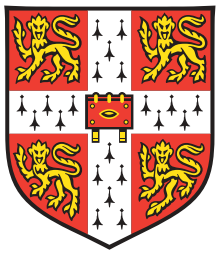
Transcriptional and Post-transcriptional Regulation of Gene Expression: Computational Analysis of Microarray Studies in Fungal Species

Katherine Joanne Lawler

Darwin College
October 2009



A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

European Molecular Biology Laboratory,
European Bioinformatics Institute

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using LaTeX2ε according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Katherine Lawler
Cambridge, October 2009

Transcriptional and Post-transcriptional Regulation of Gene Expression:
Computational Analysis of Microarray Studies in Fungal Species

Summary

DNA microarrays remain a powerful tool for identifying changes in gene expression between different environmental conditions or developmental stages. Coordinated regulation of gene expression is typically studied by identifying groups of genes with correlated changes in mRNA abundance across different experimental conditions. Recent computational methods attempt to reconstruct networks of gene regulation from global expression patterns. In time series studies, the temporal changes in mRNA abundance typically measured by DNA microarrays are the result of a balance between transcription and mRNA degradation. There is a growing body of evidence that a global gene expression response can be regulated at both the transcriptional level and at a post-transcriptional level through the regulation of mRNA stability.

This thesis presents two related computational studies of the genome-wide regulation of gene expression based on the analysis of microarray datasets. The first study concerns the dynamics of a global gene expression response. The regulation of mRNA abundance by both transcriptional and post-transcriptional control implies a range of possible strategies for shaping gene expression in response to a stimulus. Strategies for shaping gene expression are investigated, and I examine the strength of evidence for regulated mRNA stability from microarray time series. In particular, I investigate the role of regulated mRNA stability in the gene expression response to oxidative stress in the fission yeast *Schizosaccharomyces pombe*. A dynamic model of mRNA abundance is applied to simultaneous time series of mRNA abundance (DNA microarray) and transcription rate (RNA polymerase II ChIP-chip) datasets. Candidate genes are identified for which the gene expression response appears to be driven by a change in mRNA stability rather than by transcriptional control. The dynamic gene expression response of stress-induced and -repressed genes and stress response regulators is described.

The second study is concerned with patterns of near-steady-state gene expression levels in the recently sequenced fungal crop pathogen *Fusarium graminearum*. I derive gene expression patterns from all available genome-wide expression datasets, covering various life cycle stages and growth conditions. Expression analysis is combined with recently predicted transcription-associated proteins to identify genes coexpressed with putative DNA-binding transcription factors. The distribution of coexpressed genes on the genome is investigated: localized regions of the *F. graminearum* genome are found to be enriched for coexpressed genes, and functional annotation is provided for these localized regions based on sequence homology to annotated protein families.

Acknowledgements

This work was carried out in the Microarray Informatics Group at the European Bioinformatics Institute (EBI), where I was supported by an EMBL predoctoral fellowship. The EMBL Core Course during the first three months of the EMBL PhD programme provided an invaluable opportunity for a mathematics graduate to get a first glimpse of life, work, and open questions in the vast and complex world of molecular biology.

First I would like to thank my supervisor, Alvis Brazma, for his support and guidance throughout my time at EBI and for introducing me to the world of gene expression microarrays and high-throughput genomic studies. Thanks also to my thesis advisory committee Jürg Bähler, Wolfgang Huber, Luis Serrano and Michael Ashburner for help and advice at various stages of this work.

Thanks to Richard Coulson (EBI/CIMR) and Samuel Marguerat (WTSI/UCL) for scientific guidance, for many fruitful discussions, and for so generously and patiently sharing their knowledge. Thanks also to the Fission Yeast Genomics Group at WTSI, in particular Falk Schubert for statistical advice, and to Kim Hammond-Kosack and colleagues (RRes) for datasets and expertise.

The Microarray Informatics Group at EBI has been a fascinating group of people to work amongst. Particular thanks to Ekaterina Pilicheva, Misha Kapushesky, Juok Cho and Johan Rung for stimulating discussions and technical advice. The EBI community of predocs and postdocs has provided a bedrock of enthusiasm and expertise. Particular thanks to Joern Toedling for a crash course in writing and thinking simultaneously and for such thorough and constructive criticism.

To James, to CL (1978-2007), and to my family for their patience, support and encouragement. Many thanks.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis describes three studies in which DNA microarray datasets have been analysed, resulting in insights into the regulation of transcriptional and post-transcriptional control of gene expression on the scale of whole genomes.

This chapter introduces some general aspects of the biology of gene expression and describes established technologies for the measurement of changes in gene expression on a genome scale. The microarray platforms and associated preprocessing methods which had been used to generate the datasets used in this work are described, followed by an overview of gene annotation used in subsequent chapters. Subsequent chapters contain additional background information relevant to the studies presented in each chapter. Finally, an overview of this thesis is presented.

1.1 Regulation of gene expression

This section briefly introduces key aspects of transcriptional and post-transcriptional regulation in eukaryotic cells which are referred to throughout this thesis. References are given to review articles on aspects of gene expression and regulation.

The diversity of cell phenotypes which are produced from identical genomes is primarily due to differences in gene expression, whether between different cell types in a multicellular organism, or as a result of diverse gene expression responses be-

tween different physiological conditions or developmental stages. The complement of proteins present in the cell is the product of a complex set of mechanisms by which proteins are produced from genes encoded in DNA. Gene expression occurs when DNA is transcribed into RNA which is then translated into protein. The rate of production of functional proteins in the cell is regulated at many stages of gene expression, primarily at the level of transcription but also at post-transcriptional levels. It is convenient to describe gene expression as a series of sequential steps, from transcription to post-translational protein modifications; however, many transcriptional and post-transcriptional mechanisms of gene expression are interdependent and simultaneous [1] and the extent of this interdependence is not fully understood.

1.1.1 Transcriptional regulation of gene expression

Initiation of transcription requires the assembly of the pre-initiation complex at the transcription start site. RNA polymerase is recruited and moves along the DNA producing an RNA transcript (elongation). Transcription is terminated and the RNA transcript is post-transcriptionally modified. RNA polymerase II transcribes the majority of eukaryotic protein-coding genes. A rate-limiting step for transcription is the recruitment of RNA polymerase II to the core promotor [2]. RNA polymerase II stalling, in which the polymerase pauses on the gene during transcript elongation, has been observed *in vivo* in mammalian cells [3] and on *Drosophila* heat-shock genes [4]. The aggregation of RNA polymerase II close to the transcription start site appears to be a widespread mechanism for stress response regulation and recovery. RNA polymerase II has been found to be located upstream of hundreds of genes which are subsequently rapidly induced upon exit from *S. cerevisiae* stationary phase [5], and more recently upstream of genes related to development in *Drosophila* [6] and human stem cells [6].

1.1.1.1 Transcription factors

In this thesis, the term *transcription factor* is used to refer to the specific transcriptional activators and repressors that activate or repress the transcription of target genes *via* specific binding to promotors regions. Feedback loops involving a handful of transcription factors can control complex gene expression responses and developmental processes. Classical examples of combinatorial transcriptional control by a small number of transcription factors include the development of polarity and segmentation in the *Drosophila* embryo [7], and sea urchin embryogenesis [8]. On a genome-wide scale, initial genome-wide networks of transcriptional control have been constructed for *S. cerevisiae* [9, 10]. These early network models of transcriptional regulation suggest that transcriptional control is rich in feedback events and combinatorial control (reviewed in [11, 12]).

1.1.1.2 Chromatin structure and chromatin remodelling

Transcriptionally active chromosomal regions have been correlated with chromatin structure and associated chromatin modifications. Transcriptional activation and repression by specific transcription factors, and transcription initiation involving the general transcription factors, requires that proteins are bound to specific regions of DNA. Nuclear DNA in eukaryotic cells exists as highly structured chromatin, of which the basic structure is the nucleosome consisting of DNA wrapped around a central core of eight histone proteins. Chromatin structure is dynamically altered by the presence of histone variants and by several covalent modifications of histone tails, including acetylation, methylation and phosphorylation of specific residues [13, 14]. Histone tail modifications act dynamically and in combination to alter the local chromatin structure, either opening up regions of the chromatin and therefore permitting a higher rate of binding by transcriptional activators, or condensing chromatin into a transcriptionally inactive state. For transcription to be initiated, regulatory and general transcription factors must bind to DNA in promotor regions

and around the transcription start site. Regions of open chromatin are more likely to be available for binding and therefore to be more transcriptionally active than regions of densely compacted chromatin. Histone modifications and nucleosome positioning have been correlated with transcriptionally active and inactive chromosomal regions, for example in *S. cerevisiae* [15, 16, 17], human and mouse cell lines [18, 19], and during human heart cell development [20]. Such correlations are consistent both with chromatin modifications causing transcriptional activation, and with existing transcriptionally active regions being marked as such and promoting transcription of an already active region. In multicellular organisms, regions of chromatin are condensed to the point of being transcriptionally silent [21]. Chromatin modifications and associated DNA methylation can be inherited across cell divisions, contributing to the maintainance of differentiated cell lines in multicellular organisms and the differential expression of genes between different cell types and disease states [22, 23].

1.1.1.3 Spacial arrangement of transcriptionally active chromosomal regions

Recent studies using chromosome conformation capture (3C, 4C, 5C) technology indicate that the location of chromosomal regions in the nucleus is highly predictable and is reproduced across successive cell cycles [24, 25]; reviewed in [26]. Given the apparent importance of relative chromosomal positions in the nucleus, a resulting hypothesis is that the arrangement of transcribed regions on the chromosome may be constrained by the regulation of transcription.

1.1.2 Co-transcriptional and post-transcriptional regulation of gene expression

Concurrently with transcription, nascent pre-mRNA transcripts are modified by the addition of a 5′ m^7G cap structure, and following transcript termination the 3′ end of the transcript is polyadenylated. The 5′ cap structure and 3′ poly-A tail are important binding targets for proteins involved in mRNA stability, translation initiation,

and nuclear export. Pre-mRNA contains introns which are removed by the spliceosome complex at consensus sequences at exon-intron boundaries to produce mature mRNA. Alternative splicing contributes significantly to the diversity of proteins in higher eukaryotes. Functional mature mRNA is exported from the nucleus where it is translated in association with ribosomes. mRNA transcript stability is regulated by RNA-binding proteins and small RNA molecules. Families of RNA-binding proteins, such as the AU-rich element (ARE)-binding proteins, bind specifically to mRNA sequence motifs or mRNA structure motifs. RNA-binding protein sequence motifs are often but not always located in the 3' untranslated region (UTR). MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are two classes of small RNAs which have recently been implicated in the regulation of mRNA stability and translation [27]. Post-transcriptional regulation of gene expression is reviewed in [28] with emphasis on the regulation of translation and mRNA stability.

1.2 DNA microarray platforms

The microarray datasets used in the studies described in this thesis had been produced using two DNA microarray platforms: cDNA microarrays (for both expression analysis and ChIP-chip assays) and Affymetrix GeneChip oligonucleotide arrays. There are several other established microarray platforms for expression analysis, including exon arrays [29], high resolution tiling arrays [30], and Illumina bead arrays [31]. All microarray datasets must be preprocessed in order to correct for systematic technical effects observed in raw probe intensities. Depending on the platform and array diagnostics, preprocessing may include background correction, within-array normalization or between-array normalization.

1.2.1 cDNA microarrays

Two-channel cDNA microarrays are used in expression studies to directly compare the abundance of mRNA transcripts in two different cell populations [32]. cDNA

probes correspond to selected regions of the genome – typically regions within known or predicted open reading frames for expression studies. Probes are spotted onto the array surface so that a specific mRNA transcript preferentially hybridizes to a set of replicate spots. mRNA is purified from each sample and reverse transcribed into the more stable DNA complementary to the mRNA transcripts[1]. Each sample is tagged with a different fluorescent dye, usually Cy3 and Cy5, and the reverse transcribed DNA fragments compete for binding to the probes which are fixed on the microarray surface. Hybridized arrays are scanned to produce an array image corresponding to each of the two channels. Arrays are scanned at two wavelengths corresponding to the excitation wavelengths of the two dyes, and the intensity of emitted light at two emission wavelengths is captured for each pixel. An intensity is reported for every spot for each of the two channels by summarizing the pixel intensities within each spot. The median of pixel intensities is typically used to summarize the spot intensity. A quality score may be assigned to each spot for each channel, allowing low quality spots to be flagged: a spot may be flagged as low quality if the spot intensity is too close to the background intensity, the standard deviation of pixel intensities within the spot is too high, or the detected spot has an unexpected shape in the array image. After normalization of all summarized intensity values, a normalized ratio or *fold-change* between the channel intensities is typically reported for each spot as a measure of the ratio of mRNA abundance in the two samples.

It is not possible to quantitatively compare two different mRNA species using hybridization methods. Each mRNA species has a different reverse transcription efficiency, leading to reverse transcription bias in which the relative amounts of each cDNA are not the same as the relative amounts of mRNA in the sample. Labelling efficiency and hybridization efficiency also vary between different transcripts due to variations in GC content, presence of various levels of truncated cDNA due to

[1]Detailed protocols for *Schizosaccharomyces pombe* RNA extraction, mRNA purification, reverse transcription, fluorescent labelling and hybridization are described by Bähler and colleagues [33], for example.

reverse transcription bias, and differences in the stability of the RNA secondary structure [34].

1.2.1.1 Normalization of cDNA microarray intensity values

In order to compare the hybridization of an mRNA species between two different cDNA channels, the raw summarized intensity values must be normalized to control for any systematic technical effects observed in the reported intensity values. This includes correcting for spatial artifacts on the microarray slide, correcting for imbalances in the measured intensities of the two dyes, or removing signal intensity-dependent bias in intensity ratio values [35, 36]. There are several established statistical methods for the normalization of cDNA microarray data including lowess (locally weighted scatterplot smoothing) whole-array or print-tip normalization which corrects for intensity-dependent bias in ratio intensities [37], and variance-stabilizing normalizations such as VSN [36]. The choice of normalization procedure is guided by the microarray platform and array design, the experimental design, and the biological assumptions behind a particular experiment.

The *S. pombe* cDNA microarray datasets on which Chapter 4 is based were normalized using a local median-of-ratios method. This normalization procedure was designed and implemented by Bähler and colleagues for use with the *S. pombe* cDNA microarrays which have been designed, produced and hybridized by the Fission Yeast Genomics Group, Sanger Institute [33]. The dominant technical effect on these arrays is spatial variation of raw intensity values, as illustrated in Figures 1.1 and 1.2. To correct for spatial variation on the array, a sliding window is moved across the array and the intensity ratio of the centre spot is adjusted by a normalization factor calculated such that the median of ratio intensity values in the window is equal to one. An underlying assumption made in applying this method is that there is a balance of up- and down-regulation of mRNA abundance between the two channels or that the majority of genes are not differentially expressed between the two samples, so that the median of spot intensity ratios is expected to be equal to one within

any spatial window on the array. In experiments for which this assumption is not valid an alternative normalization should be used, such as normalization based on external bacterial controls [33, 38].

The *S. cerevisiae* cDNA microarray datasets [5] used in Chapter 2 had been normalized using external RNA controls which were added across a range of concentrations. Spacial normalization had been applied to each array by fitting loess curves to external control spots within each print-tip region of the array and applying the resulting normalization functions to all spots [38]. External control normalization had been used in this case in order to capture global changes in mRNA abundance [5].

1.2.2 Affymetrix GeneChip oligonucleotide arrays

Affymetrix GeneChip arrays are another widely used platform for gene expression analysis [39]. Chapter 5 is based on an analysis of several Affymetrix GeneChip experiments. The array design consists of probesets which are designed to be complementary to a region of each mRNA transcript, usually at the 3′ end of the transcript due to degradation of mRNA from the 5′ end. Each probeset consists of a set of 11-20 perfect match (PM) probes which are typically 25 nucleotides long, together with an equal number of mismatch (MM) probes which are identical to the PM probes except for a single nucleotide substitution in the centre of the probe.

1.2.2.1 Normalization and probeset detection for Affymetrix GeneChip oligonucleotide arrays

There are a number of established preprocessing methods for Affymetrix GeneChip arrays. GeneChip probe intensities must be background corrected and normalized within or between arrays, and then summarized to produce an intensity measurement for each probeset. All between-array and within-array normalization procedures involve a trade-off between noise reduction and the introduction of bias. In

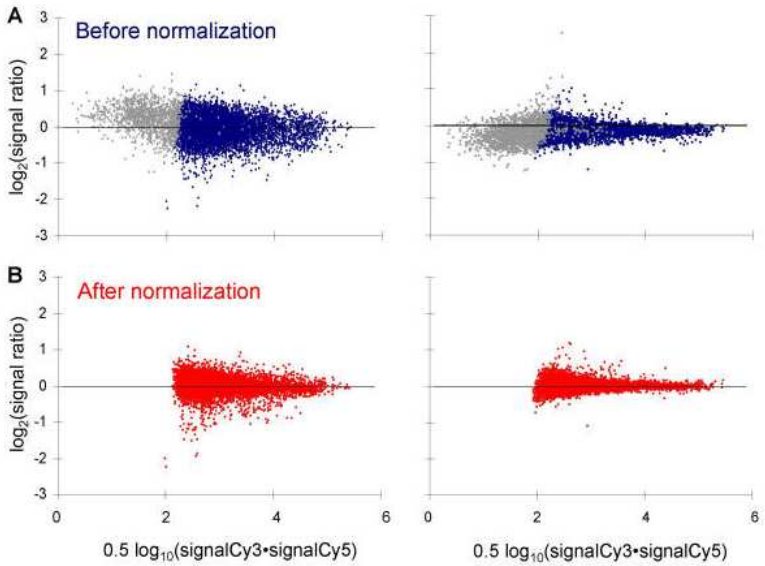


Figure 1.1: MA-plots of *S. pombe* cDNA microarray datasets show that there is no apparent intensity-dependent bias in the ratio of channel intensities. (A) Left: cells grown at 25C (Cy5) compared to cells grown at 30C (Cy3). Right: a self-self hybridisation. Grey spots are those with intensities close to background and are discarded before spatial normalization. (B) As in (A) after spatial normalization. This figure is from Figure 2 in Lyne et al. 2003 [33]

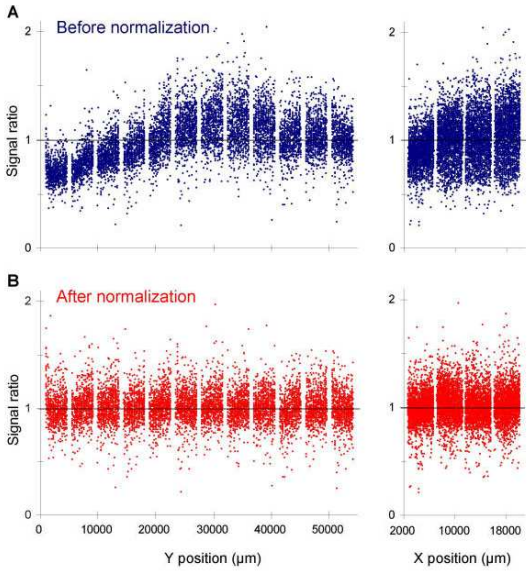


Figure 1.2: Correction of spatial artifacts by local median-of-ratios spatial normalization. (A) Distribution of signal ratios along the Y- (left) and X-axis (right) of the microarray before normalization. The data are from array shown in Figure 1.1 (left side). The groups of spots separated by small gaps reflect the 12 x 4 sub-grids of the array, each printed with a different spotting pin. (B) Distribution of signal ratios as in (A) after local median-of-ratios normalization of the data. This figure is from Figure 3 in Lyne et al. 2003 [33]

Chapter 5, several Affymetrix GeneChip datasets are preprocessed and analysed and the selection of appropriate normalization probeset detection methods for this study is discussed.

The Affymetrix MAS 5.0 [40] normalization and probeset detection algorithms make use of the signal intensities from mismatch probes to correct for the contribution of non-specific binding to background intensity levels. The dChip [41] normalization and summarization algorithm similarly uses mismatch probe to correct for non-specific binding but also models probe-specific hybridization affinity. The RMA (log scale robust multi-array analysis) summarization method [42] discards information from mismatch probes. The RMA expression measure for probeset-based arrays is calculated from raw probe intensities in three steps: background correction, normalization within and/or between arrays, and probeset summarization. RMA summarization has been shown to outperform the MAS 5.0 summarization method, particularly for low intensity probesets [42]. and is currently the preferred standard method for differential expression calls. *RMA* refers only to the probeset summarization method and this can be combined with any background correction and normalization procedures before the probeset summarization step.

Quantile normalization between arrays [43] assumes that the relative abundance of each hybridizing RNA species on each array, and therefore the distribution of intensities on each array, are approximately equal. A virtual reference array is constructed by pooling all probe intensities from all arrays. For each array, each probe intensity is mapped to the intensity corresponding to the same quantile on the reference array. Variance-stabilizing normalizations [36, 44] aim to reduce the dependence of the intensity variation on the mean of channel intensities. If the signal intensities on different arrays are not expected to follow a similar distribution, it may be appropriate to fit a normalization function on a subset of array probes which is then applied to the whole array [36, 38].

A recent method calculates a probeset detection 'barcode' [45] by pooling hundreds

of publicly available hybridizations of a given microarray platform in order to estimate the undetected intensity distribution for each probeset on each platform. This method estimates the probeset- and platform-specific intensity distributions of undetected probesets and has been shown to perform better than MAS 5.0 for the detection of low intensity signals, but this method relies on the availability of a large number of hybridisations to estimate the undetected distribution and an associated 'detected' intensity threshold.

1.2.3 ChIP-chip datasets

In addition to differential expression studies, microarrays (cDNA, oligonucleotide, and high resolution tiling arrays) have been used to detect DNA-binding proteins in promoter regions [46], to locate RNA polymerase II [6, 47] and nucleosome positions [48] , and to identify chromosomal regions containing histone or chromatin modifications [20]. ChIP-chip assays consist of chromatin immunoprecipitation (ChIP) against the DNA-binding protein or modification of interest, followed by hybridization of IP-enriched DNA fragments and a genomic DNA control sample. An RNA polymerase II ChIP-chip assay was used to estimate changes in transcription rate in *S. pombe* in Chapter 4. A detailed protocol for the assay can be found in Lackner *et al.* [47] and is briefly described here. Cells are crosslinked in formaldehyde, then chromatin is extracted and sonicated into random fragments, and immunoprecipitation is performed using an antibody specific to the protein or modification of interest. To target RNA polymerase II, for example, an antibody specific to the carboxyl-terminal domain of RNA polymerase II is used. Crosslinking is reversed and the enriched DNA is purified, fluorescently labelled, and hybridized to an array. For the *S. pombe* RNA polymerase II assays, the sample was hybridized to a two-channel cDNA array and a control sample of genomic input DNA was fluorescently labelled and hybridized to the same array. Local normalization was used as described for the RNA (expression) hybridizations.

