefine the residuals after each round of estimation, allowing genes and metabolites in the selected pathway to be re-used if they are present within the set of genes and metabolites in another highly ranked pathway. However, the risk which such a procedure is that we keep selecting overlapping pathways at the expense of neglecting other pathways which may be of biological interest.

The R^2-based criterion for global pathway selection focuses on the amount of explained variance *within* each pathway. We combine the gene and metabolite level models by scaling the R^2-values to impose equal weights (or other predefined weights) between the gene expression and metabolite data. A BIC criterion to select pathways could be suggested, but in our experience the BIC criterion is unstable in this setting. This is due to the fact that each pathway is small compared to the size of the overall data sets, and the log-likelihood will therefore dominate the criterion resulting in the selection of null models.

Pathway selection is partly dependent on cross-validation (from the model selection in the metabolite model) and the procedure therefore suffers from some inherent variability. The variability only affects pathway ranking, since the same overall set of pathways almost always is selected. The differences we observe are mainly of the form where overlapping pathways "swap" places with each other in the active set of pathways.

Several definitions of altered behavior at the pathway level are possible. For example, a pathway can be classified as active at the gene level if only a few genes show strong differential expression between two conditions. Alternatively, activity could be thought to involve a larger set of genes showing moderate differential expression. The three simulation scenarios based on data from the NCI-60 data set attempt to mimic these different types of pathway activity.

The NCI-60 data is used as a test case to investigate the proposed model. We chose two cancer types, Leukemia and NSCLC, to act as control and treatment groups respectively, but any other choices of cell lines would have worked equally well. The approach with spiked (subsets of) pathway genes and metabolites, was motivated by our desire to generate a more controlled setting with known signals in the data (although the original data may of course be informative to differentiate between cancer types). To enhance existing signals in a real data also has the benefit of preserving correlations between genes and metabolites. One might argue that we have spiked the data "too much", thus making the pathway decision trivial. However, we show that e.g. GSEA cannot pick up the signal in all data sets, and thus our simulated scenarios function as proofs of concept for our modeling approach.

The aim of the proposed model, is to pick up correlations between the different data sources, and use these to make a pathway decision. For one of the pathways (`hsa:00400`), null models were frequently selected for the member genes. Hence, the pathway received a low R^2-score for the gene model, and subsequently also for the metabolite model. In contrast, with GSEA, GSA, or correlation analysis, we observe that the pathway (`hsa:00400`) is ranked highly (especially in the correlation analysis) in some of the simulated data sets. Since none of the genes within this pathway are scored highly among the differentially expressed genes, we conclude that with the correlation analysis, the metabolites within this pathway likely are included in the clustered set due to correlation with *other* (non-pathway) genes. Therefore, it seems likely that the correlation analysis is "too generous" in certain scenarios. However, it should also be noted that the correlation analysis is an ad-hoc method since no consensus on how to do integrated correlation analysis on transcript and metabolite data is reported in the literature. We conjecture that the method might be improved with more restrictive p-values and alternative clustering schemes.

We observe that both our model, GSEA, GSA, as well as the correlation analysis have problems in detecting the active pathways in the a third active (set 2) data set. This is probably due to the limited amount of spiked in signal in this data set, which makes the signal hard to pick up. It should be noted that when omitting the smallest pathways on the metabolite level (containing just one metabolite), all four spiked pathways are scored among the top ten for this data set with our model (while GSA and GSEA have the inherent property of not ranking such small sets highly in an enrichment context). this demonstrates the strength of using an integrated approach when there are different data sources available for analysis.

For GSEA, at least for the all active data set (simulated set 1), `hsa:00400` is scored among the top pathways in the metabolite data. Nevertheless, since the enrichment of this pathway on the gene level is non-significant, we would probably not consider this pathway interesting in a biological context. Our model makes this decision directly based on the fact that the genes cannot predict the metabolite expression. However, if we wish to find pathways with large differential expression on the metabolites, but not necessarily with correlations between gene expression and metabolite data within the same pathway, GSEA is a good choice for analysis.