1.2.4 High-density microarray platforms and high-throughput sequencing technologies

Although not the focus of this thesis, it is important to note that recent advances in microarray technology and ultra high-throughput sequencing technology have resulted in platforms which can be used to map the transcriptome at a higher resolution than cDNA or GeneChip oligonucleotide arrays, and which can detect transcripts originating from any region of the genome. High-density tiling arrays contain a set of probes mapping to a higher proportion of the genome than the more established cDNA microarrays and Affymetrix (GeneChip) oligonucleotide arrays. Tiling arrays consist of short probes which target the whole genome at a high resolution and can therefore be used to detect all transcribed regions of the genome or to map ChIP targets to the genome at a high nucleotide resolution [30]. For example, the *Saccharomyces cerevisiae* Affymetrix tiling array used by Steinmetz and colleagues consists of 25-mer probes tiled at eight nucleotide intervals [49], and has been used to show that more than 95% of the *S. cerevisiae* genome is transcribed. The *S. pombe* Affymetrix tiling array contains 25-mer probes at 20-nucleotide intervals on both strands and has been used by Bähler and colleagues [50], and others, to define the *S. pombe* transcriptome under various conditions. Tiling arrays provide an unbiased survey of the entire transcriptome, in that the probes are not designed to match specifically to previously identified protein-coding sequences or other specific regions of the genome.

Ultra high-throughput cDNA sequencing (RNA-Seq) is also becoming established as a sensitive method for detecting all transcribed genomic regions. RNA-seq has a high dynamic range of transcript detection and in particular is capable of detecting transcripts that are expressed at a low level and are not detected by arrays [50]; reviewed in [51]. There remain significant challenges in identifying the genomic origin of transcripts and quantifying the expression levels of transcribed genomic regions using tiling arrays and high-throughput sequencing [49, 30], but there is no doubt that information gained from new high-throughput technologies about the

transcriptome and its regulation will increasingly be used to complement datasets from the more established microarray platforms.

1.3 Functional annotation: Gene Ontology and enrichment analysis

The Gene Ontology (GO) project [52] provides three ontologies which are used for the systematic description of gene products: biological function, cellular component, and molecular function. Each ontology forms a rooted directed acyclic graph in which each node is associated with a GO identifier (or *GO term*). A gene annotated with any given GO term is also annotated with all ancestral GO terms, allowing for desciptions of the gene preduct at varying levels of specialization. Generic and species-specific versions of the Gene Ontology are continuously updated based on experimental or electronically derived evidence [52].

1.3.1 GO annotation of *S. pombe*

GO annotation for *S. pombe* genes [53] is available from the GeneDB *S. pombe* genome database [54]. As of October 2009, 5178 *S. pombe* genes have been annotated with GO terms from the three ontologies[2]. In this thesis, hypergeometric tests for overrepresentation of *S. pombe* GO terms were performed using Gene List Analyser 1.0 [55]. Independent hypergeometric tests were performed for each GO term [56] and no distinction was made between reported annotation from different sources of evidence. Gene List Analyser 1.0 was also used to perform the tests for overrepresentation of genes in other *S. pombe* gene lists of interest which are reported in Chapter 4. Reported p-values ($p \leq 0.05$) have been corrected for the testing of multiple GO categories using the false discovery rate (FDR) multiple testing correction [57].

[2] *S. pombe* GO annotation may be browsed using AmiGO [52] at the following url: http://www.genedb.org/genedb/pombe

1.3.2 GO annotation of *F. graminearum*

There is currently no published species-specific GO annotation for the recently sequenced *Fusarium graminearum* genome considered in Chapter 5. GO annotation was therefore assigned to predicted *F. graminearum* genes using an existing mapping of GO annotation to protein families. A curated mapping of GO annotation to protein domains or protein families is available from the Interpro database [58]. Predicted *F. graminearum* proteins were mapped to Interpro protein families using hidden Markov model (HMM) searches (HMMER [59]) of existing PFAM HMM protein domain models [60]. GO annotation was transferred from an Interpro protein domain or protein family to a predicted *F. graminearum* gene if a match to the associated HMM was detected. GO annotation of predicted *F. graminearum* genes was performed by Richard Coulson[3]. Using this method, 5024 genes (36% of predicted genes) were annotated by one or more GO terms. In an attempt to improve the coverage of GO annotation, the protein sequences of predicted *F. graminearum* genes were compared directly to eukaryotic proteins contained in UniProt [61]. Protein sequences were compared pair-wise using Blastp [62] and clusters of proteins with similar protein sequences were detected using Markov clustering [63]. GO annotation was transferred to a *F. graminearum* gene if a high degree of similarity to a previously annotated gene was detected; see Chapter 5 (Methods, page 115) for further details.

1.4 Thesis Aims

A long-term goal is to use post-genomic datasets to understand how the cell integrates diverse signals in order to coordinate a transcriptional, post-transcriptional, and metabolic response, whether in response to environmental changes or during a transition between developmental stages or phases of the life cycle. With the increasing availability of genome-scale datasets, it is possible to model the cell at var-

[3]Microarray Informatics Group, EMBL-EBI

ious levels of complexity, from the regulation of transcription [64], to the regulation of proteins and protein-protein interactions [65], to dynamic models of metabolic and biosynthetic pathways [66].

This thesis focuses on one aspect of the regulation of a cell's internal environment: the genome-wide regulation of mRNA abundance, either in response to environmental changes or during the transition between life cycle phases. The primary datasets used in this thesis are gene expression microarray datasets which capture changes in mRNA abundance between conditions on a whole-genome scale. As a first approximation coexpressed genes may be assumed to be coregulated [67]. As described in Chapter 2, a number of regulatory models have been developed in order to explain observed mRNA abundance profiles as the result of coregulation by DNA-binding transcription factors. Regulatory effects which are unobserved but potentially shared by genes with similar gene expression profiles, such as DNA-binding transcription factors activated by post-transcriptional modifications, may be modelled as hidden variables, revealing potential shared regulatory effects between coexpressed genes (*e.g.* [68, 69]).

The abundance of an mRNA species is the result of a balance between transcription rate – the rate of production of mRNA transcripts – and mRNA degradation. Recent models of transcriptional regulation have accounted for mRNA degradation by assuming that each mRNA species is degraded at a constant, gene-specific, rate [70, 71]. However, the contribution of regulated mRNA stability to the regulation of gene expression levels on a genome-wide scale is not well understood. In order to investigate the contribution of mRNA degradation to changes in mRNA levels, a time course of changes in transcription rate was considered here alongside a time course of mRNA abundance. This transcription rate time course had been generated using a recently developed RNA polymerase II ChIP-chip assay in *S. pombe*, and has permitted the first genome-wide analysis of the contribution of regulated mRNA stability to a dynamic gene expression response in *S. pombe*.

In contrast to high time-resolution dynamic studies of changes in mRNA levels in response to a stimulus, comparisons of mRNA levels between different steady state conditions can reveal differences in transcriptional programs between different conditions. Groups of coexpressed genes may share transcription regulatory properties, for example coregulation by DNA-binding transcription factors (which may be detected as co-occuring motifs in promotor regions) or coregulation of local chromatin structure. This thesis concludes with a study of differential expression between steady state conditions in *Fusarium graminearum*, providing a first inter-experiment map of differentially expressed genes and coexpressed predicted transcriptional regulators in this crop pathogen.

1.4.1 Thesis overview

This thesis continues with the following chapters:

Chapter 2 presents a case study on the reconstruction of transcription regulatory networks during stationary phase exit and entry in *Saccharomyces cerevisiae*. A high resolution microarray time series is used to derive hypotheses about the transcriptional control of gene expression during stationary phase exit and re-entry. This study motivated the subsequent investigation into the shaping of a dynamic gene expression response by the regulation of both transcription rate and mRNA turnover, which is the subject of Chapters 3 and 4.

Chapter 3 considers how changes in mRNA abundance can be controlled by both dynamic transcription rates and regulated mRNA turnover, and presents a framework for the detection of regulated mRNA stability using microarray time series experiments.

In Chapter 4, I investigate the contribution of mRNA stability and transcriptional control to shaping a gene expression response to oxidative stress in *Schizosaccharomyces pombe*. The models developed in Chapter 3 are applied, along with other methods, to identify genes which are candidates for regulated mRNA stability in

response to environmental stress. The dynamic transcriptional response is also investigated.

Finally, Chapter 5 describes an integrative transcriptomics study of gene expression in *Fusarium graminearum* using multiple gene expression datasets and genome annotation, and presents new observations and hypotheses about the transcriptional regulation of gene expression during the *F. graminearum* life cycle and crop infection.

Chapter 2

Inferring transcription regulatory control from a microarray timecourse: a case study in *Saccharomyces cerevisiae*

This chapter presents a case study on the reconstruction of the control of transcription regulation by DNA-binding transcription factors during stationary phase exit and entry. The study considers a timecourse of exit from stationary phase and subsequent re-entry into stationary phase from exponential growth in the budding yeast *Saccharomyces cerevisiae*. Hypotheses about the transcriptional regulation of gene expression are presented, based on the analysis of a timecourse dataset and generated under two specified models of transcriptional control.

2.1 Introduction

The budding yeast *S. cerevisiae* is the most well-studied single cell eukaryotic model organism on a genome-wide scale. Many of the housekeeping functions in *S. cerevisiae* are conserved in higher eukaryotes, but complexities such as alternative splicing and families of small regulatory RNAs which are present in higher eukaryotes are not as important in the transcriptional control of *S. cerevisiae*. The relative simplicity of *S. cerevisiae* compared to vertebrate and mammalian cells means that we

can attempt to reconstruct aspects of transcriptional regulation using simple *cis*-regulatory models of transcription. In this chapter I assume that observed changes in mRNA level can be explained by a linear combination of active forms of DNA-binding transcription factors which are present with unobserved concentrations or *activity profiles* in the cell population. By applying two previously described linear models of transcriptional regulation, the unobserved activity profiles of transcription factors and their effects on target genes are inferred.

2.1.1 Transcription regulation during stationary phase exit and re-entry in *S. cerevisiae*

When starved of nutrients a population of *S. cerevisiae* cells enters stationary phase composed of quiescent cells, an inactive state in which the cell can survive adverse conditions. *S. cerevisiae* quiescence may be used as a model for mammalian G0 cells [72], the cellular state in which mammalian cells spend most of their lifetime. Stationary phase cultures produced by different nutrient-limiting conditions are associated with different transcriptional programs [73], and changes in mRNA abundance during distinct phases of stationary phase exit or entry have been observed on a genome-wide scale in various nutrient-limited conditions [74, 75, 73]. Radonjic and coworkers [5] performed a nine-day study of exit from stationary phase and subsequent re-entry into stationary phase. The initial and final quiescent cell states in that study were induced using glucose-limited conditions: an initial, glucose-limited stationary phase culture was resuspended in fresh medium (2% glucose) and grown for nine days until a second glucose-limited stationary phase was reached. The resulting dataset is the first available high-resolution genome-wide timecourse measuring changes in mRNA abundance in *S. cerevisiae* during stationary phase exit and subsequent re-entry. The authors identified clusters of genes with coherent changes in mRNA abundance during the 9-day study [5]. A rapid transcriptional response was observed upon redilution of quiescent cells with fresh medium. This response includes 16 transcription regulators which are rapidly and transiently induced at the

level of mRNA abundance upon exit from stationary phase, upregulated $\geq$ 4-fold within 3 mins of dilution with fresh medium [5]. A further 20 clusters of genes were defined by the authors as having similar mRNA abundance profiles during distinct time periods during stationary phase exit, growth phases and subsequent re-entry into stationary phase (see Figure 3 in [5]). The clusters contain DNA-binding transcriptional regulators: the upregulation of mRNA abundance of transcriptional regulators suggests there may be a functional role for such regulators at respective timepoints, whereas the downregulation of transcriptional regulators may indicate no concurrent functional role or may be a result of active repression *via* a network of transcriptional activators or repressors [9]. Further, RNA polymerase II was shown to be present upstream of hundreds of genes before the initial rapid transcriptional upregulation [5], suggesting that rapid transcriptional upregulation may not be rate-limited by the recruiment of RNA polymerase II as suggested by current models [2].

2.1.2 Transcription factor activity profiles

The mRNA level of a transcription factor may not be an adequate indicator of changes in transcriptional regulatory activity *via* binding to the promoter regions of target genes. Transcription factors are activated by post-translational modification or ligand binding, and it is the activated form of a transcription factor which controls the transcription rate of a target gene [12]. I assume here that the transcription factor expression profile is not informative about the regulatory activity of a transcription factor, and instead treat the transcription factor activity – the quantitative effect of a transcription factor on the mRNA level of each target gene – as a hidden variable of interest. Given a gene expression timecourse, the task is now to describe the observed gene expression profile of each gene as a function of the hidden regulatory effects of transcription factors.

2.1.3 Study aims

The aim of this study is to attempt to explain the observed timecourse of gene expression levels during *S. cerevisiae* stationary phase exit and entry [5] as regulated gene expression profiles mediated by unobserved populations of active transcription factors with specific binding to target genes. In order to incorporate prior knowledge about potential transcription factor control of specific target genes, a transcription factor-target gene connectivity matrix was retrieved from previously analysed ChIP-chip datasets [76]. The connectivity matrix represents a consensus map of *in vivo* transcription factor-promoter binding specificity in *S. cerevisiae* across diverse conditions. Two previously described linear models of gene expression are applied, incorporating prior knowledge of potential transcription factor binding events. The observed gene expression profiles are interpreted as the result of transcriptional control by populations of active transcription factors with specific binding to target genes.

2.2 Datasets

2.2.1 Stationary phase exit and entry: a gene expression timecourse

I considered a nine-day microarray timecourse of *S. cerevisiae* glucose starvation stationary phase culture including exit from and entry to quiescence [5]. The timecourse contains 34 timepoints and is therefore potentially rich in transcription regulatory events related to stationary phase exit and subsequent re-entry (Figure 2.1). For this case study a processed external control normalized dataset was retrieved from ArrayExpress [77], accession number E-UMCU-12.

The arrays are two-channel cDNA arrays, as described in [38]. Genes are represented by duplicate spots on the array. Raw intensity data had been preprocessed using customized lowess print-tip normalization procedures as described in [5, 38]. An external control normalization procedure had been used in order to capture the

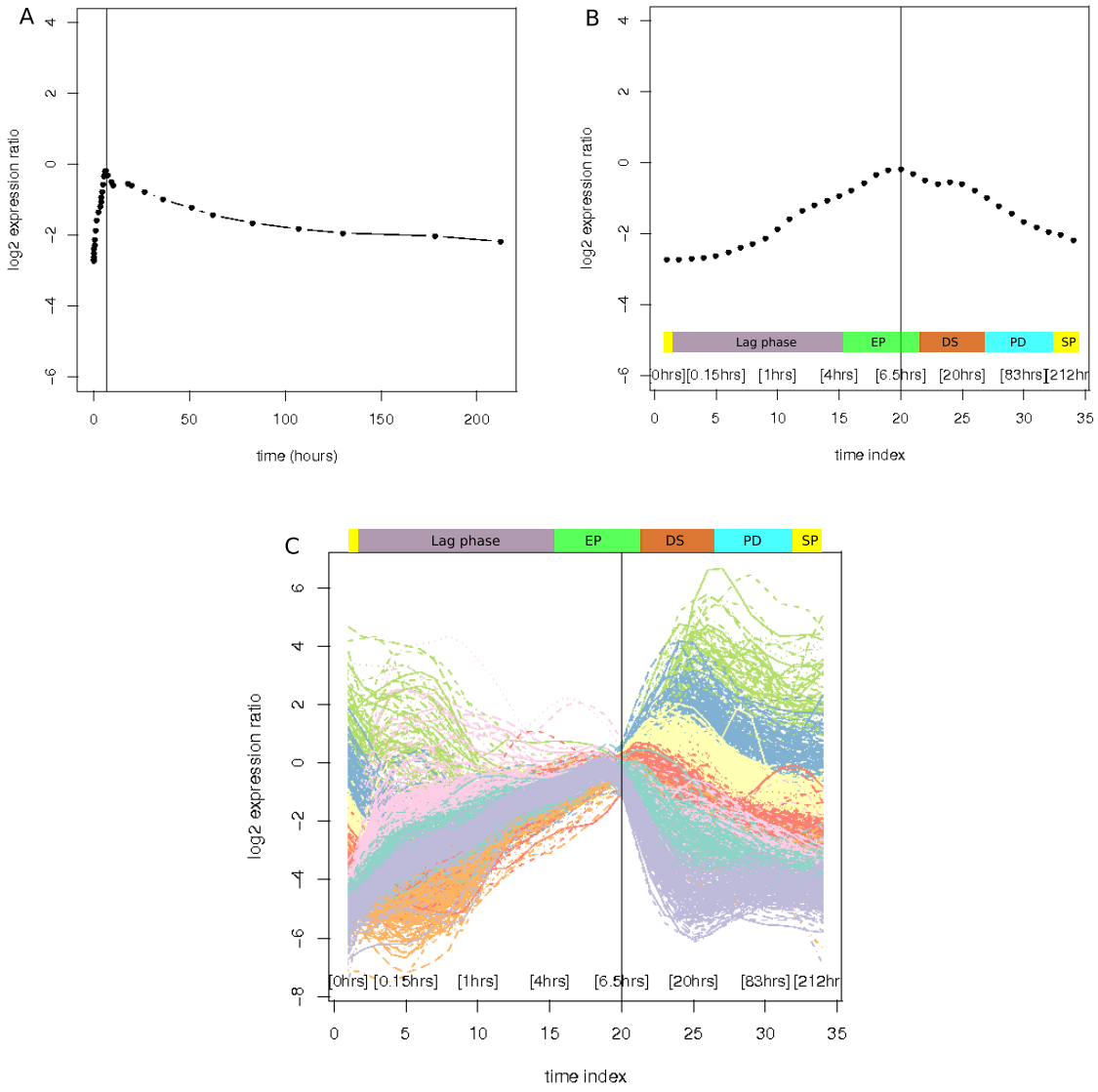


Figure 2.1: Gene expression profiles during stationary phase exit and re-entry. Expression profiles are shown as $\log_2$ ratio values compared to the mid-log reference timepoint ($t = 6.5$hrs, vertical line). **A**: Mean of all gene expression profiles versus time. Points indicate sampled timepoints. **B**: Mean of all gene expression profiles versus timepoint index. Corresponding times (hours) are shown in square brackets. The horizontal bar shows phases of the timecourse: lag phase, exponential phase (EP), diauxic shift (DS), post-diauxic shift (PD), stationary phase (SP). **C**: Representative clusters of genes with similar gene expression profiles. Shown are genes with $\geq$ 4-fold expression change, clustered by k-means clustering ($k = 8$). Colours indicate cluster membership. The dataset contains groups of genes with distinct patterns of gene expression during different phases of the timecourse [5].

global reduction in mRNA levels expected at stationary phase compared with exponential growth phase. The external control normalization was based on an invariant set of nine control RNAs spiked into all total RNA samples at known concentrations and spanning three orders of magnitude [38], and lowess lines had been fitted to external control spots within each print-tip subgrid on the array. Following normalization to external controls, the data had been processed further:

1. spot signals lower than 50 were replaced with 50;
2. signals were scaled using factors representing total RNA per cell at each timepoint
3. the resulting expression profile of each gene were smoothed using the cubic spline R function `smooth.spline` with `spar=0.4`;

The resulting 6357 $\times$ 34 (*genes* $\times$ *timepoints*) matrix of processed $\log_2$-scale expression profiles was used as the expression dataset in this study.

2.2.2 Transcription factor binding network

A binary (0-1) regulatory network was used to define the specificity of transcription factor binding to promoter regions of target genes. The network defines potential *in vivo* transcription factor binding events of the promoter regions of target genes. A draft *S. cerevisiae* transcription regulatory network structure had been defined by Harbison *et al.* [10] using ChIP-chip assays to detect *in vivo* binding (directly or indirectly) of 203 proteins to promoter regions in different environmental conditions. MacIsaac *et al.* [76] reanalysed the *S. cerevisiae* ChIP-chip data to produce a regulatory map of conserved sequence motifs representing transcription factor binding specificity. Each transcription factor-promoter interaction in the regulatory map is associated with a significance level (P) and a conservation score (C), the number of *sensu strictu* yeast in which this binding specificity is conserved. Here, I took the network defined at a significance level of $P < 0.001$ and a conservation score of

$C = 2$ [76] as an *a priori* binary network structure of potential transcription factor-promoter specific binding interactions in *S. cerevisiae*. This defines a connectivity matrix X, where

$$\begin{aligned} x_{gf} &= 1, \quad \text{if transcription factor } f \text{ binds to the promoter of gene } g, \\ x_{gf} &= 0, \quad \text{otherwise.} \end{aligned} \tag{2.1}$$

All rows and columns of X contain at least one non-zero element, so that all genes g indexed in X are bound by at least one transcription factor, and all indexed transcription factors f have binding specificity to at least one gene. Genes in X but not represented on the gene expression array were removed from X. This defined a connectivity matrix X of 117 transcription factors and 1909 genes. The matrix is sparse, with 1.7% of all elements being non-zero.

2.3 Methods

2.3.1 Transcription factor activity and hidden factor analysis

To explore hidden transcription factor activites and the effect on target genes during *S. cerevisiae* stationary phase exit and entry, I applied two previously described models [69, 68] of genome-scale regulation of gene expression. Both models aim to decompose a *genes* $\times$ *experiments* matrix of gene expression levels into two parts, a time-invariant weighted connectivity matrix describing the regulatory effect of each transcription factor on target genes, and a matrix describing the active levels of each transcription factor over experiments. The models differ in the formulation of the model and in the method of inference, and the two methods are complementary for exploratory analyses of genome-wide gene expression regulation during yeast stationary phase exit and entry.