The metabolite data in the NCI-60 set contains characterization of a relatively large set of metabolites, although many of the features, i.e. peaks in the spectra, could not be identified uniquely. Unfortunately it is often the case that missingness in the data is present. The missing metabolite concentration values can be imputed from the other replicates within treatment groups, which also has been done partially in the NCI-60. As an alternative to imputation our model and estimation procedure can handle moderate levels of data missingness, as long as there are a sufficient number of data points left to do estimation for each gene and metabolite (with of course larger variance in prediction as a result). Incompleteness in the pathway information also induces some problems in the modeling. In effect, we reduce the sizes of the pathways to match the detectable gene and metabolite sets, but this is the standard procedure in enrichment analysis.

6. Conclusions and Future Work

The model we propose in this paper aims to identify perturbed pathways by the integration of two different 'omics' data types. We integrate the pathway level decision in the modeling procedure, and show with simulations that the model generally performs better at identifying active pathways than for example enrichment methods performed separately on the transcript and metabolite data.

A potential application of the current model is to extend it to jointly analyze transcription factor binding site data coupled to gene expression or microRNA expression, for which a similar biological ordering is inherent. Another possibility is to generalize the model to handle several treatment groups and include more data sources, for example adding copy number variation data to the current setting with transcript and metabolic data. To extend the model to handle more than two treatment groups can be done in several ways, but one possibility is to re-parametrize the model to penalize contrasts of regression coefficients (for example in the metabolite model).

In the current work we focused on selecting predictors via an elastic net penalty in the metabolite model, but if more prior information is available concerning specific links between genes and metabolites, it is possible to further restrict the set of potential predictors that we allow to influence the metabolite expression. Such prior information can also be used to validate the predictors selected by the elastic net. We intend to implement the possibility of including such prior information in future version of the model estimation procedure.

Appendix A. Detailed Parameter Estimation Procedure

A.1. **Data Normalization.** Initial centering of the data was done by removing overall sample means for each gene and metabolite over all the replicates. The replicates for each gene and metabolite were also scaled to a variance equal to 1.

A.2. **Estimation Procedure.** A stepwise regression procedure in the c_k parameters were adapted. The outline of the procedure is as follows. For the initial round of estimation, let the indicator $c_k = 1$ for each pathway $k = 1, \dots, K$ in turn and perform steps 1-3 below for each k.

1. Find the set of genes with $a_{ik} \cdot c_k = 1$ (i.e., genes that are members of pathway k). Estimate the parameter β_i for these genes (using ordinary least-squares). Calculate the residual sum-of-squares SS_E and BIC for each gene for the null and one-parameter models. Calculate a set of rate-distortion slopes (see Appendix B) and for each slope, calculate the overall BIC for the chosen model. Choose the slope (and hence models) that give the lowest overall BIC (by summing the individual BICs). The result is a subset of genes in which β_i is estimated for each gene. For the remaining genes, the null model is chosen. Keep track of the number of parameters estimated and for which genes the non-null model was estimated.
2. Find the set of metabolites with $b_{jk} \cdot c_k = 1$ (members of pathway k). The genes in pathway k, i.e. the genes for which $a_{ik} \cdot b_{jk} \cdot c_k = 1$ are allowed to influence the metabolite expression, excluding the genes for which only the null model was estimated in step 1. Define a predictor matrix X consisting of the gene expression (the responses) for these genes. The matrix X can have a large number of predictors, and the number may exceed the number of replicated measurements (i.e., a $p > n$ situation), so some regularization is needed in order to model the metabolite responses as dependent on the predictors in X. Impose an elastic net constraint (we choose the elastic net parameter $\alpha = 0.8$) and estimate the solution paths for different levels of regularization (indexed by the penalty parameter λ). Calculate a sequence of rate-distortion slopes and choose the slope that minimizes the overall prediction error (over all metabolites) according to cross-validation.
3. Calculate the residual sum-of-squares SS_E, and the total sum-of-squares SS_T for both the gene and metabolite models. Calculate the coefficient of determination $R^2 = 1 - (SS_E/SS_T)$ for both models (giving R^2_g and R^2_m).
4. Calculate a combined coefficient of determination score R^2_{comb} for each k as

$$R^2_{comb}|_k = w\frac{R^2|^g_k}{\max_k\{R^2|^g_k\}} + (1-w)\frac{R^2|^m_k}{\max_k\{R^2|^m_k\}}$$

The weight parameter $0 \leq w \leq 1$, defines how much the gene and metabolite model influences the combined R^2, which is used to pick the most important pathway to add. If $w = 1$, only the gene model influences the choice, and conversely if $w = 0$, only the metabolite model does so. Equal contribution of the two data sources implies $w = 0.5$ as a sensible choice.