2.3.1.1 Bayesian sparse hidden components analysis

First, I considered a sparse hidden components model proposed by Sabatti and James [69] and originally applied to the regulation of gene expression in *Escherichia coli*. Given a $G \times T$ matrix Y of log-scale gene expression measurements, where G is the number of genes and T is the number of experiments, the expression profile y_g of gene g is modelled as

$$y_{gt} = \sum_f a_{gf} p_{ft} + \gamma_{gt} \tag{2.2}$$

where a_{gf} is the regulatory strength of transcription factor f on gene g, p_{ft} may be interpreted as the concentration of the active form of the transcription factor f, and $\gamma_g t$ is an error term with $\gamma_{gt} \sim \mathcal{N}(0, \sigma_g^2)$. To apply this model to yeast stationary phase exit and entry timecourse, the T experiments in the model were taken to be the 34 timepoints in the timecourse.

The model is in general overparameterized, and is therefore solved using a Bayesian inference method. The model also infers an *a posteriori* connectivity matrix Z where

$$\begin{aligned} z_{gf} &= 1, \quad \text{if transcription factor } f \text{ binds to the promoter of gene } g, \\ z_{gf} &= 0, \quad \text{otherwise.} \end{aligned} \tag{2.3}$$

. Firstly, I used the connectivity matrix X, described above (Datasets, page 23), to define the prior distribution on the learned network structure Z. Following Sabatti and James [69], I set the prior distribution on Z to

$$\begin{aligned} Pr(z_{gf} = 1) &= 0.5, \quad \text{if } x_{gf} = 1, \\ Pr(z_{gf} = 1) &= 0, \quad \text{otherwise.} \end{aligned} \tag{2.4}$$

This prior is only mildly informative for the *a posteriori* structure of the connectivity matrix. Secondly, the posterior mean and variance of a_{gt}, p_{ft}, $\gamma_{gt} \mid Z$ is estimated, for given prior distributions. The prior distributions for a_{gt}, p_{ft}, γ_{gt} are independent

Gaussian distributions,

$$a_{gf} = 0 \quad \text{if } z_{gf} = 0, \quad a_{gf} \sim \mathcal{N}(0, \sigma_a^2 = 10000) \quad \text{otherwise;}$$
$$p_{ft} \sim \mathcal{N}(0, \sigma_p^2 = 1), \quad \gamma_{gt} \sim \mathcal{N}(0, \sigma_g^2) \tag{2.5}$$

where the error variance σ_g is modelled as an inverse gamma distribution with parameters $\alpha = 0.7, \beta = 0.3$.

Parameter inference was performed using a Markov Chain Monte Carlo method designed and implemented in R [78] by Sabatti and James [69]. The method outputs estimates for the mean and standard deviation of the posterior distribution of p_{ft}, the active transcription factor concentrations, and a_{gf}, the regulatory strength of transcription factor f on gene g. The sign of a_{gf} and p_{ft} is interchangable, so in addition the output is summarized as two quantities, (i) pav_{ft}, the average effect of transcription factor f over all the genes it regulates, and (ii) ave_{gf}, the average regulatory strength of transcription factor f on gene i over all experiments:

$$pav_{ft} = \frac{\sum_g a_{gf} p_{ft}}{\sum_g I(a_{gf} \neq 0)}, \quad ave_{gf} = \frac{\sum_t a_{gf} p_{ft}}{T} \tag{2.6}$$

where $\sum_g I(a_{gf} \neq 0)$ is the number of genes regulated with a non-zero control strength by transcription factor f, and T is the number of timepoints ($T = 34$). Quantities pav_{ft}, ave_{gf} are estimated as the *a posteriori* mean with an associated standard deviation.

2.3.1.2 Probabilistic dynamic model of transcription factor activity

Second, I adopted a dynamic linear model of gene expression proposed by Sanguinetti and co-workers [68] and tested by the authors on *S. cerevisiae* cell cycle mRNA and metabolic datasets. The model was applied here in order to estimate the regulatory effect of transcription factors on target genes during stationary phase exit and entry.

The expression level of each gene over time is modelled as a linear combination of

the regulatory effects of each binding transcription factor. An binary connectivity matrix X of transcription factor-promoter interactions is assumed to be known *a priori* (see Datasets, page 23) given by:

$$\begin{aligned} x_{gf} &= 1, \quad \text{if transcription factor } f \text{ binds to the promoter of gene } g, \\ x_{gf} &= 0, \quad \text{otherwise.} \end{aligned} \tag{2.7}$$

Given a $G \times T$ matrix Y of log-scale gene expression measurements, where G is the number of genes and T is the number of timepoints, the expression profile y_g of gene g is modelled as

$$y_g(t) = \sum_f x_{gf} b_{gf} c_f(t) + \mu_g + \epsilon_{gt} \tag{2.8}$$

where $t \in 1...T$, x_{gf} is the binary connectivity (0 or 1) of transcription factor f to gene g, μ_g is the baseline expression level for gene g (in the absence of transcription factor binding), and ϵ_{gt} is a gene-specific, time-specific error term. The weight b_{gf} of transcription factor f on the expression of gene g may be interpreted as the regulatory strength of transcription factor f acting on gene g. In this model, b_{gf} is constant in time. The time-varying part of the model, $c_f(t)$, is a property of transcription factor f and may be interpreted as the time-varying concentration of an active form of the transcription factor.

This model is again overparameterized but can be solved using an approximate Bayesian inference approach, described in detail in [68]. In this approach the posterior probabilities of transcription factor concentrations $c_f(t)$ and regulatory strengths b_{gf} are inferred given a prior distribution for each parameter in the model. Sanguinetti *et al.* [68, 79] placed two constraints on the prior distribution of $c_g(t)$: (i) the prior distribution is stationary in time, and (ii) the distribution of $c_g(t)$ depends only[1] on $c_g(t-1)$. Thus $c_f(t)$ is modelled as

$$c_f(t) = \gamma_f c_f(t-1) + \eta_{ft}. \tag{2.9}$$

where $\gamma_f \in [0,1]$ captures the time correlation between adjacent timepoints and η_{ft}

[1] This is the Markov property, used here to capture correlations between adjacent timepoints.

is the process noise, with prior distributions taken to be

$$\eta_{ft} \sim \mathcal{N}(0, 1 - \gamma_f^2), \quad c_g(1) \sim \mathcal{N}(0, 1). \tag{2.10}$$

The priors on the regulatory strengths b_{gf}, the baseline expression levels μ_g and the error term ϵ_{ft} are taken to be Gaussian,

$$b_{gf} \sim \mathcal{N}(0, \alpha^2), \quad \mu_g \sim \mathcal{N}(0, 1), \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{2.11}$$

Parameter inference was performed using a variational Expectation-Maximization inference method [80], designed and implemented by Sanguinetti and co-workers, in the MATLAB toolbox ChipVar [68]. The method generates estimates for the mean[2] of the posterior distribution of $c_f(t)$, the active transcription factor concentrations; and b_{gf}, the regulatory strength of transcription factor f on gene g.

2.4 Results

2.4.1 Bayesian sparse hidden components analysis

Using hidden components analysis, 96 of the 117 transcription factors in the prior connectivity matrix (X) were found to have non-zero control strength on one or more target genes ($abs(a_{gf}) > 0$). The average control strength ($pav_f(t)$) of each transcription factor (f) on its *a posteriori* target genes is summarized as a cluster diagram (Figure 2.2; Table 2.1).

For the transcription factors with variable average control strength ($p_g(t)$) over the timecourse, it is interesting to look at the average control strength over time for each transcription factor acting on each inferred target gene. In particular, it is possible to identify genes which are controlled by combinations of transcription factors during

[2]The current software implementation, ChipVar v0.11, does not report estimates for the variance of the posterior distribution of $c_f(t)$, b_{gf}. Analysis is therefore restricted here to estimates of the size (posterior mean) of $c_f(t)$, b_{gf}, and the statistical significance of control of transcription factor on each target gene is not estimated.

the timecourse under this model. Table 2.2 lists commonly identified pairs of transcription factors which control multiple ($\geq$ 5) target genes with non-zero control strength ave_{gf}. There are 18 such transcription factor pairs (involving 25 transcription factors) identified *a posteriori* for the gene expression timecourse, compared to 163 such pairs (involving 68 transcription factors) in the original connectivity matrix X.

2.4.2 Probabilistic dynamic model of transcription factor activity

The time-varying concentration of the active form of each transcription factor in the connectivity matrix X were inferred under the probabilistic dynamic model of gene expression. Time-independent regulatory control strengths of each transcription factor on each potential target gene was also inferred. The inferred active concentrations of the 117 transcription factors in connectivity matrix X are shown as a cluster diagram (Figure 2.3; Table 2.3)[3]. Note that due to the model formulation (Eqn. 2.8) there is a sign ambiguity for the active concentration, $c(t)$ and associated regulatory strengths. Comparing Tables 2.1 and 2.3 it can be seen that there is consistency between:

(i) $pav_f(t)$, the average effect of transcription factor f on each target gene, inferred by sparse hidden components analysis.

(ii) $c_f(t)$, the concentration of the active form of transcription factor f, inferred under the probabilistic dynamic model; and

For example, *OPI1* and *SIP4* are inferred to have a differentially active role during stationary phase exit and from post-dauxic phase into stationary phase, compared to the mid-log reference timepoint. Under both models, *DAL81* is inferred to have an active control in stationary phase exit compared to mid-log phase, but an opposite regulatory effect during dauxic shift.

[3]The regulatory strengths b_{gf} of transcription factor f on gene g were not analysed further due to the absence of variance estimates for the posterior distributions of b_{gf}.

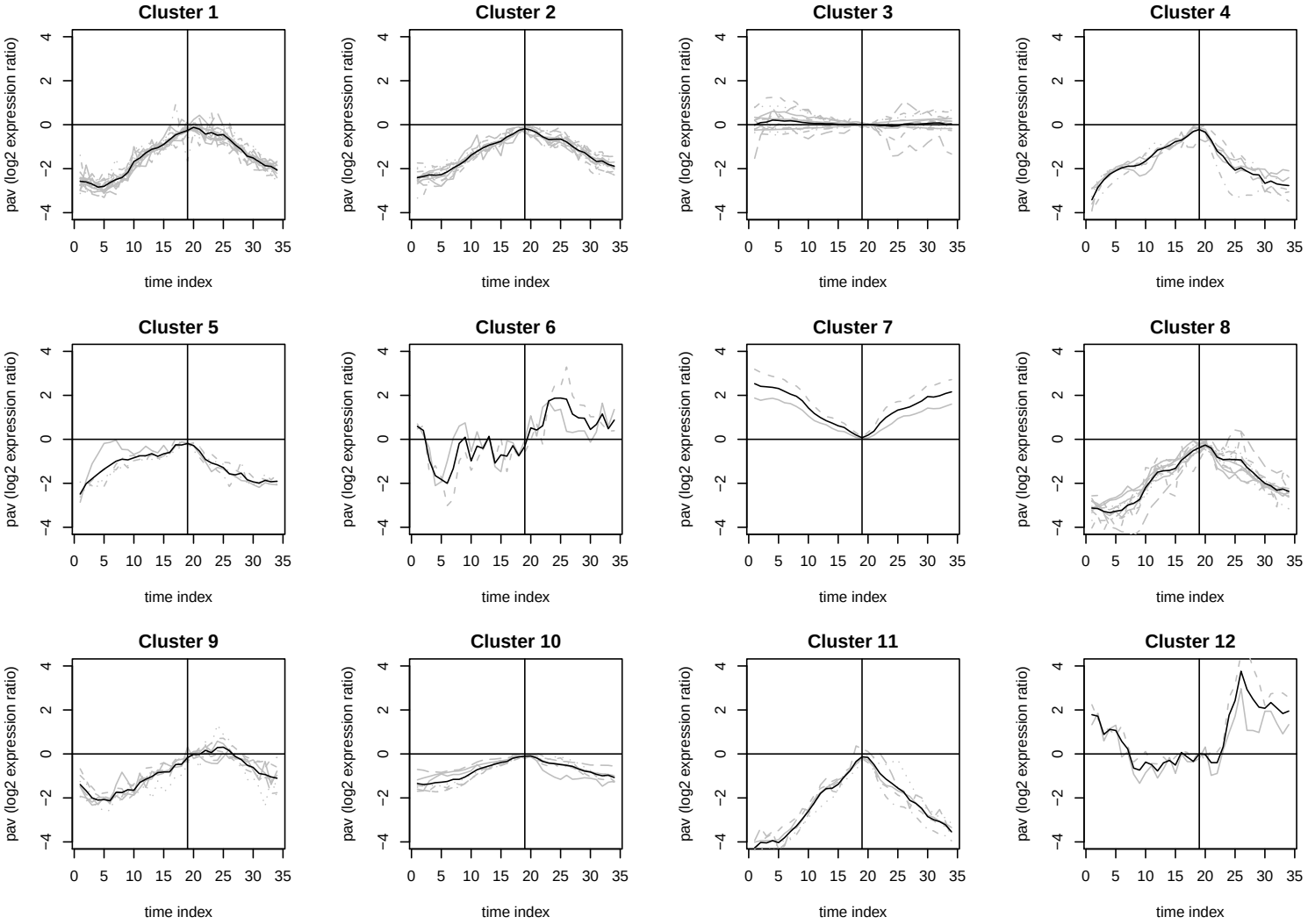


Figure 2.2: Transcription factors are clustered according to average control strength ($pav_f(t)$) of each transcription factor f on its target genes, inferred using Bayesian sparse hidden components analysis. Only the 96 transcription factors with non-zero *a posteriori* control strengths on one or more target genes are shown. Black lines depict cluster means. Grey lines depict $pav_f(t)$ for transcription factors f. The mid-log reference timepoint is indicated by a vertical line. Clustering was performed using k-means clustering with $k = 12$, using the R package `kmeans`. Transcription factor names are listed in Table 2.1.

Table 2.1: Table of transcription factors with similar average control strength ($pav_f(t)$) inferred using Bayesian sparse hidden components analysis (see Figure 2.2)

Cluster 1	ACE2, ARG80, CAD1, CIN5, DIG1, FKH2, GAT1, HAC1, IME1, MATA1, MCM1, NDD1, PHO2, SPT2, SPT23, STB2, STP1, TYE7, ZAP1
Cluster 2	ADR1, CHA4, DAL82, FKH1, GCN4, GTS1, MBP1, NRG1, PDR3, RCS1, RDS1, STE12, SUM1, YAP7
Cluster 3	HAP2, HAP5, MET32, PHO4, PUT3, RTG3, SFP1, SKN7, SKO1, SNT2, SUT1, THI2, UME6, YOX1
Cluster 4	FHL1, GCR2, INO2, LEU3, RLM1, YDR520C
Cluster 5	ROX1, STB5, XBP1, YAP1
Cluster 6	DAL81, MIG1
Cluster 7	MSN4, SOK2
Cluster 8	ABF1, ARR1, BAS1, CST6, GZF3, HAP4, MAC1, MET31, PHD1, REB1, RPN4, STB1, TEC1, YAP5
Cluster 9	DAL80, GAL4, HAP3, MOT3, MSN2, STB4, YAP6
Cluster 10	AFT2, CBF1, GLN3, HAP1, INO4, RFX1, SWI6
Cluster 11	HSF1, RAP1, RME1, SWI4, SWI5
Cluster 12	OPI1, SIP4

Table 2.2: Table of transcription factors pairs that control 5 or more target genes (control strength $ave_{gf} > 0$), inferred using Bayesian sparce hidden components analysis.

$\geq$ 6 genes	5 genes
SKN7 / SWI4	STE12 / SWI4
SWI6 / SWI4	PHD1 / SWI4
TYE7 / CBF1	RAP1 / UME6
MSN4 / MSN2	SKN7 / CBF1
SUT1 / SKN7	STE12 / SWI6
CIN5 / SKN7	STE12 / SKN7
PHD1 / SKN7	NRG1 / SKN7
FHL1 / RAP1	SWI6 / STE12
DIG1 / STE12	
MBP1 / SWI6	

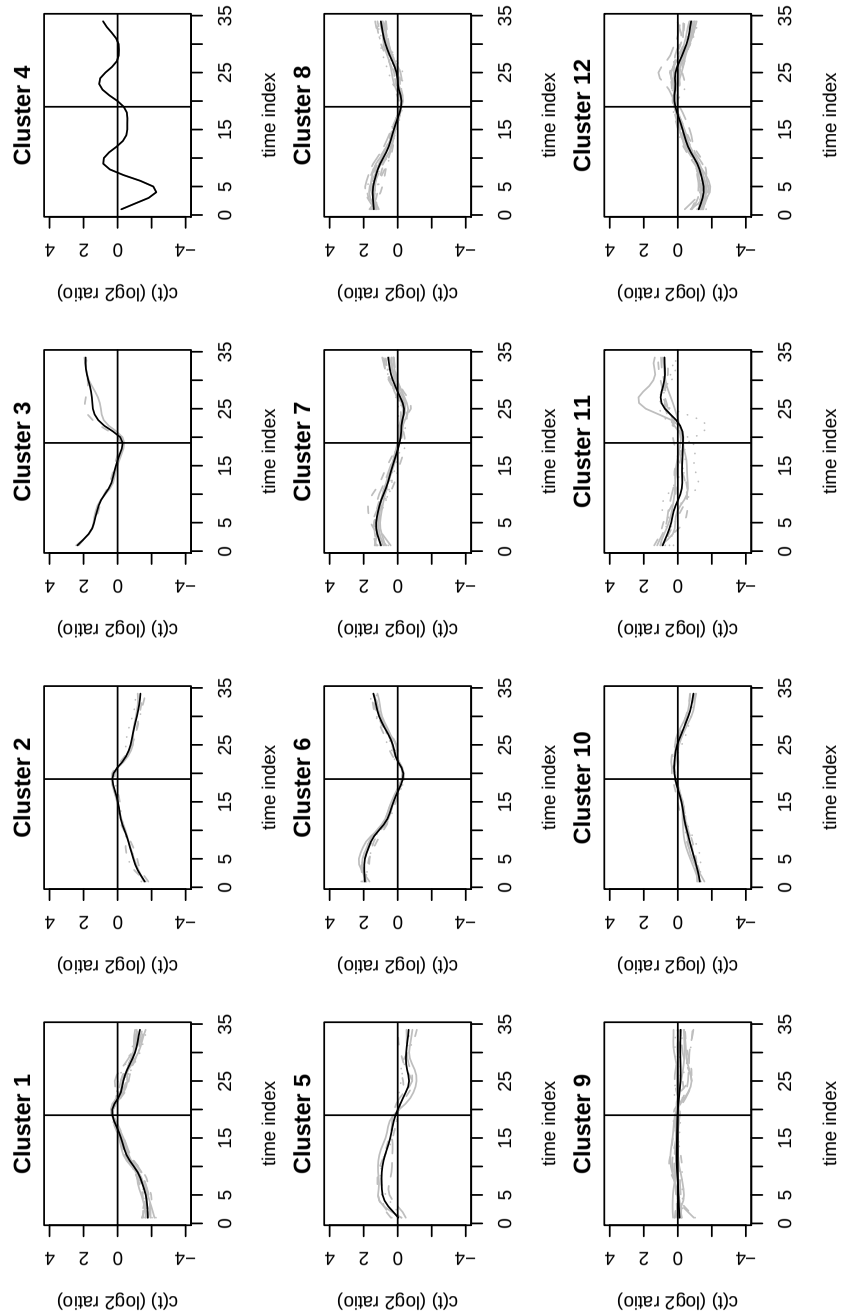


Figure 2.3: Transcription factors are clustered according to active transcription factor concentration $c(t)$, inferred using a probabilistic dynamic model [68]. Black lines depict cluster means. Grey lines depict $c(t)$ for each transcription factor. The mid-log reference timepoint is indicated by a vertical line. Clustering was performed using k-means clustering with $k = 12$, using the R package `kmeans`. Transcription factor names are listed in Table 2.3.

Table 2.3: Table of transcription factors with similar active concentrations over time ($c_f(t)$) inferred using a probabilistic dynamic model (see Figure 2.3)

Cluster 1	ABF1 SWI4 CBF1 RAP1 CIN5 STB1 SWI5 FKH1 RME1 TYE7 BAS1 YAP5 IME1
Cluster 2	GCR2, UME6, HAP5, YDR520C
Cluster 3	LEU3, FHL1
Cluster 4	DAL81
Cluster 5	PUT3, MSN2, RPH1, STP4, MIG1
Cluster 6	RPN4, STE12, INO4, PHD1, REB1, MBP1, GCN4, HAP1
Cluster 7	HAP3, SPT2, GZF3, PDR3, SKN7, MOT3, SNT2, DAL80, PHO4, GLN3, GAL4, HAP4, GTS1
Cluster 8	ARG80, MAC1, ADR1, CHA4, SPT23, DAL82, YAP1, AFT2, HAC1, MSN4, TEC1, HAP2, MET31, RDS1
Cluster 9	YHP1, AZF1, YML081W, PDR1, MET4, GAT3, SKO1, IXR1, ROX1, SFP1, ASH1, SMP1, MET32, UGA3, GCR1, XBP1, YAP3, YRR1, RIM101, ARO80
Cluster 10	RLM1, MATA1, STB2, CAD1, INO2, ARG81
Cluster 11	YOX1, STB5, GAL80, OPI1, RFX1, SIP4, RGT1, RLR1
Cluster 12	PHO2, ACE2, ZAP1, SOK2, FKH2, SUM1, GAT1, ARR1, RCS1, SUT1, YAP6, NDD1, MCM1, CST6, HSF1, RTG3, THI2, DIG1, SWI6, STB4, YAP7, NRG1, STP1

2.5 Discussion

The active transcription factor concentrations $c(t)$ inferred using the probabilistic dynamic model are qualitatively similar to the average control strengths (a_{gf}) inferred using the hidden components analysis. This suggests that the decomposition of the gene expression matrix into a time-varying part ($c_f(t)$ or pav_{ft}) and a gene-specific part (b_{gf} or a_{gf}) is robust to differences in model formulation and method of inference: the gene expression matrix can be expressed as a linear combination of transcription factor activities, weighted by gene-specific control strengths, consistently by two different methods.

The Bayesian sparse hidden components model as implemented does not assume that the experiments form a timecourse, and therefore treats the active concentrations of a transcription factor (p_{gt}) at sequential timepoints as *a priori* independent. As noted by Sabatti and James [69], this can in principle be modified to allow for covariance of p_{gt}, p_{gs} between timepoints by selecting an alternative prior[4] but adds computational complexity.

The two models considered here exploit the sparsity of the prior connectivity map, X, to explain the observed gene expression profiles as a linear model of transcriptional control by transcription factors with binding specificity to target genes. The connectivity matrix X does not necessarily represent direct DNA-binding by the relevant transcription factor. The ChIP-chip datasets from which the connectivity matrix was generated includes transcription factors which are not directly DNA-binding. An example of indirect connectivity of a transcription factor to a promoter binding site is the Gal80-Gal4 complex noted by MacIsaac and co-workers [76]. Gal80 inhibits the regulatory control of Gal4 on its target genes by binding to Gal4, but Gal80 does not itself have a DNA-binding domain. The connectivity matrix is therefore not a direct indication of DNA-binding specificity, but indicates that a given transcription factor is associated with specific promoter motifs. Simi-

[4] for example, $p_g \sim \mathcal{N}(0, \Gamma)$ where Γ represents covariance between p_{gt}, p_{gs}.

larly, the models applied here do not necessarily imply that the stated transcription factors have a direct effect on inferred controlled target genes by binding to the promoter region of the target genes. Rather, the models are used to explain the expression levels of target genes in terms of motifs which had previously been found in associated promoter regions, and these motifs in turn are associated with direct or indirect DNA-binding *in vivo* by transcription factors based on ChIP-chip assays.

The original ChIP-chip experiments by Harbison and co-workers were performed in a number of different experimental conditions [10]. The authors of that study found that there was variation in binding specificity between different conditions [10], and it is likely that further binding specificities would be found under a different set of conditions [81]. The *a priori* connectivity matrix, X, may therefore be missing transcription factor-gene interactions that are important during any of the intermediate phases included in the timecourse covering stationary phase exit or entry by glucose limited conditions. The set of 207 transcription factors studied by Harbison and colleagues [10] represented known and predicted transcriptional regulators and is not an exhaustive list of all possible transcriptional regulators. The transcription factor *GIS*1, for example, was recently confirmed to be a key transcriptional regulator in nutrient-limited conditions [82] but is not one of the 207 transcription factors in the ChIP-chip dataset [10]. Furthermore, only 1909 of the 6357 genes contained in the gene expression dataset were considered to be potential targets of one or more of the 207 transcription factors. More than half of all genes were therefore not included in the two models considered here because no potential transcription factor-gene specific binding had been predicted *a priori* for these genes in the connectivity matrix, X.

The problem of quantifying on a genome-wide scale the regulatory control of target genes by transcription factors across a limited number of experiments or timepoints has been studied extensively in recent years. The models applied in this study [69, 68] are two examples amongst a number of related models and associated inference methods which use linear models to explain observed mRNA levels.

Detailed kinetic models of changes in mRNA abundance during a timecourse study have been applied to small networks involving a handful of genes [83, 70]. Barenco and coworkers [70] modelled mRNA abundance as a function of transcription rate, mRNA degradation rate and hidden transcription factor activity, limited to the construction of small networks of transcriptional regulation using a gene expression timecourse dataset. Nachman and coworkers [83] proposed a dynamic Bayesian network model of mRNA abundance as a function of transcription factor protein concentrations.

On a whole-genome scale, the use of such kinetic models is hampered by the resulting complexity of detailed kinetic models, the lack of kinetic data on the genomic scale, and the problem of estimating parameters in an overdetermined system. Thus a number of simpler models of transcriptional control have been proposed on the genomic scale. Segal and coworkers [84] proposed fitting a module network – a simplified Bayesian network model of genome-wide mRNA abundance which assumes that there are groups of similarly regulated genes – in order to reduce the number of regulatory relationships that must be inferred from a genome-wide dataset. The model seeks to explain gene expression in terms of dependencies on the gene expression profile of given transcriptional regulators. The ARACNE algorithm [85] uses mutual information between expression profiles to define a network of dependency and putative transcriptional control. In contrast, the two models applied in this chapter [69, 68] do not take into account the mRNA abundance of transcription factors. Instead, these models assume that gene expression profiles can be expressed as a linear combination of an unobserved property of transcription factors, which may be interpreted as the concentration of an active form of the transcription factor. A number of similar linear models of gene expression have been proposed which use prior knowledge of transcription factor binding specificity or binding motifs shared between coexpressed genes [86, 71, 87]. Each model of transcriptional control of gene expression exploits a subset of prior knowledge in an effort to uncover insights about transcriptional regulation. A model which includes both

the changes in mRNA abundance of transcription factors and the unobserved post-transcriptional activation of transcription factors, acting on common sets of genes, may in future prove to be a more enlightening model of genome-wide regulation of gene expression.

This chapter has considered only observed changes in mRNA abundance over time, and has sought to explain genome-wide changes in mRNA abundance as a linear combination of the effects of DNA-binding transcription factors. Under current models, activation or repression of transcription by transcription factors alters the rate of transcription of a given mRNA species, and the modulation of the transcription rate in turn affects the mRNA abundance. Observed mRNA abundance is the result of a balance between the rate of transcription and the rate of mRNA degradation. To understand how an observed timecourse of mRNA abundance may be controlled by modulating the transcription rate *via* transcription factor-DNA interactions, it is necessary to consider the observed changes in mRNA abundance as the result of both a transcription rate and an mRNA degradation rate, either of which may potentially vary with time. The next chapter is concerned with dynamic models of mRNA abundance, taking into account the potential effect of regulated mRNA degradation rates.

Chapter 3

Detecting regulated mRNA stability using microarray measurements: models and applications

This chapter describes an approach for identifying changes in mRNA stability in response to an environmental stress, using simultaneous microarray timecourse measurements of changes in mRNA abundance and transcription rate. This approach is used in Chapter 4 to identify genes which are candidates for regulated mRNA stability in response to an oxidative stress in *Schizosaccharomyces pombe*.

Given two simultaneous timecourses of changes in transcription rate and mRNA abundance in response to stress, are they consistent with constant mRNA stability, or is there evidence of post-transcriptional control of mRNA levels? This chapter presents a model-based approach towards understanding transcriptional and post-transcriptional contributions to shaping a gene expression response. A simple model of mRNA kinetics is adapted so that it becomes applicable to two simultaneous microarray time courses measuring changes in mRNA abundance and transcription rate.

3.1 Introduction

It has recently become feasible to produce microarray timecourse datasets of changes in transcription rate using run-on assays [88] or, as considered in the subsequent study, recently designed RNA polymerase II ChIP-chip experiments. High time resolution datasets can capture changes in both mRNA levels and transcription rates during a global gene expression response. This presents the possibility of directly modelling the observed mRNA dynamics, without recourse to transcriptional inactivation which is disruptive to the cell and has compounding effects. Previous studies have simulated an mRNA abundance timecourse for a given observed transcription rate timecourse together with a sample of possible mRNA halflives [89, 90] and unknown scaling values to account for unknown absolute values [91], and have then compared the simulated changes in mRNA levels to observed timecourse measurements. In another approach, first-order mRNA decay was fitted to sequential time intervals [92, 89] but this did not yield biologically plausible results. Jiang and colleagues [91] calculated a timecourse of changes in mRNA degradation rates for five genes, and noted that the result was dependent on the unknown relative scalings between measured mRNA abundance and transcription rates. There has not yet been a systematic attempt to fit analytic solutions of mRNA degradation models to observed timecourse data, whilst explicitly taking into account the unknown relative scaling of transcription rate and mRNA abundance measurements which are a consequence of measuring relative values.

3.1.1 Chapter outline

First, this chapter describes the simple kinetic models used to study mRNA turnover and the effect on mRNA abundance on a genome-wide scale. Previous genome-wide studies of regulated mRNA stability are described. Second, a framework is presented for identifying candidate genes for regulated mRNA stability using simultaneous microarray measurements of changes in transcription rate and mRNA

abundance.

3.2 Previous work: genome-wide mRNA degradation rates and mRNA abundance

Recent genome-wide studies of mRNA turnover aim to (i) rank the relative decay rates of mRNA species within a population of mRNA in a fixed condition, or (ii) identify changes in the decay rate of mRNA species in response to changes in environmental or physiological conditions. The relative ranking of mRNA stability amongst a population of mRNA species can be estimated by stopping transcription and estimating the subsequent rate of loss of transcripts using a timecourse of microarray measurements. Changes in mRNA stability may be detected by comparing steady-state estimates of mRNA stability or, for a system which is not at steady-state, a dynamic model of mRNA abundance may be fitted to simulteneous timecourses of mRNA abundance and transcription rate measurements.

3.2.1 mRNA degradation as a first-order decay process

For each mRNA species, mRNA abundance *E(t)* is determined by the transcription rate *R(t)* and the degradation rate *D(t)*. Assuming that the growth rate of cells is negligible, and that there is no time delay between transcription and contribution to extracted mRNA abundance (e.g. negligible time delay between transcription initiation and transcript termination) then mRNA abundance *E(t)* is simply related to the transcription rate *R(t)* and the degradation rate D(t) by:

$$\frac{dE}{dt} = R(t) - D(t) \tag{3.1}$$

The assumption of negligible growth rate is valid for the oxidative stress response described in Chapter 4 (in which cell growth is arrested throughout the stress response) and the case of non-negligible growth rate is not considered. A number of further simplifying assumptions are implied in Eqn 3.1. mRNA abundance *E(t)*

is the abundance of mature polyadenylated mRNA extracted from the sample and hybridizing specifically to a complementary probe on the array. mRNA present at various stages of transcript processing or degradation is not explicitly modelled. The transcription rate *R(t)* is here the rate at which hybridizing RNA is produced, and the degradation rate *D(t)* is simply the overall rate at which hybridizing RNA is lost from the cell. In particular, if the transcription rate is estimated as the rate of production of nuclear pre-mRNA (for example using RNA polymerase II ChIP-chip or nuclear run-on assays), the degradation rate $D(t)$ may include loss or decay of transcripts by any mechanism including intermediate steps from nuclear transcription to mature polyadenylated transcripts. The kinetics of intermediate steps in mRNA metabolism and specific degradation pathways are not explicitly modelled (see Cao and Parker [93] for a 21-parameter model of the metabolism of a single mRNA species).

The mRNA degradation rate in a given condition is typically modelled as a first-order decay process [94, 95, 96, 97]; that is, the degradation rate *D(t)* is proportional to mRNA abundance:

$$D(t) \equiv kE(t) \tag{3.2}$$

where k is the decay *rate constant*. Assuming zero-order mRNA synthesis and first-order mRNA decay, mRNA abundance *E(t)* and transcription rate *R(t)* satisfy

$$\frac{dE}{dt} = R(t) - kE(t) \tag{3.3}$$

The mRNA halflife $\tau_{\frac{1}{2}}$ is typically used to summarize mRNA stability and is related to the decay rate constant k by $\tau_{\frac{1}{2}} = ln(2)/k$. Assuming first-order degradation with decay rate constant k, then given any time-varying transcription rate *R(t)*, Eqn 3.3 may be solved for the mRNA abundance *E(t)*:

$$E(t) = \int_{t_0}^{t} R(t')e^{k(t'-t)}dt' + E_0e^{-kt} \tag{3.4}$$

At steady state (*ss*), dE/dt = 0 so that Eqn. 3.3 becomes:

$$E_{ss} = \frac{R_{ss}}{k} \tag{3.5}$$

The decay rate constant of an mRNA species can be estimated directly by blocking transcription. mRNA abundance then follows an exponential decay, $E(t) = E_0 e^{-kt}$, from an initial mRNA level E_0. Hargrove *et al.* [94] and recent genome-wide studies [96, 98] have found that mRNA decay curves are exponential, indicating that first-order decay is a reasonable model for mRNA decay kinetics. Recent studies have reported genome-wide mRNA decay rates, or ranked mRNA decay rates within a population of mRNA species, in several organisms and cell types by inhibiting RNA polymerase II transcription and fitting an exponential decay to the subsequent loss of mRNA. mRNA decay rates have been estimated in *S. cerevisiae* [96, 98, 99], *S. pombe* [47], human cell lines [97], *Arabidopsis* [100], and *Mus musculus* embryonic stem cells. Decay rates have been variously associated with protein functional classes (*e.g.* [97, 100]; transcription factors have short mRNA half-lives, whereas biosynthesis genes have long mRNA halflives); membership of protein complexes [96]; gene length and number of introns [47, 100, 101]; and the presence of mRNA motifs, including AU-rich elements and PUF binding motifs which are known to have functional significance for mRNA stability [100, 101].

3.2.2 Detecting changes in mRNA stability

3.2.2.1 Comparison of mRNA decay rates

Recent studies have compared mRNA halflife estimates using transcriptional inactivation across conditions. There is evidence that the control of mRNA decay rates contributes to the genome-wide regulation of gene expression in response to environmental stress. Shalem and co-workers [98] compared mRNA halflives and mRNA abundance profiles between two stress conditions (weak oxidative stress, DNA damage) and exponentially growing cells in *S. cerevisiae*. Genes with rapid transient accumulation (loss) of mRNA are reported to be destabilized (stabilized)

compared to exponentially growing cells, consistent with rapid relaxation of mRNA levels to pre-stress levels observed for transiently induced or repressed genes. In contrast, the transcripts of persistently induced (repressed) genes tended to be stabilized (destabilized) compared to exponentially growing cells. mRNA halflives were estimated for one time period per stress response (oxidative stress: from 25 mins; DNA damage: from 40 mins after stress); an interesting question is whether mRNA stabilization/destabilization occurs immediately upon induction of a stress response or some time after stress induction, and whether changes in mRNA stability are localized in time or represent continual changes in stability. Molin *et al.* [99] studied the timing of changes in mRNA decay rates during a transient stress response (weak salt stress) by stopping transcription before stress induction and at two timepoints (6, 30 mins) after stress induction. The authors find that for some – but not all – genes which are highly and transiently induced (repressed), mRNA is rapidly stabilized (destabilized) at the onset of stress but subsequently destabilized (stabilized) close to the peak in mRNA levels. This indicates that, for some transient stress-response genes, rapid accumulation (loss) of transcripts may be facilitated by rapid early stabilization (destabilization) of mRNA, whereas the return to pre-stress mRNA levels may be driven by mRNA stabilization (destabilization) close to peak mRNA levels.

A limitation of using transcription inhibition experiments to study changes in mRNA decay rates is the impact of the transcriptional inactivation on the cell. RNA polymerase II transcriptional inactivation is achieved by exposure to a transcription-blocking drug such as 1,10-phenanthroline or using a temperature-sensitive RNA Pol II mutant, and can itself induce a specific stress response [102]. In addition, inhibiting transcription during a primary stress response results in compounded responses: any post-transcriptional processes which are continuously modulated in response to stress will continue to respond to the primary stressor during the period of transcriptional inhibition.

3.2.2.2 Comparison of mRNA abundance and transcription rates

An alternative approach to detecting regulated mRNA turnover, without inhibiting transcription, is to measure changes in both mRNA abundance and transcription rate and determine whether this is consistent with constant mRNA turnover, or whether the data imply a role for post-transcriptional control of mRNA levels. Changes in transcription rate can be measured on a genome-wide scale by combining nuclear run-on assays with microarray hybridization [88, 103] or using a recently developed RNA polymerase II ChIP-chip assay (S. Marguerat, personal communication; see Chapter 4).

In steady-state conditions it is possible to identify genes with putative altered mRNA decay rates by comparing fold-changes in transcription rates to fold-changes in mRNA abundance. An observed discrepancy between the fold-change in transcription rate and the fold-change in mRNA abundance from stressed to unstressed cells indicates that the mRNA decay rate is modulated between different conditions (Eqn. 3.5) and suggests post-transcriptional control of the gene expression response to stress.

Using this approach, genes with putative modulated mRNA decay rates in response to stress have been identified in *S. cerevisiae* [90], human lung carcinoma cells [104] and tobacco plant and *Chlamydomonas* plastids [105, 106]. However, this approach is only applicable to the comparison of steady-state conditions (in which there is no change in mRNA level, transcription rate, and any other kinetic parameters which affect mRNA metabolism). Although this approach has been used to identify changes in mRNA decay rate across a stress response timecourse under the assumption of sequential steady states (*e.g.* [90]), in general a timepoint-by-timepoint comparison of sampled non-steady-state measurements ignores the dynamics of transcription and degradation which shape the mRNA abundance response over time. Given the mounting evidence that mRNA decay rates of a subset of stress-induced and stress-repressed genes are modulated in response to stress, it is of interest to

find at what time a change in decay rate takes place, whether immediately upon exposure to stress or later in the stress response, and whether there is a rapid localized change or a gradual, continual change in transcript decay rate.

3.2.2.3 Modelling timecourse datasets of mRNA abundance and transcription rates

A special case of stress response timecourse datasets is the detection of changes in either transcription rate or mRNA levels together with no detected change in the other. Cheadle and colleagues [103] tracked the genome-wide transcription rate (using nuclear run-on arrays) and mRNA levels of human Jurkat T-cells for 60 minutes after activation. Most genes fell into two groups: (i) a change in mRNA levels with no detected change in transcription rate, and (ii) rapid transcriptional induction with no detected increase in mRNA levels. The first group may be regulated post-transcriptionally. The authors note that the second group may have been identified due to differences in detection between changes in transcription rate and mRNA levels. Such differences may be caused by restricted dynamic range of detected intensity values, for example, which can result in loss of detection of small changes, or underestimation of large changes due to intensity saturation.

More generally, continual changes in both transcription rate and mRNA abundance may be observed. In an early study, Jiang *et al.* [91] measured the transcription rate (nuclear run-on assay) and mRNA accumulation (Northern blot) in response to stimulus for five genes induced in the human acute response (inflammation). Firstly, the authors simulated mRNA levels using the observed transcription rate changes and a sample of possible mRNA halflives, showing that three genes did not appear to be consistent with a constant halflife. Secondly, they simulated the effect of an immediate rapid fold-increase in mRNA stability, and noted that three of the five genes were consistent with a rapid early change in mRNA stability. Finally, noting that all measurements are relative to a reference value and that absolute values are not known, an *abundance ratio* was defined which describes the relative scaling of

mRNA levels and transcription rate measurements. An mRNA degradation rate was estimated for each timestep, for a range of possible abundance ratio values. While three of the genes support a gradual destabilization for a subset of possible abundance ratio values, in all cases the mRNA degradation behaviour was found to depend on the unknown abundance ratio.

Molina-Navarro and co-workers [92] performed a similar analysis of microarray-based transcription rate (nuclear run-on) and mRNA abundance in response to oxidative stress in *S. cerevisiae*. Unknown absolute values were not accounted for, however, as the analysis was based on externally normalized values which were assumed to represent absolute values. A decay rate constant k was calculated for each time interval by assuming a first-order decay process (Eqn 3.3) within each time interval, and solving for the decay rate constant k_i within each time interval $\Delta t_i = t_{i+1} - t_i$ ([92]; summarized here based on Eqn VIII in [89]):

$$p - k_i \left(R_{i+1} - k_i E_{i+1}\right) = \left[p - k_i \left(R_i - k_i E_i\right)\right] e^{-k_i \Delta t_i} \tag{3.6}$$

where $p \equiv (R_{i+1} - R_i)/\Delta t$ is the observed change in transcription rate over the time interval, and E_i, R_i the mRNA level and transcription rate measurements at timepoint t_i. Most genes were found to have non-constant k over the timecourse. Estimates of k_i are noisy due to the use of sequential timepoints to estimate the local change in transcription rate. The solution permits and often finds negative values of k, which is not biologically plausible when interpreted as an mRNA decay rate constant. The authors find coherent transcription rate, mRNA level and resulting $k(t)$ behaviours amongst functionally related genes, indicating that there are distinct patterns of transcriptional and possibly post-transcriptional control in response to oxidative stress.

Together these studies indicate that the mRNA stability of some genes is regulated in response to stress, and that this is true across a range of organisms and stress conditions. There is also evidence that coherent patterns of transcriptional response and mRNA abundance are associated with protein functional classes [92, 90, 104],

suggesting that the gene expression response to stress is specifically regulated and associated with protein function. Where the transcription rate and mRNA abundance both vary in response to stress, the interpretation of transcriptional and post-transcriptional contributions to mRNA abundance is complicated by unknown absolute values.

Theoretical strategies for regulating mRNA levels by combining the regulation of transcription rate and mRNA stability have been described previously, focusing on instantaneous changes in transcription rate and mRNA stability from an initial steady state [94, 91, 98, 89]. Perez-Ortin [89] showed that the speed of approach and relative magnitude of a new steady state in mRNA abundance can be tuned by instantaneously modulating both the transcription rate and mRNA stability from an initial steady state. Shalem and colleagues [98] observed that transient upregulation of mRNA levels is consistent with a rapid transcriptional response (and accompanying rapid approach to a peak in mRNA abundance) followed by mRNA destabilization, allowing rapid return to the initial mRNA level. Both cases assume a simple exponential approach behaviour in mRNA levels which results from instantaneous changes in transcription rate and/or mRNA stability from an initial steady state and assuming first-order mRNA degradation. Solving Eqn. 3.4 for simultaneous instantaneous changes in both stability ($k_I \rightarrow k_F$) and transcription rate ($R_I \rightarrow R_F$) from initial steady-state at time $t = 0$, the mRNA abundance $E(t)$ follows:

$$E(t) = \frac{R_F}{k_F} + \left(\frac{R_I}{k_I} - \frac{R_F}{k_F} \right) e^{-k_F t} \tag{3.7}$$

More generally, both transcription rate and mRNA abundance change over time. Previous studies have compared observed mRNA levels to simulated timecourses for a given mRNA halflife and have estimated changes in mRNA degradation for consecutive time intervals. However, there has not been a systematic attempt to fit analytic solutions for changes mRNA abundance to an observed timecourse, given a model of mRNA degradation and observed changes in transcription rate, and explicitly modelling the unknown absolute values.

The rest of this chapter presents a model-based approach to address the following questions: given a timecourse of genome-wide changes in mRNA abundance and changes in transcription rate, are the data consistent with constant mRNA stability? Alternatively, is there evidence for post-transcriptional control of mRNA abundance? In particular, the examples considered here are relevant to transient and persistent stress response behaviours.