5. Pick the k, say k', with the dominating R^2_{comb}. Calculate the residuals for the genes and the metabolites. Note which genes have been included in the model (divide the genes into active and inactive sets). Remove k' from the set of pathways for which to do estimation in subsequent rounds.
6. Stopping the stepwise procedure. Multiple choices are available on how to stop the stepwise procedure. For example, after a fixed number of rounds, or after the combined R^2 for any pathway falls under a certain threshold.

For subsequent rounds of estimation, do the following modifications to the estimation procedure.

Step 1. Remove the genes which are already in the active set. Do the estimation on the remaining genes.

Step 2. For metabolite j, find the genes for which $a_{ik} \cdot b_{jk} \cdot c_k = 1$ and that were chosen according to rate-distortion as before. Remove any genes that are in the active set, *and* use the expression values for the remaining genes as columns in the predictor matrix X.

A.3. **Degrees of Freedom and the Elastic Net.** To employ BIC to do model selection, we need to know the degrees of freedom of the models we choose between. For a linear regression model with n observations and p predictors, $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, $i = 1, \dots, n$, the degrees of freedom for the total sum-of-squares is $n - 1$. For the regression sum-of-squares the degrees of freedom equals the number of predictors $p+1$ in the model.

For the gene model, the degrees of freedom for the regression sum-of-squares is the number of genes for which we include the intercept term separating the treatment and control groups.

The calculation of the degrees of freedom for the regression sum-of-squares for the metabolite model is more complicated since an elastic net constraint is imposed which results in an adaptively fitted model. In the elastic net regression model we wish to do the following optimization, with a mix of a ridge and lasso penalties.

$$\hat{\beta}_{elasticnet} = \underset{\beta}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \Big\}.$$

The degrees of freedom can be estimated by the following formula [30, 25].

$$\mathbf{H}_{\lambda_2}(\mathcal{A}) = \mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} + \lambda_2\mathbf{I})^{-1}\mathbf{X}_{\mathcal{A}}^T$$

and setting

$$\widehat{\mathrm{df}} = \mathrm{Tr}(\mathbf{H}_{\lambda_2}(\mathcal{A}))$$

where $\mathcal{A}$ denotes the active set of predictors, i.e. the predictors included in the model for a given pair of λ_1 and λ_2. $\mathbf{X}_{\mathcal{A}}$ contains the columns of the predictor matrix $\mathbf{X}$ corresponding to the active set. As we recommend cross-validation for the model selection in the metabolite model, we do not explicitly use the degrees of freedom in the metabolite model, but we state the formula here for completeness. The elastic net problem stated above can be reformulated into a problem involving an elastic net penalty α which controls the compromise between the ridge and lasso type penalties, and an overall penalty parameter λ.

A.4. **Implementation.** The estimation procedure was implemented in the open source software R, and the elastic net paths in the metabolite model calculated via the glmnet algorithm [31].

Appendix B. Model Selection Criteria

The rate-distortion theory, originally intended for data compression in the information theory field, can be adjusted to work as a model selection criterion in significance testing or cluster analysis on high-dimensional data like gene expression, or on metabolite data ([32]). Suppose we have N genes that act as responses (each observed n times in a microarray experiment). For each gene we have a set of predictors for which we would like to select a subset of features that influences the gene expression the most. We can use the lasso or some other shrinkage method to estimate a solution path for each gene, and calculate a residual sum-of-squares (SS_E) for each gene model, indexed by the penalty parameter λ.

We wish to minimize the overall distortion (i.e., the residual sum-of-squares) under the constraint that the total number of parameters used in the overall modeling must not exceed a certain bound (related to how many parameters we must "pay" to achieve some overall explanatory power in the set of model). Figure B.1 illustrates the rate, equal to $1/\lambda$ and indicating model complexity, and distortion (SS_E) curves for two genes.