3.3 Fitting an mRNA kinetic model to microarray data

3.3.1 First-order mRNA degradation

I assume here that a time course of transcription rate and mRNA abundance is measured. Observed values are fold-changes in transcription rate and mRNA abundance relative to a reference value. There is therefore an unknown scaling between measurements of relative transcription rate and relative mRNA abundance, and there may also be a baseline shift in measured changes in transcription rate or mRNA abundance. I therefore assume that, in the absence of noise, the normalized ratio measurements from the arrays, $y(t)$, $f(t)$, are related to absolute mRNA abundance $E(t)$ and transcription rate $R(t)$ as follows:

$$\text{mRNA abundance} \quad y(t) = \alpha_y E(t) + \beta_y \tag{3.8}$$

$$\text{transcription rate} \quad f(t) = \alpha_f R(t) + \beta_f \tag{3.9}$$

where the scaling (α) and shifting (β) constants are allowed to differ for each gene. Substituting into Eqn 3.3, the kinetic equation relating mRNA abundance measurements $y(t)$ and transcription rate measurements $f(t)$ given first-order decay becomes:

$$\frac{dy}{dt} + ky(t) = A'f(t) + B' \tag{3.10}$$

where constants A', B' may differ for each gene. The solution for mRNA abundance

under first-order degradation with decay rate constant k becomes:

$$y(t) = A \int f(t')e^{k(t'-t)}dt' + B + Ce^{-kt} \tag{3.11}$$

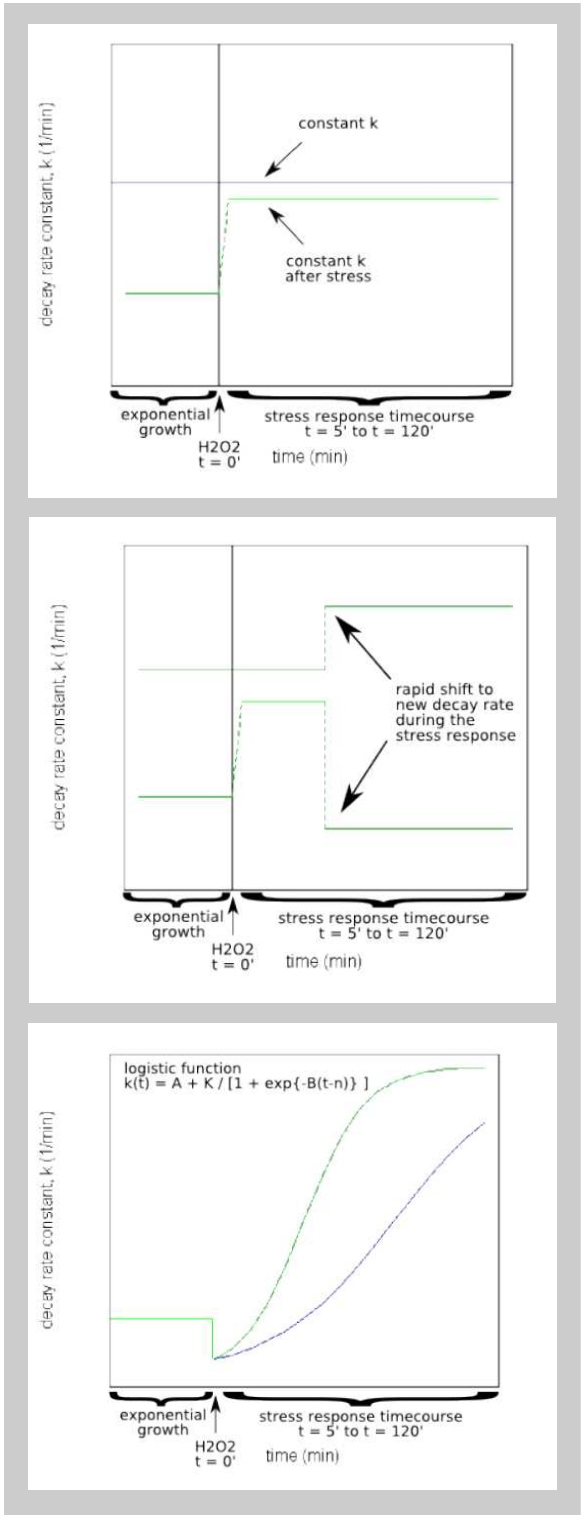
where A, B, C are gene-specific constants. Figure 3.1A illustrates a first-order model of mRNA degradation.

3.3.2 First-order mRNA degradation: interpretation of parameters

The inclusion of constants A', B' in the first-order decay model (Eqn. 3.10) reflects a limitation of using microarray measurements to investigate mRNA kinetics and therefore working with relative, not absolute, changes in transcription rate and mRNA abundance. (This is also the case for low-throughput assays [91].) An expression profile may be explained by more than one set of parameters A, B, C (Eqn. 3.11) so that it is impossible to distinguish between modes of response, whether transcriptional, post-transcriptional, or a combination. To illustrate, consider the simple case that both the mRNA abundance and the transcription rate follow an approximately exponential decay. This may be explained by either (i) a dominating transcription rate combined with a constant short halflife (large A, large k)[1], or (ii) a dominating decay rate (small A).

If the mRNA abundance response follows an exponential approach behaviour and the transcription rate has any other behaviour, this can always be explained by a dominating first-order decay process (so that A is small). In this example it is not possible to distinguish between (i) degradation dominating the kinetics, so that abundance is unaffected by the relatively small absolute changes in transcription rate; and (ii) transcription rate (*i.e.* the production of new transcripts) contributing to changes in RNA abundance over time. By including unknown constants A, B, C in the model solution (Eqn. 3.4) and searching over those parameters for good fits to the observed mRNA abundance profile, we may search over all possible unknown relative scalings A' and relative shifts B'. Where there are a wide range of parameter

[1]Note that *large/small* is determined by the scaling of the governing equation (Eqn. 3.11).



Exponential (first-order) decay, with rate constant k throughout the stress response.

Exponential (first-order) decay in two regimes, with rate constants k_1 ($t < t_{switch}$) and k_2 ($t > t_{switch}$) during the stress response.

Exponential (first-order) decay in two regimes, with rate constants k_1 ($t < t_1$) and k_2 ($t > t_2$) separated by a period of adjustment (e.g. caused by an asynchronous cell population).

A period of adjustment between two decay rate constants can be modelled as, for example, a linear shift (**a**) or a logistic function (**b**).

Figure 3.1: Models of mRNA degradation $D(t)$. (**A**) Exponential decay; (B) Exponential decay in two regimes. Both (**A**) and (**B**) are compatible with a rapid change in decay rate at the start of the stress response. Alternative models of mRNA degradation may be considered subject to the constraints of the dataset (time resolution and noise) and parameter identifiability; for example (**C**), $D(t) = k(t)E(t)$ where $k(t)$ follows a generalized logistic model.

sets which can describe a given mRNA abundance profile, the observed timecourse is not sufficient to specify the model precisely; however, it is possible to test an alternative model against a model of constant first-order degradation by comparing a measure of the global best fit between models. This is discussed in the following sections.

3.3.3 Alternative models of mRNA degradation

mRNA stability may be regulated during a stress response. To describe changes in mRNA degradation rate, an alternative model of mRNA degradation may be fitted to the stress response timecourse: an alternative model may capture key aspects of the mRNA degradation response that are missed by a first-order degradation model. To illustrate, I describe here two alternative models of mRNA degradation during a stress response.

Firstly, a straightforward extension of first-order degradation is to allow a single instantaneous change in mRNA degradation at a (unknown) time t_{switch} after stress induction (Figure 3.1B). This model is consistent with observations that mRNA degradation may be modulated at discrete times after stress induction and may therefore be involved in regulating a transient response [98, 99].

Solving $\frac{dE}{dt} = R(t) - D(t)$ with

$$D(t) = \begin{cases} k_1 E(t); & t \leq t_{switch} \\ k_2 E(t); & t > t_{switch}, \end{cases}$$

the analytic solution for a single change in degradation rate at time t_{switch} is:

$$y = A \int f(t') e^{k_1(t'-t)} dt' + B + De^{-k_1 t} \quad \text{for t} < \text{t}_{\text{switch}}, \tag{3.12}$$

$$y = A \int f(t') e^{k_2(t'-t)} dt' + B\frac{k_1}{k_2} + Ge^{-k_2 t} \quad \text{for t} \geq \text{t}_{\text{switch}}, \tag{3.13}$$

where G is chosen to satisfy continuity[2] of $y(t)$ at t_{switch} ($\equiv t_s$ for brevity):

$$G = e^{k_2 t_s}\left(A\int^{t_s} f(t')e^{k_1(t'-t_s)}dt' + B + De^{k_1 t_s} - A\int_{t_s} f(t')e^{k_2(t'-t_s)}dt' - B\frac{k_1}{k_2}\right) \tag{3.14}$$

The solution for an abrupt change in stability has six parameters ($k_1, k_2, A, B, D, t_{switch}$), compared to the four parameters of the constant decay rate model (k, A, B, C in Eqn 3.11).

Secondly, consider the case that there is an increase or decrease in mRNA degradation rate but that this is not instantaneous. This model is a natural extension of an instantaneous change in mRNA abundance, where the change in mRNA stability is spread over time either due to the mechanism of mRNA stability regulation for a mRNA species, or as a consequence of rapidly modulated mRNA stability within a cell but observed in a cell population with an asynchronous population response. Continuous growth (which may be localized in time and may be increasing of decreasing) may be modelled, for example, using a logistic growth function [107]. The mRNA degradation rate $D(t)$ may be modelled as $D(t) \equiv k(t)E(t)$ where $k(t)$ follows a logistic growth behaviour (Figure 3.1C):

$$k(t) = \alpha + \frac{\gamma}{\left(1 + e^{\beta(\nu - t)}\right)} \tag{3.15}$$

Solving $\frac{dy}{dt} + k(t)E(t) = A'f(t) + B'$ [*cf.* Eqn. 3.10] for mRNA abundance $y(t)$:

$$y(t) = e^{-G(t)}\left[D + \int_{t_0}^{t} e^{G(t')}\left(Af(t') + B\right)dt'\right] \tag{3.16}$$

where

$$G(t) = (\alpha + \gamma)t + \frac{\gamma}{\beta}\left[ln\left(1 + e^{\beta(\nu - t)}\right)\right], \tag{3.17}$$

and there are up to seven parameters: constants $\alpha, \beta, \gamma, \nu$ describe the mRNA degradation rate model ($k(t)$) and constants A, B, D are a consequence of unknown absolute values.

[2]Continuity at t_{switch} may not always be an appropriate constraint; for example, if the time resolution of the timecourse is low compared to the frequency of fluctuations in transcription rate close to time t_{switch}. In the absence of a continuity constraint, the two time regimes may be fitted separately resulting in a 7-parameter model.

3.3.4 Comparison of mRNA degradation models

A standard method to compare the significance of support for two nested models for a given dataset is to use the likelihood ratio test: twice the difference between log-likelihood for the two models is tested against a χ^2_k distribution, where k is the difference between the degrees of freedom in the two models [108]. For non-nested models, or models which are nested but with a parameter in the nested model fixed at a boundary value of the containing model, the asymptotic distribution of the likelihood ratio will not in general be χ^2_k (*e.g.* see [109] for discussion of significance for specified nested models with parameters of the nested model fixed on the boundary of the containing model). A straightforward alternative method for comparison of support for any two models with different numbers of parameters (*i.e.* different model complexity) is to compare an adjusted-R^2 statistic, such that the goodness-of-fit of a model is penalized for model complexity. In particular, an adjusted-R^2 statistic is used in the subsequent chapter to compare two models, M_1: constant mRNA decay rate (Eqn. 3.4), and M_2: an instantaneous change in mRNA decay rate (Eqn. 3.14) during the timecourse; in this case, model M_1 is nested in model M_2 with $t_{switch} \in \{t_0, t_{max}\}$.

3.4 Discussion

This chapter has described previous approaches to modelling mRNA abundance as a function of transcription rate and mRNA degradation, and has presented a simple model-based, analytical approach towards understanding the contribution of transcriptional and post-transcriptional control of mRNA levels. As the number of timepoints and the time resolution feasible in microarray experiments increases, the application and limitations of a model-based approach has increasing relevance for experimental design. The range of possible values of absolute mRNA abundance and transcription rate – which are unknown in microarray studies – can affect the interpretation of measured changes in mRNA abundance and transcription rate.

The unknown scaling between mRNA abundance and transcription rate should therefore not be ignored when modelling mRNA abundance and turnover using microarray measurements. Additional (nuisance) parameters are introduced which are necessary to account for the unknown scaling of relative transcription rates and relative abundance resulting from the use of microarray measurements.

I have presented a model-based approach to address whether observed changes in mRNA abundance can be explained by the observed transcription rate profile assuming a constant first-order mRNA degradation model, or assuming alternative models of mRNA degradation. It is clear that some classes of mRNA abundance response are consistent with constant first-order degradation under many possible parameter values, so that precise estimates of kinetic parameters cannot generally be obtained using the approach described here. Nevertheless, some classes of mRNA abundance response may be better explained by an alternative model of mRNA degradation. In the general case of comparing non-nested models, an adjusted-R^2 statistic is a simple and convenient criterion for identifying genes which are better explained by an alternative degradation model and can be used to compare global optimal goodness-of-fit for each considered model.

Cao and Parker [93] built a multicompartment model of the transcription and decay of a single mRNA species, incorporating 21 kinetic parameters which are estimated from measurements or adjusted to fit observations. While this approach can be used to explain an observed timecourse of mRNA levels, kinetic data for intermediate steps in mRNA metabolism is not yet available on a genome-wide scale.

Chapter 4

Regulation of mRNA stability in response to oxidative stress in *Schizosaccharomyces pombe*

Current approaches to modelling gene regulation networks have focused on regulated transcriptional control by DNA-binding transcription factors. However, there are additional layers of transcriptional and post-transcriptional regulation that affect gene expression As discussed in Chapter 3, there is a growing body of evidence that the regulation of mRNA stability contributes to the dynamics of the gene expression response. The rate of loss or accumulation of mRNA changes in response to environmental stress.

In this chapter, I investigate the contribution of transcription rate and mRNA degradation to mRNA abundance in response to oxidative stress in the fission yeast *Schizosaccharomyces pombe*. A kinetic model of mRNA abundance is defined and fitted to simultaneous microarray timecourses measuring changes in transcription rate and mRNA abundance. Genes are identified for which the observed mRNA abundance response is better explained by a stabilization or destabilization event during the stress response than by a constant mRNA degradation rate. Candidate genes are identified for regulated mRNA stability at two stages in the early oxidative stress response: a rapid change in mRNA stability at the onset of stress, and a delayed change in mRNA stability some time after the induction of a stress response.

4.1 Introduction

4.1.1 Oxidative stress response to hydrogen peroxide

This section introduces some of the transcriptional and post-transcriptional processes involved in regulating gene expression in response to environmental stress. In *S. pombe*, the magnitude of mRNA induction or repression, duration of the stress response, and the contribution of distinct stress-response pathways are finely tuned to specific stressors, and to stress severity or dosage [110]. Yeast cells must balance growth and proliferation with protection from environmental perturbations. The response to an abrupt environmental stress includes repressing genes related to growth and inducing stress-related genes. Most stress-induced and stress-repressed genes eventually return to a steady-state in which mRNA abundance is close to that of the unstressed cell, even in persistent stress conditions [74]. Bähler and co-workers defined a core environmental stress response (CESR) involving hundreds of genes in *S. pombe* [111]. The CESR is common to many stress conditions and is conserved in *S. cerevisiae*. Stress-induced genes in the CESR include genes involved in carbohydrate metabolism and energy generation whereas stress-repressed genes in the CESR are related to growth, including ribosome biogenesis and translation. Lopez-Maury and co-authors [112] have noted that hundreds of genes are induced in response to many specific stressors but few of these genes appear to have specific functional relevance to the stress; a widespread general stress response may protect the cell from several potentially co-occuring environmental stresses or from a perturbation which has not previously been encountered by the organism, or may be the result of evolutionary drift or widespread transcriptional regulation by a small number of key regulators.

Stress responses in *S. pombe* are primarily regulated at the transcriptional level, and the genome-wide stress response is most completely understood at the level of changes in mRNA abundance (*e.g.* [113, 111, 110, 114]). The transcription factors Pap1, Prr1, Hsr1 and Atf1, and the mitogen-activated protein kinase Sty1, have

been shown to be involved in the regulation of subsets of genes which are induced or repressed at the level of mRNA abundance in response to oxidative stress.

Post-transcriptional processes involved in controlling the abundance of growth-related and stress-related proteins during a stress response include the regulation of translation, subcellular localization of transcripts, and mRNA stability. Translation of growth-related transcipts and most other transcripts is inhibited in reponse to stress *via* phosphorylation of the eukaryotic translation initiation factor eIF2 [115, 116]. A subset of stress-induced transcripts in *S. cerevisiae*, including the transcriptional activator Gcn4, escape translational repression. Protection from translational repression has been associated with the presence of translation regulatory regions in the 5' UTR [116]. Subcellular localization may also play a role in the rapid control of mRNA degradation and translation in response to stress. In *S. cerevisiae* and other eukaryotes, transcripts can rapidly be stored in P-body protein assemblies. Transcripts which are not required for translation may be degraded as part of a P-body assembly or may be stored and later released for translation [117]. For example, translational repression has been associated with transcript stabilization in response to stresses affecting the endoplasmic reticulum in human cell lines [118], suggesting a complex interplay between post-transcriptional processes in response to stress.

An altered mRNA stability in response to stress has been reported for individual genes. The mRNA of *S. pombe* transcription factor Atf1, a transcriptional regulator of the core environmental stress response, has been shown to stabilize in oxidative stress conditions, contributing to the rapid accumulation of Atf1 transcripts [119, 120]. Recent studies suggest that the control of mRNA stability on a genome-wide scale contributes to the regulation of gene expression in response to stress, as discussed in Chapter 3.

Post-transcriptional processing of eukaryotic mRNA, including nuclear export, translation, cytoplasmic location, and mRNA stability, is mediated by RNA-binding proteins [121] and in higher eukaryotes by the binding of small regulatory RNAs. Tar-

get specificity of RNA-binding proteins is determined by mRNA sequence and secondary structure. Small non-coding RNAs are thought to bind to mRNA by complementary or near-complementary base pairing to the target region, although the structure of target mRNA also affects miRNA binding affinity [122]. In higher eukaryotes, siRNAs and miRNAs play a role in the active degradation of mRNA transcripts by binding to the target transcript [27]. While the siRNA pathway is active in *S. pombe* (reviewed in [123]), the existence of miRNAs in *S. pombe* is unproven and it is not known whether miRNAs are involved in regulating mRNA stability in *S. pombe*. It is not yet clear whether the absence of a Drosha homologue in *S. pombe* is indicative of the absence of an miRNA biogenesis pathway in *S. pombe* [124]. If regulated mRNA stability is mediated by RNA-binding proteins or small non-coding RNAs, targeted mRNA transcripts may share sequence motifs or combined sequence-structure signatures [125].

4.1.2 Study aims

The aim of this study was to identify candidate genes for regulated mRNA stability in response to oxidative stress by hydrogen peroxide, using simultaneous timecourses of changes in transcription rate by RNA polymerase II and changes in mRNA abundance:

Aim. Determine whether there is a group of genes with a change in mRNA stability during an oxidative stress response.

Data. Timecourse of two simultaneous measurements: (i) change in RNA abundance, measured by two-channel cDNA microarrays; and (ii) change in transcription rate, estimated by RNA polymerase II (RNA Pol II) occupancy and measured using RNA Pol II ChIP-chip on the same two-channel array design.

Approach. The kinetics of mRNA abundance were investigated during the first 120 minutes of an oxidative stress response in *S. pombe*. I investigated whether, amongst all genes with a transcriptional or mRNA abundance response to oxidative stress,

there are:

(i) genes consistent with constant mRNA stability throughout the stress response (i.e. the response in mRNA abundance can be explained by observed changes in the transcription rate);

(ii) genes with a change in mRNA stability at some time during the stress response; and/or

(iii) genes with an immediate rapid change in mRNA stability at the point of stress induction.

4.2 Datasets: transcription arrays and expression arrays

In collaboration with Samuel Marguerat and Jürg Bähler in the Fission Yeast Genomics Group at the Wellcome Trust Sanger Institute, we obtained simultaneous timecourse measurements from RNA polymerase II ChIP-chip and mRNA abundance. All wet-lab work was performed by Samuel Marguerat.

4.2.1 Experimental procedure

S. pombe cells growing at exponential phase were subjected to oxidative stress using 0.5mM hydrogen peroxide (H_2O_2). RNA and ChIP samples were harvested immediately before stress induction and 11 further samples were harvested after stress induction. Samples were taken at approximately t = 0, 5, 10, 15, 20, 25, 30, 40, 50, 60, 90, 120 minutes. RNA samples were hybridised on arrays with pooled RNA as a reference. RNA Pol II ChIP samples were hybridised with the respective genomic input DNA as the reference sample.

4.2.2 Data preprocessing

The microarrays (arrays) used in this study were custom two-channel cDNA arrays (see Chapter 1, page 7). Raw intensity datasets from each array were processed using a normalization script designed for use with these arrays. Features are spotted in duplicate or more on the arrays. Each array was normalized separately. As described in Lyne *et al.* (2003) [33], signal variation on these arrays tends not to be related to signal intensity, but tends to show marked spatial effects. The within-array normalization used to preprocess raw intensity data from these arrays is therefore designed to correct for spatial intensity variation. Local normalization of ratio values is carried out within a sliding window on the array. A scaling factor is calculated for each window such that the median channel intensity ratio amongst spots within each sliding window on the array is 1, and used to correct the channel intensity ratio of the central spot. Quality flags are set for each spot on each array; in this study, spots flagged other that "P" (*pass*) were considered to be low quality and were discarded. mRNA abundance and RNA Pol II profiles for each gene were calculated as the mean of ratio values from replicate spots, after discarding spots with missing data (quality flag not "P") on one or more of the arrays. Any features for which fewer than two spot replicates had complete data for all timepoints (quality flag "P" on all arrays) were considered to have incomplete data and were discarded from further analysis. To display relative changes along the timecourse and to perform cluster analysis, each profile (mRNA abundance or RNA Pol II occupancy) was divided by the profile value at the first timepoint ($t = 0$). To fit the mRNA degradation models, all expression and transcription profiles were brought onto a similar scale by dividing each profile value by the mean of that profile. (Note that this rescaling is absorbed by constants A', B' which will be introduced to take account of an unknown scaling between measured changes in mRNA abundance and in transcription rate).

4.2.3 RNA polymerase II occupancy and transcription rate

The timecourse of RNA polymerase II ChIP-chip arrays measures changes in RNA polymerase II occupancy during the first 0 - 120 minutes of the stress response. RNA polymerase II transcribes eukaryotic protein-coding genes. Other polymerase enzymes transcribe ribosomal genes, transfer RNA, and mitochondrial genes. Following the recruitment of RNA polymerase II to the promoter and initiation of transcription, the polymerase traverses the gene resulting in elongation of the nascent transcript. An increase (decrease) in the transcription rate is typically mediated by an increase (decrease) in transcription initiation and a corresponding increase (decrease) in RNA polymerase II density along a transcribed gene. Regulation of the rate of transcript elongation has been observed *in vivo*, including RNA polymerase stalling, transient reversal of the direction of travel, and differential regulation of the rate of transcriptional elongation between exons.

We assumed that the transcription rate is predominantly regulated by transcription initiation. We assumed that measured fold-changes in RNA polymerase II occupancy using a timecourse of RNA polymerase ChIP-chip arrays correspond to fold-changes of a similar magnitude in the rate of transcript production. Under this assumption, the measured fold-changes should be taken as estimates of fold-change in the rate of transcript production by RNA polymerase II ('transcription rate'). The microarray probes had been designed to match to the 3' end of the transcript. Stalled transcription resulting in shortened transcripts is therefore assumed to have negligible impact on detected transcripts[1]. In addition, the effect of stochastic RNA polymerase stalling within an individual cell is assumed to have negligible effect on transcripts taken from a population of cells.

The terms 'RNA polymerase II (RNA Pol II) occupancy' and 'transcription rate' are used interchangably throughout the rest of this chapter.

[1] mRNA is transcribed starting with 5' end of the nascent transcript

4.3 Methods

4.3.1 Overview of methods

This study models the observed mRNA abundance over time assuming specified models of mRNA degradation. Each model was fitted to observed changes in transcription rate and mRNA abundance, independently for each gene with complete data on the array. Genes were assigned to one of the specified models using a goodness-of-fit criterion. The genes within each model group were then clustered by mRNA abundance profile or by concatenated transcription rate and mRNA abundance profiles. Cluster analysis identified groups of genes with similar transcription rate and mRNA abundance behaviours within some of the model groups, and was used in place of a fold-change cutoff to identify groups of genes with high amplitude responses which are candidates for response to stress in mRNA abundance and/or transcription rate. Gene clusters were analysed for Gene Ontology term enrichment (see page 13), and putative stabilized and destabilized gene clusters were tested for overrepresentation of sequence motifs which may indicate specific binding of mRNA by RNA-binding proteins or small non-coding RNAs.

4.3.2 Modelling mRNA degradation during the stress response

Using a timecourse of transcription rate and mRNA abundance it is possible to investigate whether the dataset supports a model of first-order mRNA degradation at a constant decay rate throughout the timecourse, and whether there are groups of genes which are better explained by an alternative model of mRNA degradation (for details see Chapter 3, Section 3). First, it is necessary to define the models of mRNA degradation to be compared. Two phases of mRNA stability regulation are of interest: (i) Is there a rapid change in the stability of some mRNA species, either stabilization or destabilization of mRNA, at the start of the stress response? (ii) Is mRNA stability regulated at a later time in the stress response, as the cell adapts to

oxidative stress?

The gene expression response generally peaks in mRNA abundance towards the end of the timecourse and the cells do not recover to a pre-stress state in either transcription rate or mRNA abundance (Figure 4.1). Genes which are highly induced or repressed during this timecourse show a marked gain/loss of mRNA within 60 mins and maintain an elevated or reduced level of abundance up to 120 minutes. There is no evidence that the cells enter a final recovery phase in the 120 minutes of this timecourse, during which we would expect to see a decrease in the mRNA abundance of all highly-induced stress response genes. Figure 4.1 also indicates that there are several groups of genes which peak in mRNA accumulation or mRNA loss at different times, indicating that this dataset captures several gene expression behaviours during the early stress response. I therefore focused on identifying the strongest candidates for regulated mRNA stability in these early stages of the stress response, before the cells enter a final recovery phase which would return the cells to a pre-stress state.

In Chapter 3, three examples of mRNA degradation models were presented (Figure 3.1, page 50). In the present study, degradation of mRNA during the stress response was considered under two of those models:

(i) exponential decay, characterized by a steady decay rate constant, k, throughout the stress response (Figure 3.1A); and

(ii) piecewise exponential decay, characterized by a single instantaneous change in the decay rate constant ($k_1 \rightarrow k_2$) at a specific (unknown) time after stress induction (Figure 3.1B).

Both models are consistent with an additional rapid change in mRNA stability at the point of stress induction (or within the first few minutes of the stress response, due to the time resolution of the dataset).

For each gene, we would tentatively select the model of a change in degradation rate

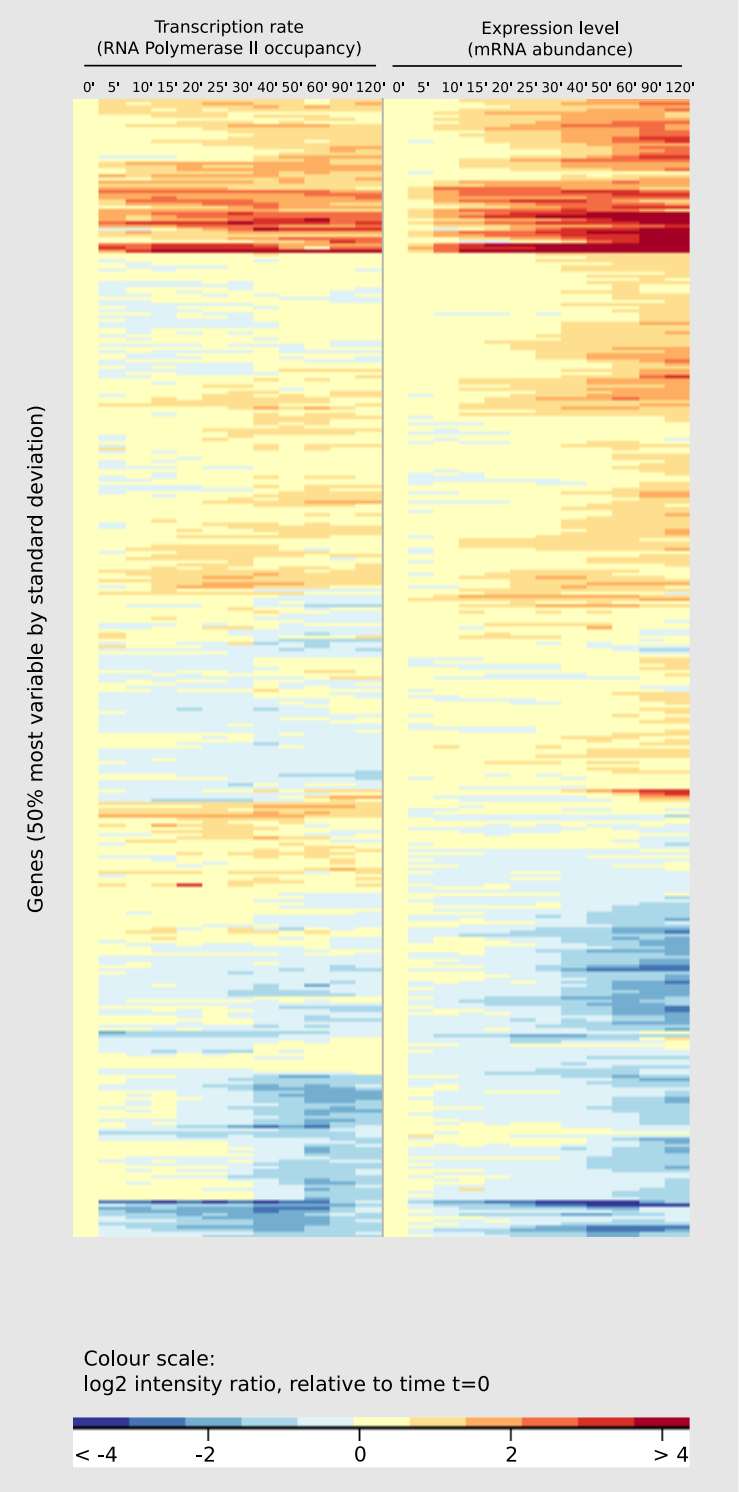


Figure 4.1: Heatmap showing the relative changes in RNA polymerase II occupancy and mRNA expression levels compared to the first timepoint for the 50% most variable genes. The heatmap shows that the cells do not recover to pre-stress mRNA levels during this timecourse: genes highly induced (repressed) in response to the stress tend to maintain a high (low) level of mRNA abundance at 120 mins. Shown are the 50% of genes with the highest standard deviation across transcription and expression profiles (1940 genes amongst the 3881 features with complete data). Plotted values are $\log_2$ channel intensity ratios relative to time t=0 (immediately before stress induction).

if adjusted-R^2 is larger for that model than for first-order degradation at a constant rate throughout the timecourse. Additional criteria were imposed, however, in order to identify only a conservative set of genes as candidates for regulated stability during the stress response: the detected change in degradation rate must be large enough; the time of the change in stability must be supported by datapoints, so must not be at the extremes of the timecourse; and genes which are already explained by a good fit to a constant degradation rate are not considered for improvement by a change in stability. Therefore a change in the first-order degradation rate was selected over the constant first-order degradation model only if:

- $adjR^2$ is larger for a change in degradation rate than for the constant decay rate;
- there is a good fit to the change in degradation rate ($adjR^2 > 0.6$);
- $adjR^2 < 0.9$ for the constant degradation rate, so that a good fit to unregulated stability is always selected;
- the change in degradation rate at time t_{switch}, from $k_{initial}$ to k_{final}, is greater than a threshold value;
- t_{switch}, the time of the change from $k_{initial}$ to k_{final}, is between 12 mins and 60 mins so that there are at least three measured timepoints in each regime.

Thresholds were chosen for the smallest permissible detected change in degradation rates ($\frac{k_{initial}}{k_{final}} < 1.4$ is discarded), and for defining a good fit to a constant degradation rate, above which we do not allow the fit to be improved by a change of stability ($adjR^2 > 0.9$). These thresholds were chosen following visual inspection of all model fits. Examples of model fits for two genes are shown in Figures 4.2 and 4.3.

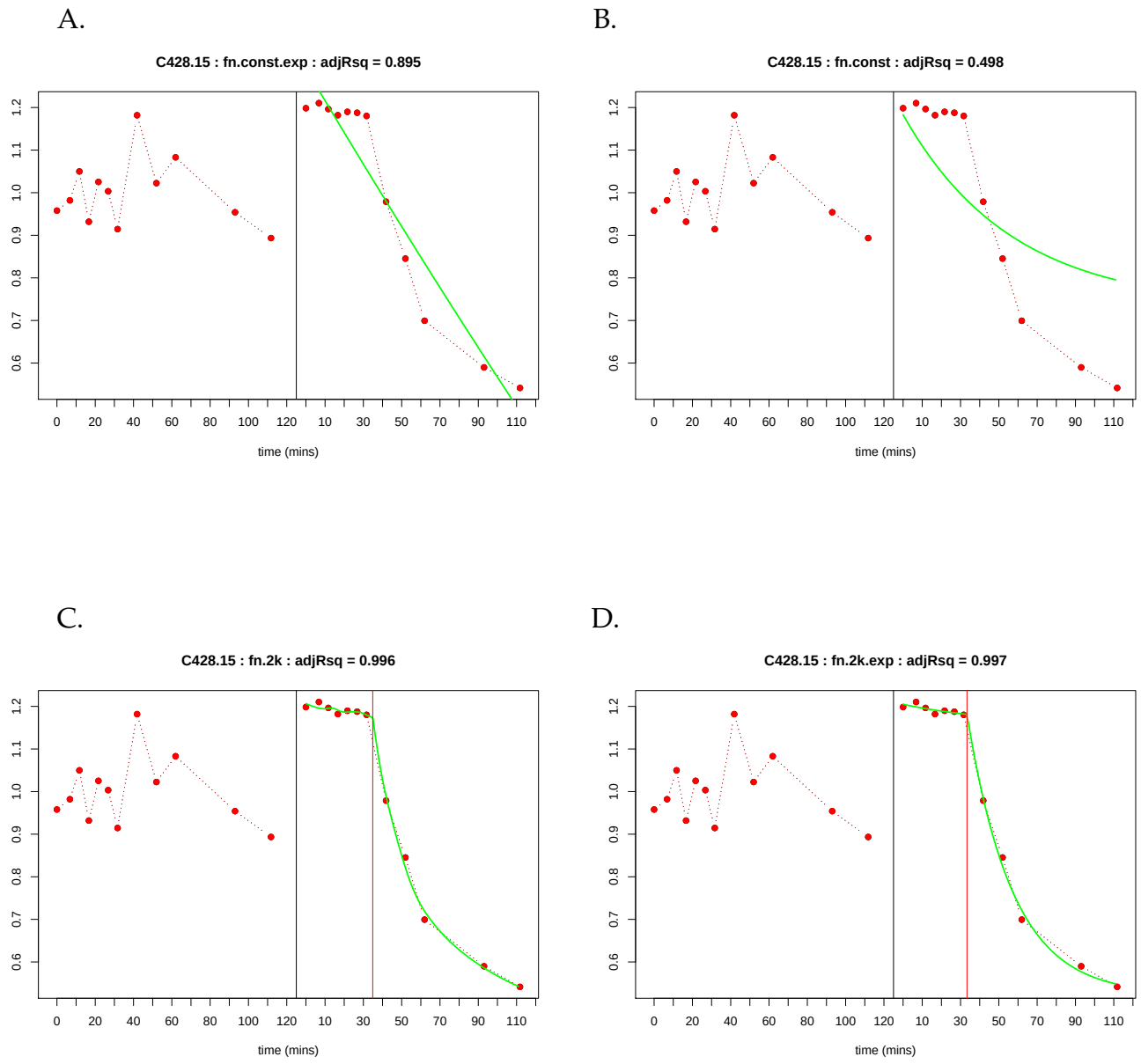


Figure 4.2: mRNA abundance for gene SPC428.15 is better explained by step-function decay (adjR^2 = 0.997) (bottom) than by a constant decay rate (adjR^2 = 0.89) (top). Transcription rate is shown on the left, mRNA abundance on the right of each plot. Fitted expression profiles y_i are shown in green. For step-function decay, t_{switch} is shown as vertical red line. **A.** constant decay with $A = 0$; **B.** constant decay with $A \geq 0$; **C.** step-function decay with $A = 0$; **D.** step-function decay with $A \geq 0$. adjR^2 = 0.895, 0.498, 0.996, 0.997, respectively.

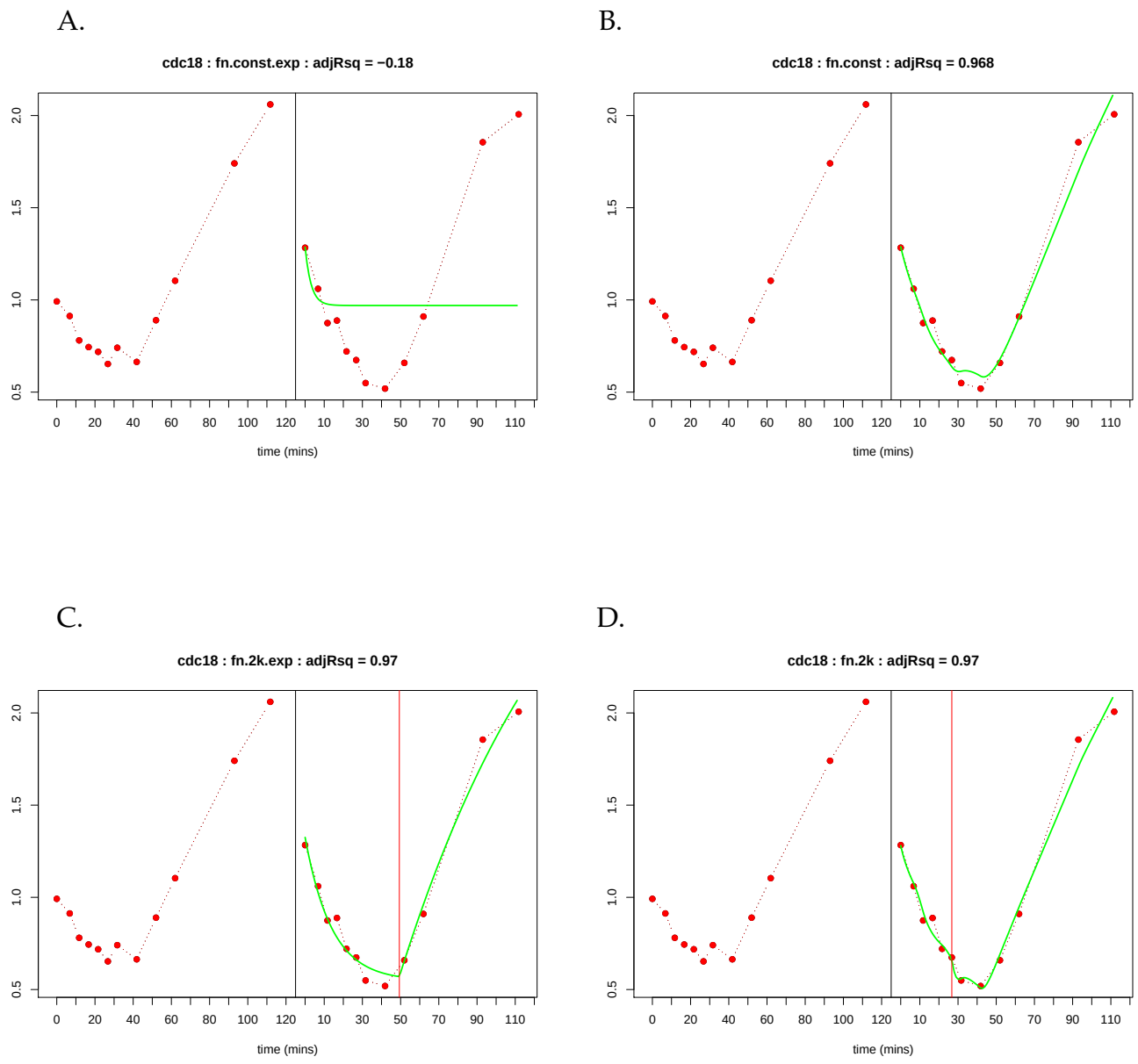


Figure 4.3: Constant decay rate is a good fit for gene cdc18 (adjR^2 = 0.97) so the step-function model (adjR^2 = 0.97) is not selected over the constant decay model. **A.** constant decay with $A = 0$; **B.** constant decay with $A \geq 0$; **C.** step-function decay with $A = 0$; **D.** step-function decay with $A \geq 0$.

4.3.3 Model fitting

For a given model of decay rate $k(t)$, the model solution for $y(t)$ was fitted to the 12 measured values of $y(t), f(t)$, independently for each gene. Parameters were sought to minimise $\sum_{i=t_1}^{t_{12}} (y_i - \hat{y}_i)^2$, where y_i, $\hat{y}_i$ are observed and fitted values of $y(t)$, respectively. Integration was performed with an adaptive quadrature method implemented in the R function `integrate` [78, 126] using linear interpolation of $f(t)$ implemented in the R function `approx` [78]. Minimization was performed using Powell's UObyQA optimization method implemented in the R package `powell` [78, 127]. The maximum number of iterations for each model fit was set to `maxit`= 10000 [127].

Goodness-of-fit

For each model fit, the goodness-of-fit was assessed using an adjusted R^2 statistic which penalizes models with more parameters:

$$\text{adjusted } R^2 = 1 - \left(1 - R^2\right) \frac{(n-1)}{(n-p-1)} \tag{4.1}$$

where

$$R^2 = 1 - \frac{SS_{err}}{SS_{total}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}; \tag{4.2}$$

$n = 12$, the number of timepoints; and $p + 1 = \#params$, the number of parameters in the model.

Initial values

The initial values used to fit the constant first-order decay model are $\{ A = 1, B = 1, C = 1, k = \frac{1}{30} min^{-1} \}$. Randomly selected initial values were also tested for a subset of array features and resulted in similar reported optimal values, suggesting that this optimization procedure together with the stated initial values are likely to find global optima for this dataset.

The initial values used to fit the piecewise constant decay model are $\{ A = 1, B = 1,$

$k_1 = \frac{1}{30}min^{-1}$, $k_2 = \frac{1}{30}min^{-1}$, $D = 1$ }. Tests with randomly selected initial values for t_{switch} showed that the reported optimal values were sensitive to the initial value of t_{switch}, indicating that the optimization procedure terminates in local optima. Values for the parameter t_{switch} were therefore searched by setting t_{switch} to a series of initial values at 5 min intervals ($t = 0$, 5, ..., 120mins) and selecting the optimum over all initial conditions.

Parameter constraints

Parameters in all models were constrained as follows:

- $A \geq 0$: the transcription rate term must be non-negative
- $k, k_1, k_2 \geq 0$: the decay rate constant must be non-negative
- B, C, D unconstrained

Exact sampling times were used to fit the models, with timepoints $t = 0$, 6.78, 11.72, 16.67, 21.67, 26.75, 31.67, 41.85, 51.95, 61.95, 93.00, 111.78 mins corresponding to the time after H_2O_2 treatment at which each sample extract was frozen.

Fitting exponential solutions of mRNA abundance

Exponential approach or decay ($E(t) = a + be^{-kt}$) is a possible solution for mRNA abundance regardless of the transcription rate profile. Therefore, both the constant decay rate model and the piecewise constant decay rate model were fitted for two cases: (i) with $A \geq 0$, and (ii) with $A = 0$. This ensured that where an expression profile could be explained solely by exponential approach to a new steady state (or piecewise exponential approach in the case of the piecewise constant decay model), this fit was found explicitly. An adjusted-R^2 value was calculated to measure goodness-of-fit for both the $A = 0$ model fit and the $A \geq 0$ model fit (using the appropriate degrees of freedom). The best fit for each of the constant decay model and the piecewise constant decay model was chosen as the optimal fit out of the $A = 0$ and $A \geq 0$ model fits. When performing model selection between the

constant decay rate model and the piecewise constant decay rate model, if the piecewise constant model was rejected based on the optimal ($A = 0$ or $A \geq 0$) fit due to the additional criteria for selection, then the other ($A \geq 0$ or $A = 0$) fit was tested. The model of piecewise constant decay rate was rejected only if both the ($A = 0$) and ($A \geq 0$) model fits were rejected.