The dashed lines indicated in Figure B.1 correspond to a fixed slope constraint Δ. For a given slope constraint, it can be shown that by selecting the points for each gene that is first intersected by the "moving" slope constraint, the overall distortion is minimized under the restriction that the total rate is less than or equal to the sum of the rates for the selected points [32].

We use BIC to select the model complexity in the gene model. Under the assumption that the model errors are Gaussian, the BIC criterion can be written

$$\begin{aligned} BIC &= -2 \cdot \text{loglik} + \text{df} \cdot \log(n) \\ &= n \cdot (\log(SS_E) - \log(n) + \log(2\pi) + 1) + \text{df} \cdot \log(n) \end{aligned}$$

with SS_E as above, n the number of replicates, and df the degrees of freedom in the model. The models we choose between for each gene is the null model, or the model with the treatment group indicator as the only predictor. Based on the rate-distortion criterion, we pick the slope Δ with the smallest overall BIC. The BIC criterion has some attractive properties, like asymptotical consistency, but also has the drawback of often selecting too sparse models in finite samples [23].

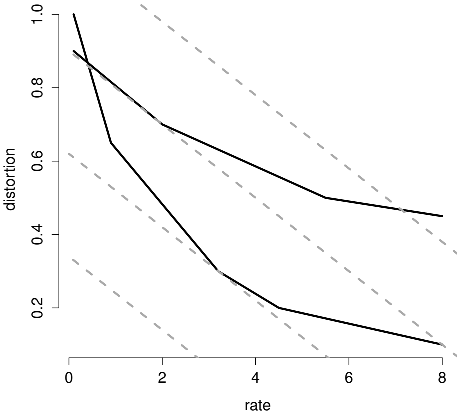
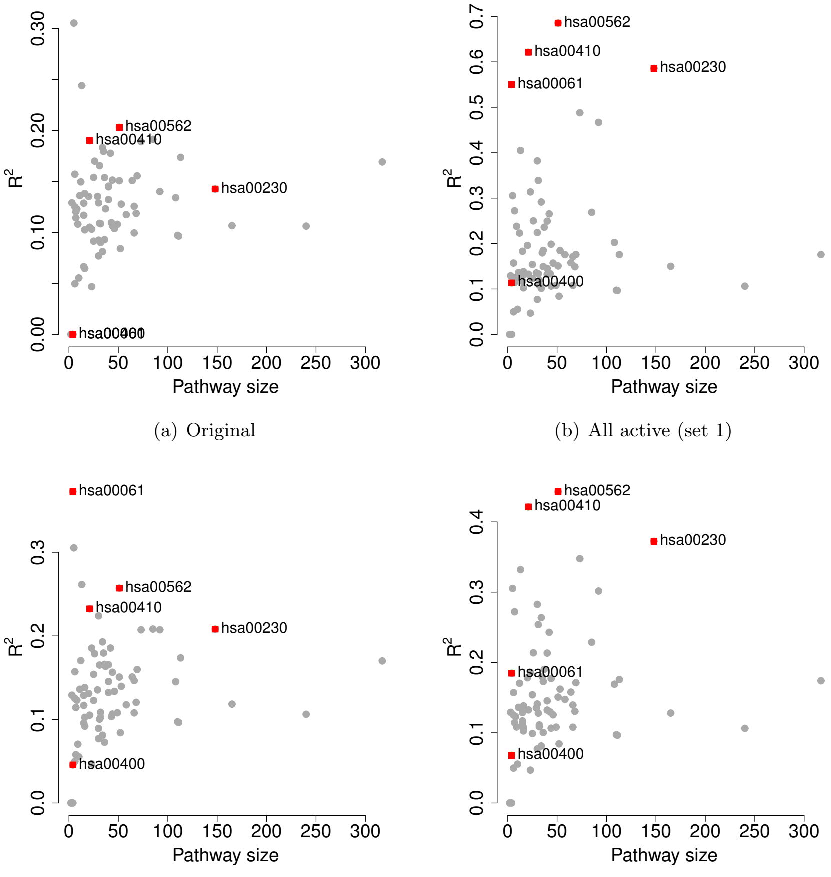


FIGURE B.1. Rate distortion curves for two genes. The dashed lines correspond to a fixed slope constraint Δ.

In the metabolite model, we chose to instead use a cross-validation procedure to select the overall rate-distortion slope. We also implemented BIC as a selection method, which we found to be very unstable in particularly $p > n$ situations. The overall trend was that sparser models was chosen for $n > p$ settings compared to the cross-validation procedure, but much more complex models chosen in $p > n$ situations, due to a dominating log-likelihood for complex models (with saturation for small values on the penalization parameter λ).

Appendix C. Supplementary Figures



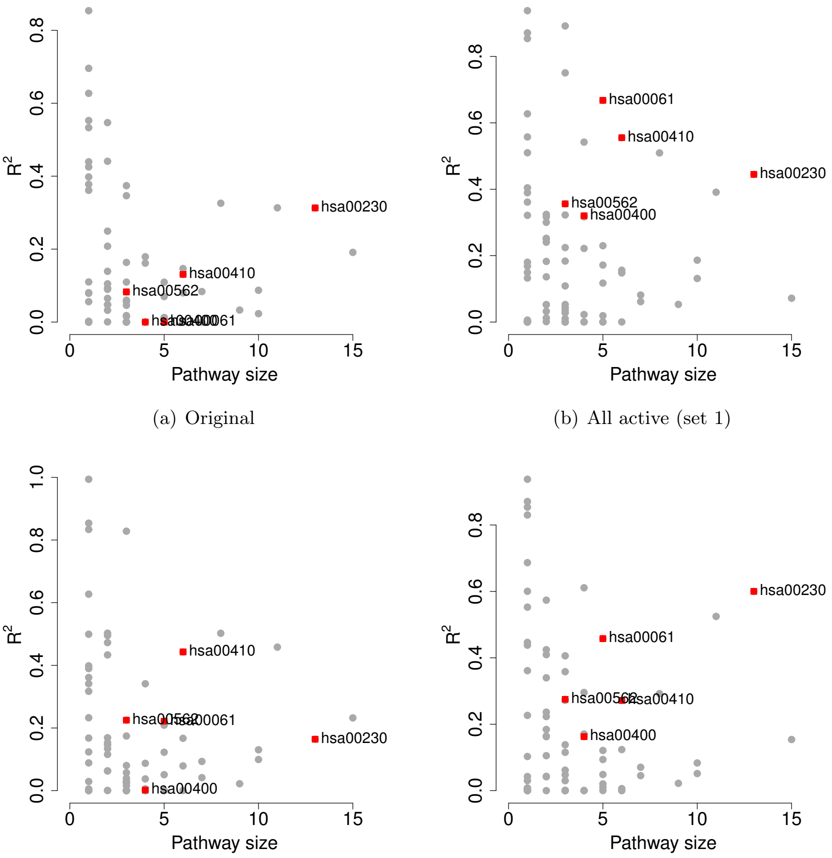
(a) Original

(b) All active (set 1)

(c) A third active (set 2)

(d) Half active (set 3)

Figure C.1. Overall R^2 scores in the first round of estimation for each pathway versus pathway size in the gene model on the original NCI-60 data and the three simulated data sets. The red squares indicate the pathways that were spiked in the simulated data sets.



(a) Original

(b) All active (set 1)

(c) A third active (set 2)

(d) Half active (set 3)

FIGURE C.2. Overall R^2 scores in the first round of estimation for each pathway versus pathway size in the metabolite models on the original NCI-60 data and the three simulated data sets. The red squares indicate the pathways that were spiked in the simulated data sets.

References

[1] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, pp. 719–724, Apr 2009.

[2] A. Sreekumar, L. M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher, and A. M. Chinnaiyan, "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression," *Nature*, vol. 457, pp. 910–914, Feb 2009.

[3] C. Abate-Shen and M. M. Shen, "Diagnostics: The prostate-cancer metabolome," *Nature*, vol. 457, pp. 799–800, Feb 2009.

[4] The Gene Ontology Consortium, "The Gene Ontology in 2010: extensions and refinements," *Nucleic Acids Res.*, vol. 38, pp. D331–335, Jan 2010.

[5] "The Gene Ontology." `http://www.geneontology.org`.

[6] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res.*, vol. 38, pp. D355–360, Jan 2010.

[7] "KEGG: Kyoto Encyclopedia of Genes and Genomes." `http://www.genome.jp/kegg`.