4.3.4 Array features with complete data and model fits

Array features were reported using custom gene IDs. Custom gene IDs were mapped to *S. pombe* systematic gene IDs using a custom mapping provided by Samuel Marguerat[2].

Of the 5434 features represented on the array, 3881 features had complete data (quality flag "P") for at least two spot replicates at all timepoints, in both the expression and RNA polymerase II ChIP-chip timecourses. The degradation models were fitted only to the 3881 features with complete data. No other filtering was performed at this stage: the degradation models were fitted for all features with complete data, regardless of the magnitude of the response in either mRNA abundance or RNA Pol II occupancy. For 214 features, the model fitting failed to return an optimal set of parameter values for at least one of the models, using the stated initial conditions and maximum iteration number.

Model fits for all considered models were returned for the remaining 3667 array features using the optimization method and initial values described above. Of these, 340 features reported adjusted $R^2 < 0.6$ for all models and were therefore considered not to be explained by either of the two degradation models based on a parameter search using the stated initial conditions. 3327 array features reported a good fit ($adjR^2 > 0.6$) for at least one of the degradation models, of which 69 features were discarded because the custom probe identifier on the array did not map to a systematic gene ID. The remaining 3258 array features were considered to be genes which

[2]Fission Yeast Genomics Group, Wellcome Trust Sanger Institute / UCL

have good fit ($adjR^2 > 0.6$) to at least one of the models. These 3258 genes were used in the model selection and the cluster analysis.

4.3.5 Clustering timecourse data using Bayesian hierarchical clustering

Clustering of the timecourse data was performed using SplineCluster (Heard *et al.* [128, 129]), a model-based Bayesian agglomerative clustering algorithm suitable for identifying homogeneous clusters of genes within a nonuniformly sampled timecourse. A detailed description is given in reference [129]. Briefly, genes are partitioned into C groups. For gene g in group k ($k \in 1...C$), the timecourse data y_{gt} at time $t \in (t_1, ..., t_T)$ is modelled as

$$y_{gt} = \mathbf{X}_g \mathbf{b}_k + \epsilon_{gt} \; ; \quad Var(\epsilon_{gt}) = \sigma_k^2 \tag{4.3}$$

The vector of coefficients $\mathbf{b}_k$ and the error variance ϵ_k are specific to group C_k. Random errors ϵ_{gt} are assumed to form an independent identically distributed Gaussian sequence. The design matrix $\mathbf{X}_g$ contains the basis function representation of the timecourse dataset. Linear spline basis functions were used, giving a continuous piecewise linear model for $y_g(t)$. Prior precision for the $\mathbf{b}$ coefficients ($1/\sigma_k^2$) was set to 10^{-7} for this dataset.

4.3.6 Sequence searches for short word occurence bias and RNA sequence/structure motifs

RNA-binding proteins mediate diverse post-transcriptional processes in eukaryotic cells, including mRNA degradation involving ARE-binding proteins. In *Arabidopsis*, *Mus musculus* and *Drosophila*, miRNA seed regions ≈4-7 base pairs in length bind to target mRNA with perfect or near-perfect complementary base pairing. Selected gene lists were searched for occurence bias of short words which may indicate targeted binding by small regulatory RNAs, using the software Sylamer [130].

mRNA transcript sequences

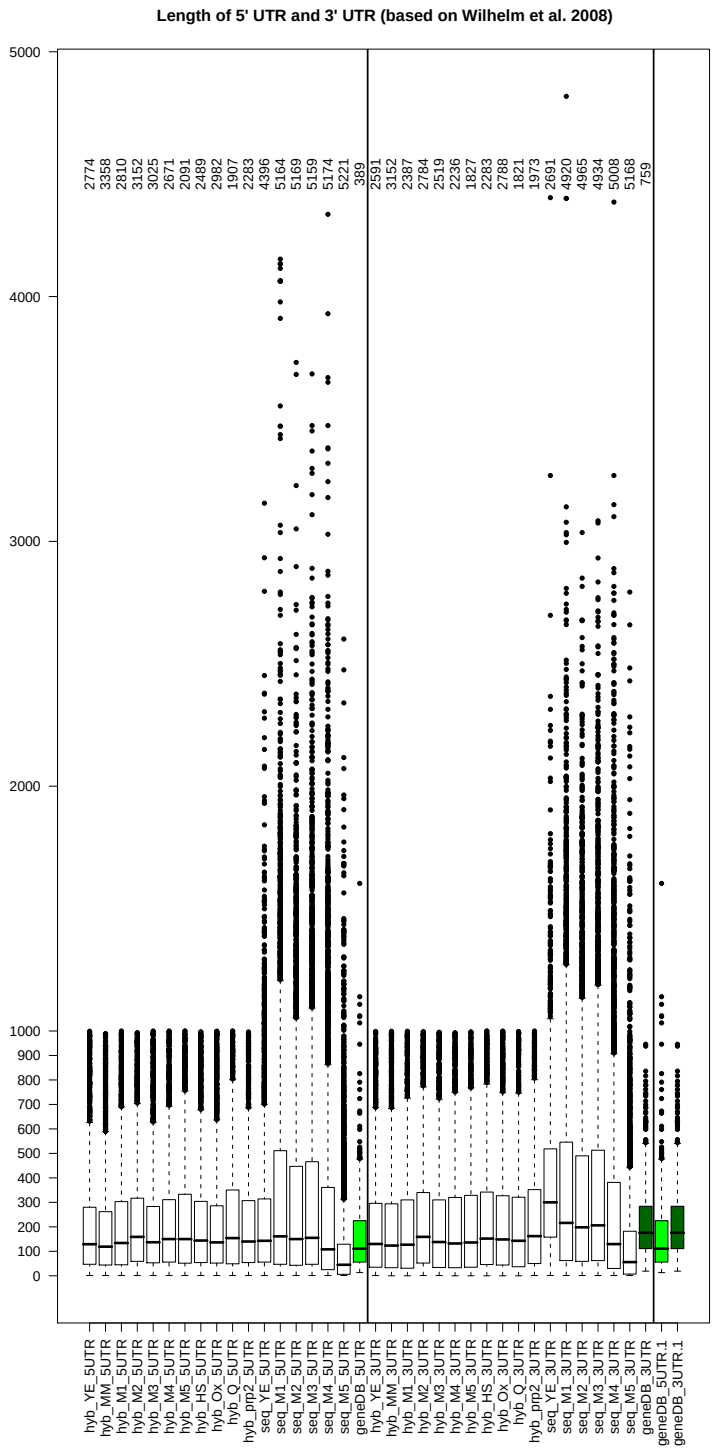
Spliced transcript sequences were obtained for all *S. pombe* genes which have an annotated protein coding sequence (CDS) in GeneDB[3]. The chromosomal location and strand origin of exons were obtained from GeneDB and spliced transcript sequences were reconstructed from the corresponding *S. pombe* genome sequence. Annotated UTR lengths are available for 397 genes (5′ UTR) / 759 genes (3′ UTR) and a recent study reported condition-specific UTR length estimates for the majority of *S. pombe* genes using high-resolution tiling array hybridizations and RNA-Seq cDNA sequencing [50]. Reported UTR length distributions are shown in Figure 4.4. I considered three definitions of UTR length in order to obtain full-length spliced transcript sequences: (i) annotated UTR lengths (GeneDB), (ii) per-gene median of UTR length estimates from all condition-specific hybridization and RNA-Seq measurements, (iii) 500 base pairs 3′ and 5′ of the protein coding sequence.

Short word occurence bias: Sylamer

A short word enrichment tool, Sylamer [130], was used to search for enrichment and depletion of 6-mers amongst spliced transcripts of selected gene lists. Sylamer searches for biases in short word occurence amongst incremental subsets of a gene list compared to a given sequence background. For fixed word lengh k, cumulative raw hypergeometric p-values are reported for each k-mer and displayed as p-value landscapes. Composition bias in the sequence background was corrected using a Markov correction (for specified $m < 6$) to estimate the expected frequency of 6-mers given observed occurences of all m-mers. Sylamer p-value landscapes were generated for all possible 6-mers amongst selected subsets of the *S. pombe* genome (m = 4, incremental step size = 1). The sequence background was taken to be all *S. pombe* genes with a GeneDB annotated protein coding sequence. Selected gene lists were tested for 6-mer occurence bias compared with the three sequence back-

[3] *S. pombe* chromosomal sequences and exon coordinates were downloaded from GeneDB (http://www.genedb.org/genedb/pombe/index.jsp) on 29 April 2009. GeneDB annotated UTR lengths were taken from Wilhelm *et al.* (2008) [50]

grounds corresponding to the three UTR length definitions.



Short name	Brief description
seq	RNA-Seq high throughput cDNA sequencing
hyb	High-density tiling array hybridization
YE / MM	Exponential growth (yeast extract / minimal medium)
M1 / M2 / M3 / M4 / M5	meiosis timepoints
HS / Ox	Stressed cells (heat shock / oxidative stress)
Q	Quiescent cells
prp2	Splicing conditional mutant

Figure 4.4: Boxplots of 5′ and 3′ UTR length distributions in *S. pombe*, detected under different conditions and platforms as reported by Wilhelm et al. (2008) [50]. The number of genes for which UTR annotation is reported is shown above each condition. Shown in green are the length distributions of previously annotated 5′ and 3′ UTRs.

4.4 Results

Table 4.1 summarizes the groups of genes identified as a result of data selection and model selection. The following sections describe the results of model fitting, cluster analysis, and sequence analysis.

4.4.1 Genome-wide fits of first-order mRNA degradation

Firstly, to what extent does a first-order degradation model explain the observed mRNA expression profiles during the first 120 minutes of the stress response? In this model the mRNA stability is assumed to be constant during the stress response timecourse, but note that this does not exclude a rapid change in mRNA stability to a new constant degradation rate at the onset of stress. Figure 4.5 displays fitted R^2 measurements for the first-order mRNA degradation model (Eqn. 3.4; goodness-of-fit to ($A \geq 0$) is shown). For most genes, an assumption of constant decay rate is a good fit to the observed mRNA abundance profile given the observed transcription rate profile. This is not the case for a random assignment of observed transcription rate profiles to observed mRNA abundance profiles (top right panel of Figure 4.5). The same model fits are shown following one random permutation of transcription rate profiles which breaks the association between transcription rate profiles and mRNA abundance profiles. There is a higher density mass in the range $R^2 < 0.6$ in the random permutation of transcription rate-mRNA abundance association than in true association. The density peak close to $R^2 = 1$ in the permuted case includes mRNA abundance profiles which are good fits to purely exponential behaviour, and which are therefore unaffected by permuting the transcription profile assignments.

Randomizing the association between transcription rate profiles and mRNA abundance profiles causes a loss of fit to a first-order mRNA degradation model. This observation confirms that a first-order degradation is a useful initial model for modelling the observed mRNA abundance and transcription rate profiles in this dataset. The first-order degradation model is able to explain mRNA abundance behaviour in

Table 4.1: Tables summarizing the results of data selection (A) and model selection (B). Model selection was used to identify genes with a putative stabilization event, putative destabilization event, and constant mRNA decay rate during the observed timecourse. Cluster analysis was subsequently applied to each group, in order to identify genes with similar mRNA abundance or transcription rate profiles within each group.

A. Data selection:
5434 features represented on the array
1553 features with incomplete data **3881 features** with complete data
214 features with incomplete model fit results **3667 features** with complete model fit results
340 features $adjR^2 < 0.6$ for all models **3327 features** $adjR^2 > 0.6$ for at least one model
69 features do not map to a systematic gene ID **3258 genes** with systematic identifiers ↪ **Model selection and cluster analysis**

B. Model selection and cluster analysis:
3258 genes considered for model selection
873 genes better fit by single instantaneous change in mRNA decay rate than by constant decay rate throughout timecourse
433 genes with putative destabilization event ↪ **Cluster analysis** **440 genes** with putative stabilization event ↪ **Cluster analysis**
2385 genes with good fit to constant mRNA decay, not better fit by single instantaneous change in mRNA decay rate ↪ **Cluster analysis**

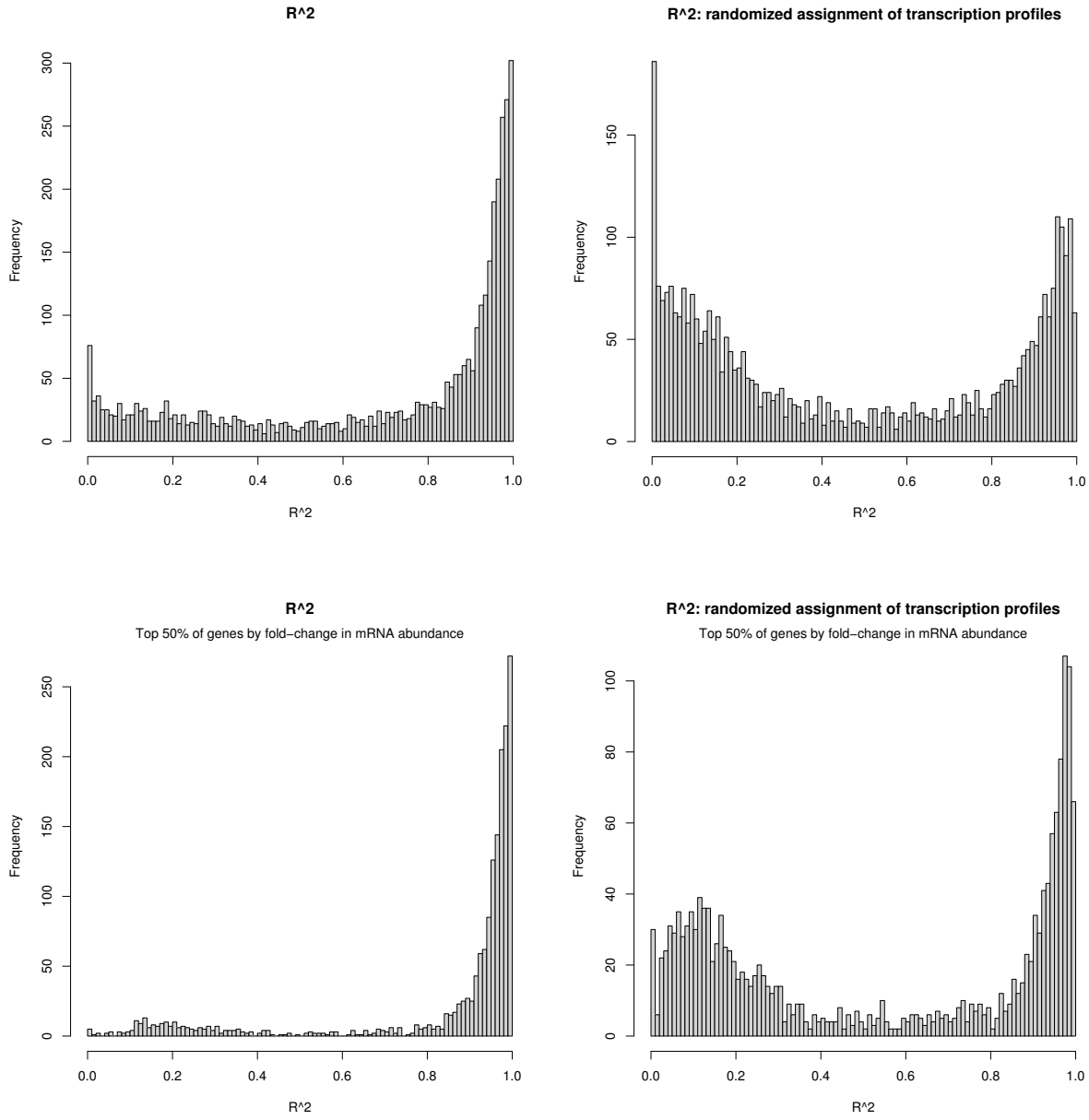


Figure 4.5: Goodness-of-fit (R^2) of a first-order mRNA degradation model (Eqn. 3.4; $A \geq 0$) to observed expression and transcription rate profiles. **Top:** array features with complete data; **bottom:** the 50% of array features with the highest fold change in expression values. **Right hand side:** transcription profiles are randomly assigned to expression profiles. R^2 values which are less than 0 are placed in the first bin.

this dataset that is lost after randomization of transcription rate-mRNA abundance associations.

4.4.2 Evidence of putative regulated mRNA stability 12-60 mins after stress induction

There are a number of genes for which a piecewise constant mRNA stability appears to be a better fit than a constant mRNA stability throughout the response, as shown in Figure 4.6. Each gene was therefore classified according to whether the observed expression profile was better explained by a model allowing an instantaneous change in mRNA stability during the stress response than by a model with a contant decay rate. Each class – constant mRNA stability or piecewise constant mRNA stability – was then examined to identify genes which are highly induced or repressed in response to stress. I also investigated the genes in each class for evidence of an additional change in stability at the start of the stress response.

An observed mRNA abundance profile was considered to be better explained by a change in stability during the timecourse than by constant stability during the timecourse if the following thresholds were satisfied:

- adjusted-R^2 greater for a change in stability than for constant stability,
- adjusted=$R^2 > 0.6$ for a change in stability,
- adjusted-$R^2 < 0.9$ for constant stability
- $\frac{k_{initial}}{k_{final}} > 1.4$ or $\frac{k_{final}}{k_{initial}} > 1.4$
- t_{switch} between 12 mins and 60 mins.

873 genes were classified as being better explained by single change in mRNA decay rate 12-60 mins into the stress response than by a constant decay rate throughout the response. Of these, 433 genes showed a putative destabilization ($k_{initial} < k_{final}$), and 440 genes showed a putative stabilization ($k_{initial} > k_{final}$).

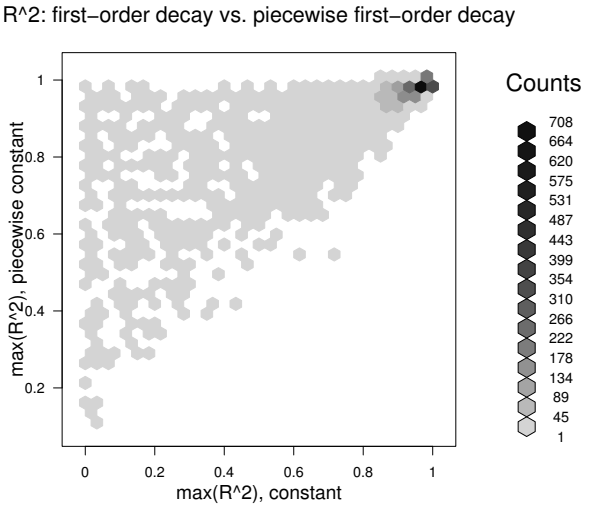


Figure 4.6: Improvement in R^2 values between the two models of mRNA degradation (first-order degradation (”constant”) or piecewise first-order degradation (”piecewise constant”). This plot indicates that for a number of genes there is an apparent improvement in goodness-of-fit by adopting a piecewise constant model over a constant model. Shown are R^2 values for the maximum of ($A = 0$) and ($A \geq 0$) fits (see Model fitting). An adjusted-R^2 value was used to classify genes according to goodness-of-fit (see Methods).

mRNA abundance profiles of putative destabilized genes and of putative stabilized genes were clustered in order to identify groups of genes with similar mRNA abundance behaviour. Clustered mRNA abundance profiles for putative destabilized and stabilized mRNAs are shown in Figures 4.7, 4.9 respectively. Several clusters represent groups of genes with low fold change in both abundance and transcription rate throughout the timecourse; these clusters represent genes with low or no response to oxidative stress and trends in these clusters may be technical rather than biological in origin.

Clusters 6, 7 and 8 of the putative destabilized genes (outlined in red in Figure 4.7) display a delayed decrease in mRNA abundance approximately 15 to 40 minutes after the induction of the stress response (t=0). Of these, clusters 6 and 8 are enriched for genes involved in ribosome biogenesis and assembly (Table 4.2). There is no coherent detected change in transcription rate for the genes in Clusters 6 and 8 (Figure 4.8) which suggests that the observed delayed decrease in mRNA abundance may

Table 4.2: GO term enrichment (biological process) for selected clusters of genes with putative mRNA destabilization 12-60 mins after stress induction. p-values calculated using Fisher's exact test with a FDR correction at $\alpha = 0.05$

Cluster	Enriched GO terms	% of cluster (% of genome) annotated	p-value
Cluster 6 (18 genes)	GO:0042254 (ribosome biogenesis and assembly)	44% (3.7%)	$p = 1.4 \times 10^{-5}$
Cluster 7 (9 genes)	No significant enrichment		
Cluster 8 (20 genes)	GO:0042254 (ribosome biogenesis and assembly) GO:0016070 (RNA metabolic process)	70% (3.7%) 50% (15%)	$p = 8.8 \times 10^{-3}$ $p < 10^{-8}$

be predominantly post-transcriptionally controlled.

4.4.3 First-order mRNA decay during the stress response: evidence of initial stabilization/destabilization

2385 genes were found to have a good fit to the first-order degradation model (adjusted-$R^2 > 0.6$) and were not better explained by a change in mRNA stability 12-60 mins after stress induction. The transcription rate and expression profiles of each of the 2385 genes were concatenated and clustered in order to identify groups of genes with similar transcription rate or mRNA abundance behaviour. Clustering reveals distinct stress response behaviours for induced or repressed genes (Figure 4.10). The amplitude of the transcription rate and mRNA abundance response in each cluster is summarized in Figure 4.11, where the amplitude of the response in a given cluster is defined as the maximal fold-change observed in the median profile of the cluster. Gene clusters with the largest amplitude in mRNA abundance tend to also have the largest amplitude in transcription rate, so that high fold-changes in mRNA abundance are associated with high fold-changes in transcription rate. However, it can be seen from Figure 4.11 that there are clusters which have a large amplitude response in mRNA abundance but a relatively small amplitude response in transcription rate. Clusters 17, 20, 22, and 32 (Figures 4.10 and 4.11) have transcription rate amplitudes amongst the lowest over all clusters but display a two- to four-fold change in mRNA abundance during the stress response. Genes in these clusters are therefore candidates for putative (de-)stabilization early in the stress response.