[8] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Res.*, vol. 38, pp. D473–479, Jan 2010.

[9] "MetaCyc." `http://metacyc.org`.

[10] K. Shen and G. C. Tseng, "Meta-analysis for pathway enrichment analysis when combining multiple genomic studies," *Bioinformatics*, vol. 26, pp. 1316–1323, May 2010.

[11] M. Ackermann and K. Strimmer, "A general modular framework for gene set enrichment analysis," *BMC Bioinformatics*, vol. 10, p. 47, 2009.

[12] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 107–129, 2007.

[13] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 15545–15550, Oct 2005.

[14] M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito, "Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, pp. 10205–10210, Jul 2004.

[15] Y. Gibon, B. Usadel, O. E. Blaesing, B. Kamlage, M. Hoehne, R. Trethewey, and M. Stitt, "Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes," *Genome Biol.*, vol. 7, p. R76, 2006.

[16] E. Allen, A. Moing, T. M. Ebbels, M. Maucourt, A. D. Tomos, D. Rolin, and M. A. Hooks, "Correlation Network Analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of *Arabidopsis*," *BMC Syst. Biol.*, vol. 4, p. 62, 2010.

[17] F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M. I. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L. J. Sweetlove, and A. R. Fernie, "Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior," *Plant Physiol.*, vol. 142, pp. 1380–1396, Dec 2006.

[18] M. Bylesjö, R. Nilsson, V. Srivastava, A. Gronlund, A. I. Johansson, S. Jansson, J. Karlsson, T. Moritz, G. Wingsle, and J. Trygg, "Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen," *J. Proteome Res.*, vol. 8, pp. 199–210, Jan 2009.

[19] P. H. Bradley, M. J. Brauer, J. D. Rabinowitz, and O. G. Troyanskaya, "Coordinated concentration changes of transcripts and metabolites in *Saccharomyces cerevisiae*," *PLoS Comput. Biol.*, vol. 5, p. e1000270, Jan 2009.

[20] I. M. Frey, I. Rubio-Aliaga, A. Siewert, D. Sailer, A. Drobyshev, J. Beckers, M. H. de Angelis, J. Aubert, A. Bar Hen, O. Fiehn, H. M. Eichinger, and H. Daniel, "Profiling at mRNA, protein, and metabolite levels reveals alterations in renal amino acid handling and glutathione metabolism in kidney tissue of Pept2-/- mice," *Physiol. Genomics*, vol. 28, pp. 301–310, Feb 2007.

[21] C. T. Ferrara, P. Wang, E. C. Neto, R. D. Stevens, J. R. Bain, B. R. Wenner, O. R. Ilkayeva, M. P. Keller, D. A. Blasiole, C. Kendziorski, B. S. Yandell, C. B. Newgard, and A. D. Attie, "Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling," *PLoS Genet.*, vol. 4, p. e1000034, Mar 2008.

[22] E. Y. Xu, A. Perlina, H. Vu, S. P. Troth, R. J. Brennan, A. G. Aslamkhan, and Q. Xu, "Integrated pathway analysis of rat urine metabolic profiles and kidney transcriptomic profiles to elucidate the systems toxicology of model nephrotoxicants," *Chem. Res. Toxicol.*, vol. 21, pp. 1548–1561, Aug 2008.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer Series in Statistics, New York, NY, USA: Springer, 2009.

[24] K. A. Le Cao, D. Rossouw, C. Robert-Granie, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Stat. Appl. Genet. Mol. Biol.*, vol. 7, p. Article 35, 2008.

[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, vol. 67, pp. 301–320, 2005.

[26] "Download page for NCI-60 molecular target data." `http://dtp.cancer.gov/mtargets/download.html`.

[27] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," *Nucleic Acids Res.*, vol. 34, pp. W729–732, Jul 2006.

[28] G. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 3, p. Article3, 2004.

[29] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.

[30] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc. Ser. B*, vol. 58, pp. 267–288, 1996.

[31] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *J. Stat. Soft.*, vol. 33, no. 1, pp. 1–22, 2010.

[32] R. Jörnsten, “Simultaneous model selection via rate-distortion theory, with applications to cluster and significance analysis of gene expression data,” *J. Comput. Graph. Statist*, vol. 18, pp. 613–639, 2009.