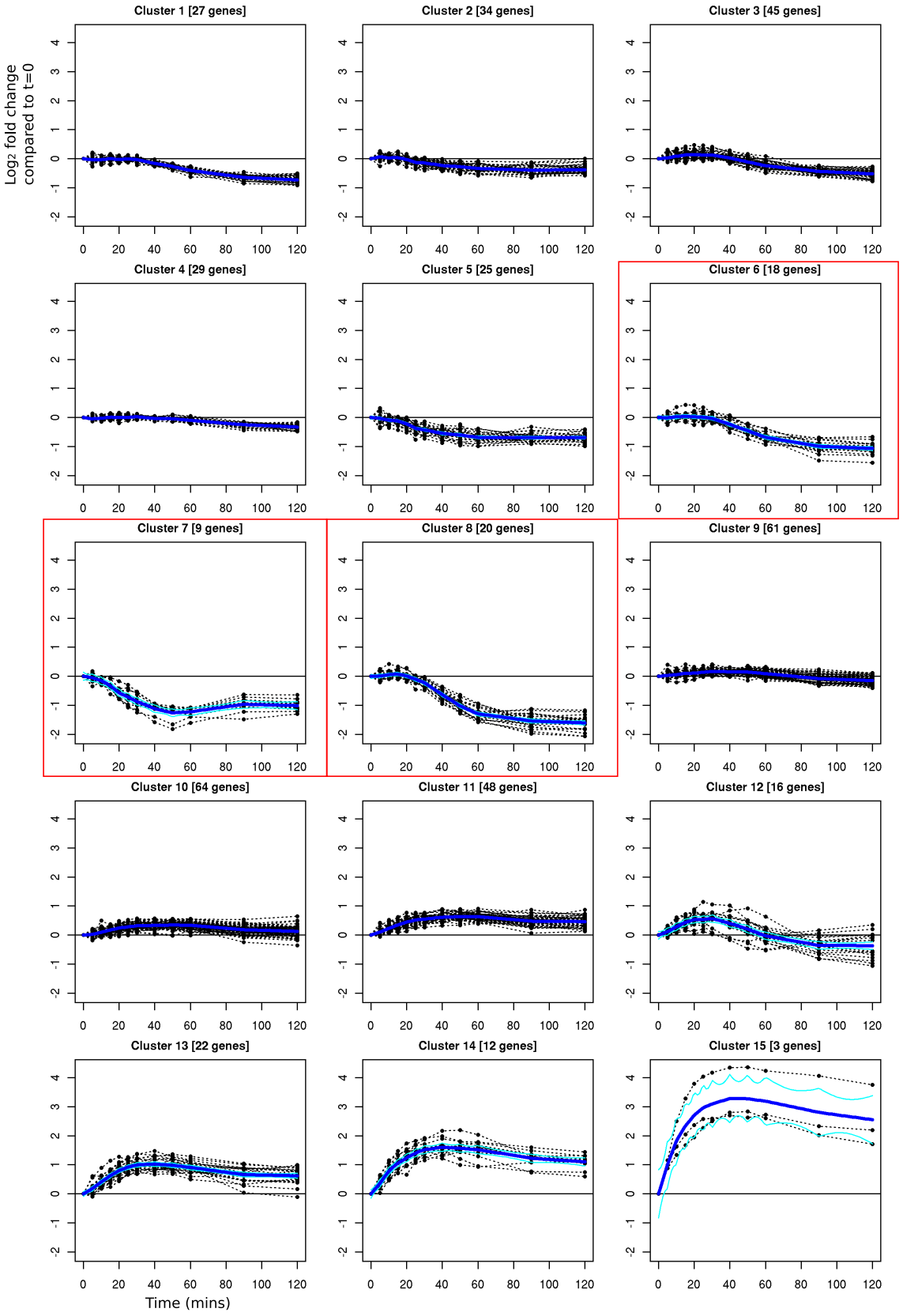


Figure 4.7: Putative delayed destabilization. mRNA abundance profiles of genes with putative destabilized decay rate 12-60 mins after induction of a stress response. Clustering was performed on $\log_2$ normalised ratios relative to the first timepoint (parameters: *prior precision* (β) = 10^{-7}). Predicted profiles are shown in blue. Clusters 6, 7, 8 outlined in red show a delayed decrease in mRNA abundance with predicted fold change >2-fold during the response and are enriched for ribosome biogenesis and assembly. Blue lines show predicted mean cluster profiles, inferred using a spline cluster algorithm (see Methods).

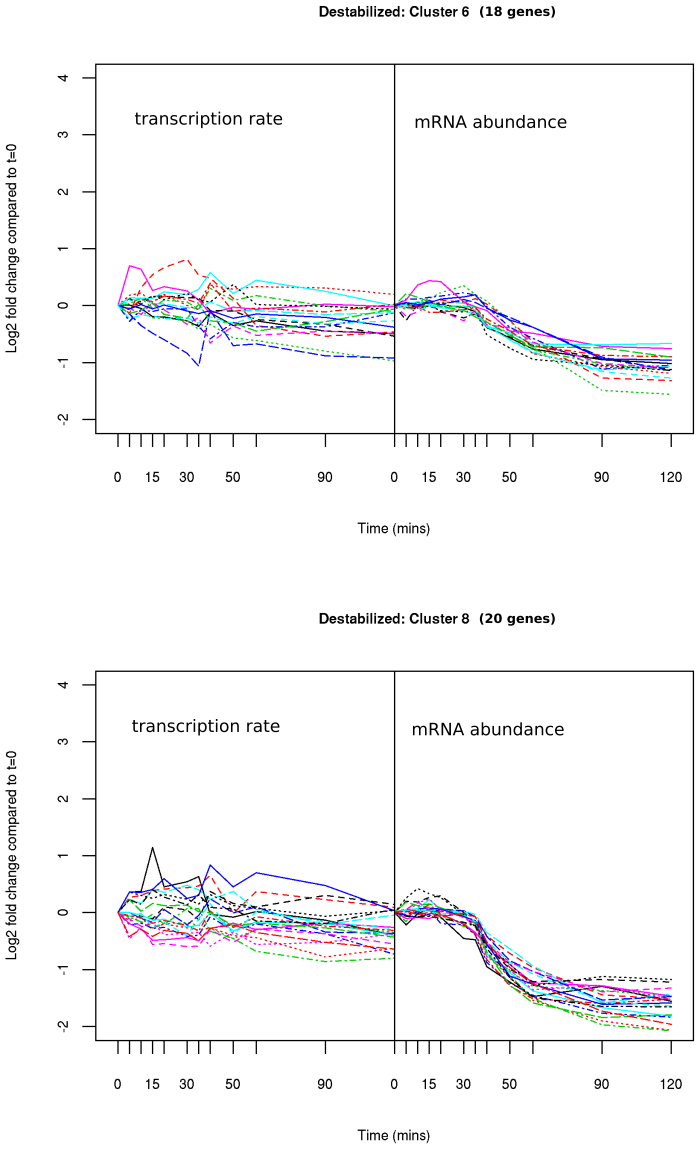


Figure 4.8: Transcription rate and mRNA abundance profiles of putative destabilized clusters (12-60 mins) which are enriched for ribosome biogenesis and assembly. (Clusters 6 and 8 from Figure 4.7). Colours represent individual genes.

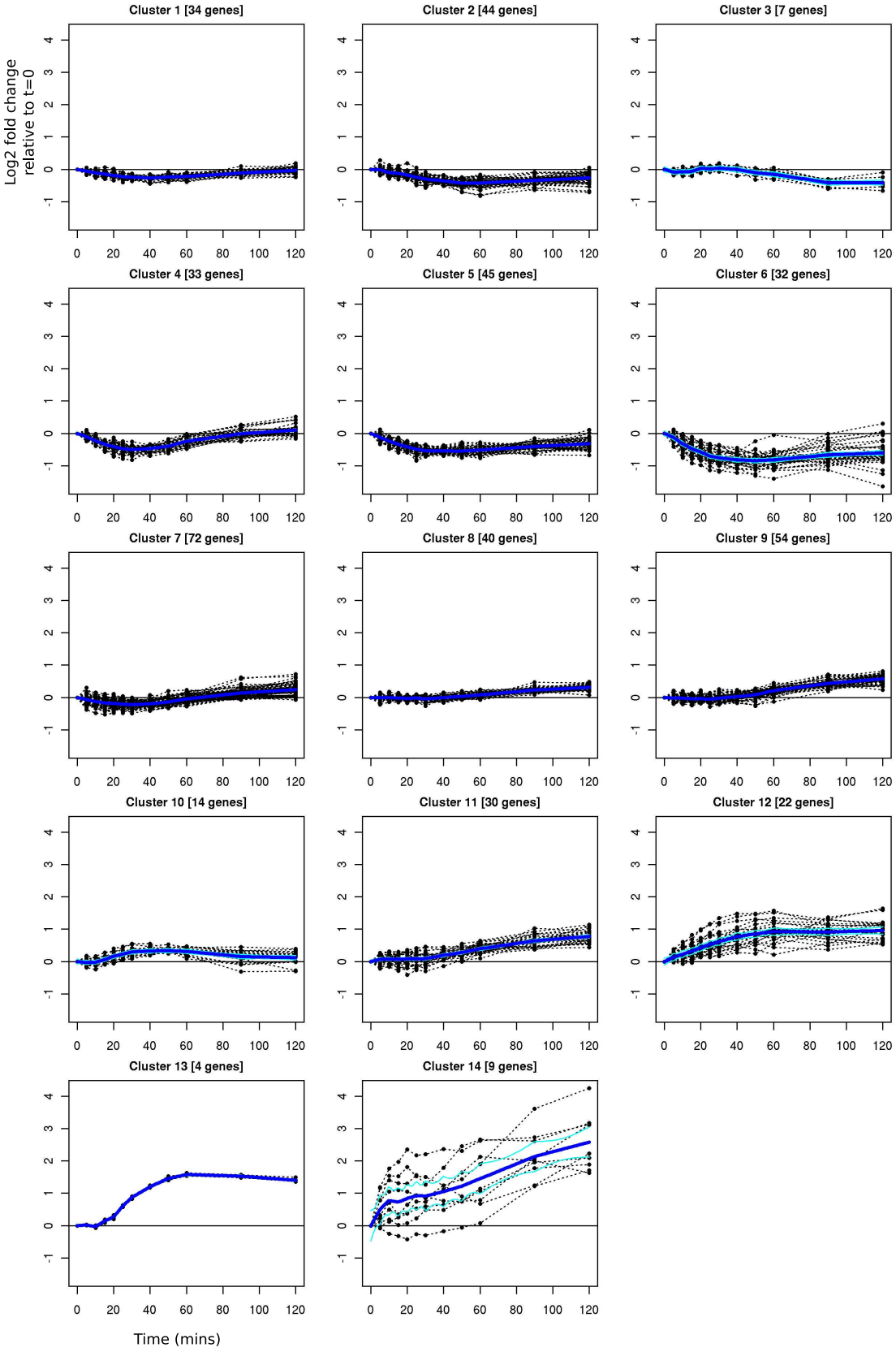


Figure 4.9: Putative stabilization: clustered mRNA abundance profiles for genes with a putative stabilized decay rate between 12-60 mins after induction of a stress response (parameters: *prior precision* (β) = 10^{-7}). All clusters show either $<$ 2-fold change in predicted profile during the stress response, or small or incoherent clusters.

Table 4.3: GO term enrichment (biological process) for selected clusters of genes with putative mRNA (de-)stabilization early in the stress response. p-values calculated using Fisher's exact test with a FDR correction at $\alpha = 0.05$. Cluster names correspond to Figures 4.10, 4.11

Cluster	Enriched GO terms	% of cluster (% of genome) annotated	p-value
Cluster 17 (19 genes)	No significant enrichment		
Cluster 20 (37 genes)	GO:0042254 (ribosome biogenesis and assembly) GO:0016070 (RNA metabolic process)	70% (3.5%) 59% (15%)	$p < 10^{-8}$ $p = 3 \times 10^{-8}$
Cluster 22 (64 genes)	GO:0042254 (ribosome biogenesis and assembly) GO:0016070 (RNA metabolic process)	28% (3.5%) 38% (15%)	$p < 10^{-8}$ $p = 1.4 \times 10^{-4}$
Cluster 32 (52 genes)	GO:0006950 (response to stress)	31% (12%)	$p = 8.6 \times 10^{-3}$

Gene Ontology enrichment analysis shows that Clusters 20 and 22, which are putative destabilized clusters, are enriched for ribosome biogenesis and assembly (Table 4.3). Similarly, clusters already identified as having a putative change in mRNA decay rate later in the stress response were also enriched for ribosome biogenesis and assembly (Figure 4.7; Table 4.2). I identified no clusters that have a relatively large transcription rate amplitude and ia small mRNA abundance amplitude: a high fold-change in transcription rate is always associated with a high fold-change in mRNA abundance during the response.

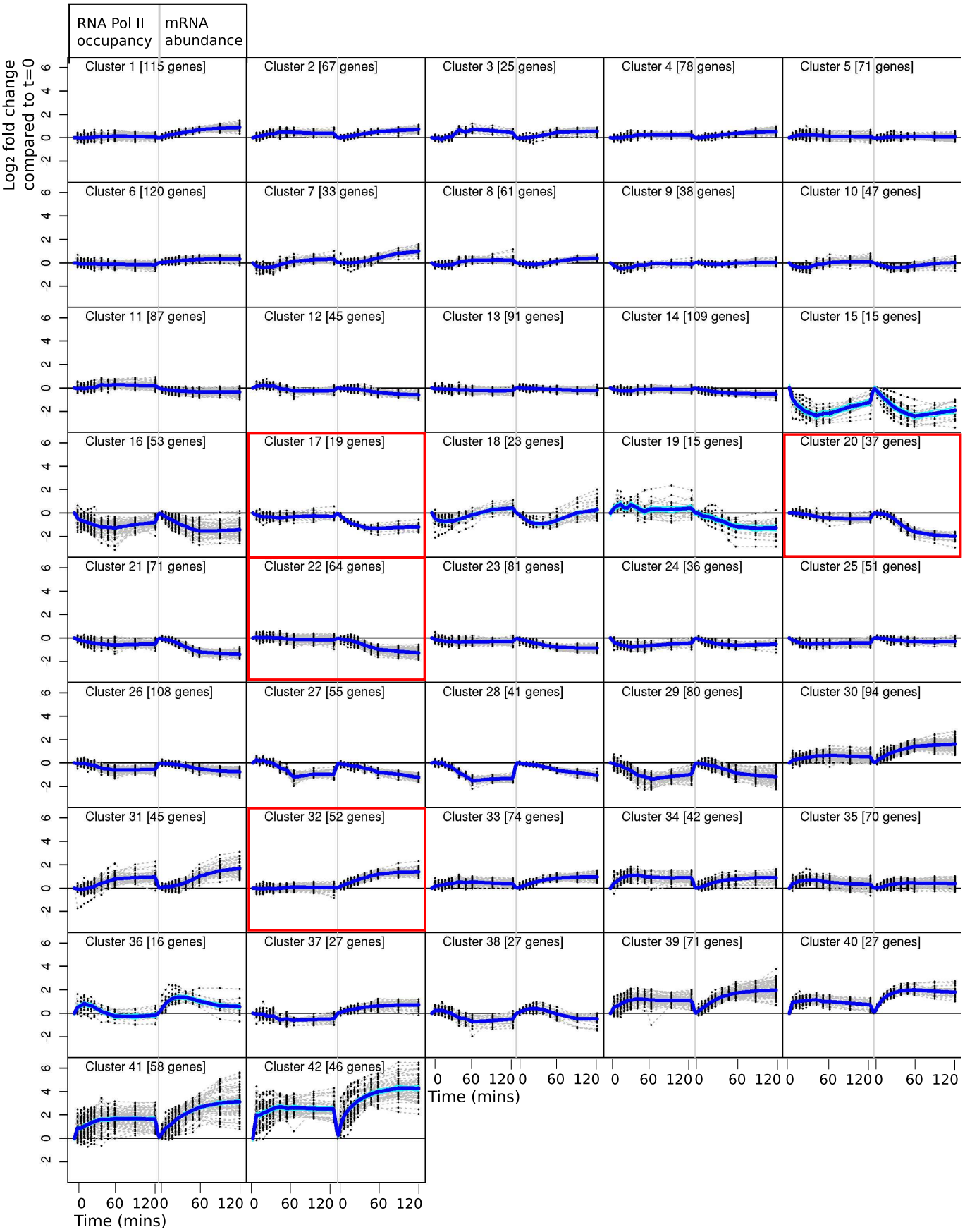


Figure 4.10: Clusters of genes assigned to first-order mRNA decay model. Transcription rate and mRNA abundance profiles were concatenated and clustered using a Bayesian hierarchical spline clustering algorithm (see Methods). **Red:** accumulation or loss of mRNA is detected, but with a comparatively small change in transcription rate.

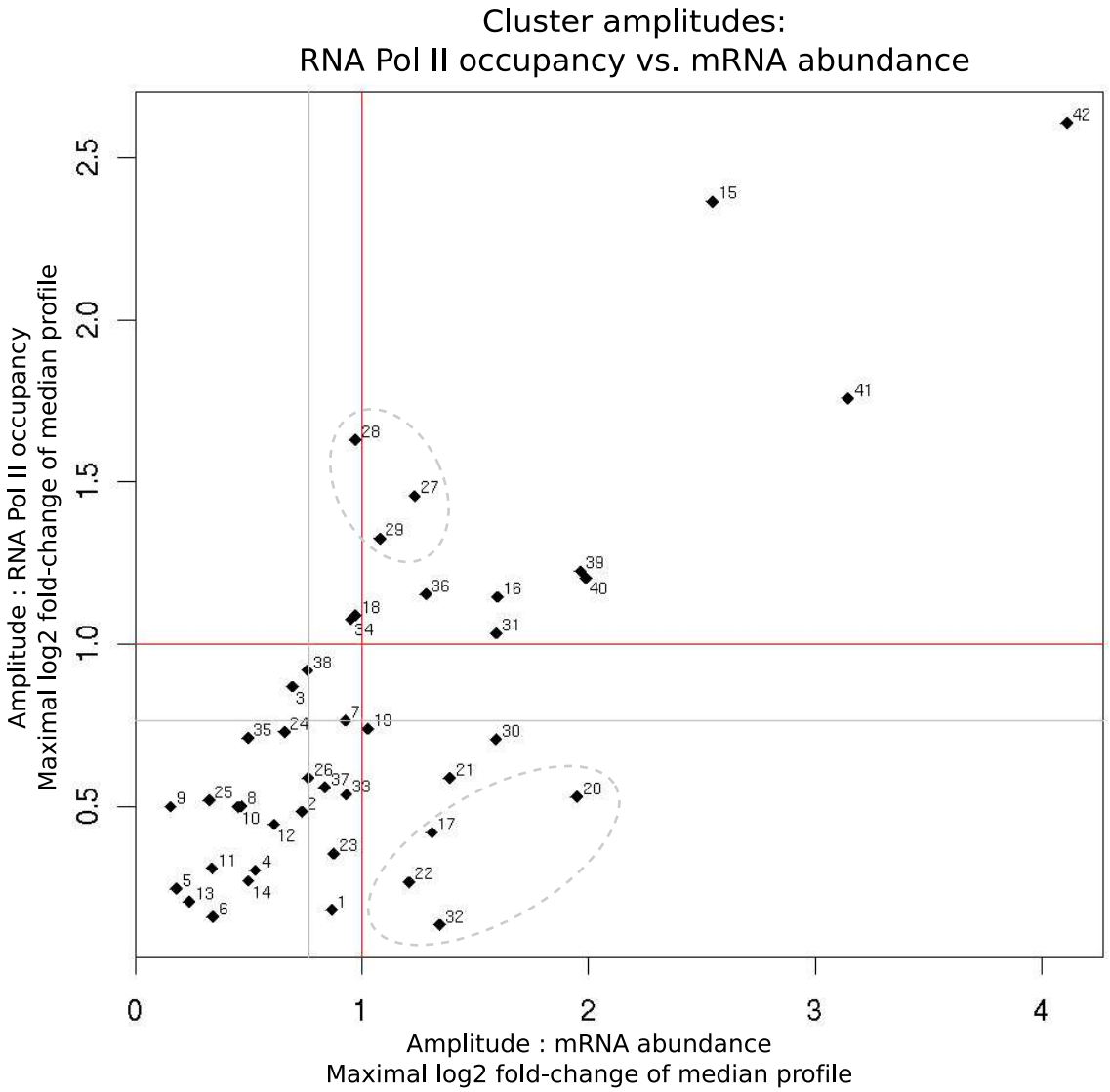


Figure 4.11: Cluster amplitudes for genes classified as having first-order mRNA degradation throughout the stress response. Each point corresponds to the predicted profile of a cluster. The transcription rate amplitude (y-axis) is plotted against the abundance amplitude (x-axis), where amplitude is defined as $(max - min)$) of the ($\log_2$-scale) predicted profile. Circled clusters deviate from the majority of clusters. Cluster 32 shows a high expression amplitude (accumulation of mRNA) but no change in transcription rate. Clusters 17, 20, 22 show a high expression amplitude (loss of mRNA) but no change in transcription rate (enriched for ribosome biogenesis and assembly). Clusters 27, 28, 29 show a similar expression amplitude (loss of mRNA) but also show a drop in transcription rate (enriched for ribosomal proteins). [Transcription rate and mRNA abundance profiles for each cluster are displayed in Figure 4.10.]

4.4.4 Dynamics of induction and repression

The preceeding sections explored the mRNA abundance and transcription rate profiles observed in response to oxidative stress. Two models of mRNA degradation were fitted to the observed profiles: first-order mRNA decay, and first-order decay with a single change in the mRNA decay rate constant. Model selection was performed in order to classify genes according to the most appropriate model. It was found that a first-order mRNA decay model is able to explain most gene profiles (adj$R^2 > 0.6$), whereas a subset of highly repressed genes are better explained by mRNA destabilization 12-60 mins after stress induction. Cluster analysis revealed that highly repressed gene clusters with immediate or delayed mRNA destabilization are enriched for genes with GO annotation related to ribosome biogenesis and assembly, whereas clusters of highly repressed genes which are also transcriptionally repressed at the point of stress induction are enriched for ribosomal proteins (Figure 4.12).

Using a cluster analysis of transcriptionally induced genes, I identified coherent clusters of transcription rate and mRNA abundance profiles (Figure 4.13). In particular, genes which display a rapid accumulation of mRNA at the onset of stress are also rapidly transcriptionally induced at the onset of stress (e.g. Figure 4.13: Clusters 1, 4, 5). Similarly, mRNA which begins to accumulate later in the stress response (approximately 20-40 mins after stress induction) is also transcriptionally induced at approximately the same time (e.g. Figure 4.13: Clusters 2, 10, 15). The small number of transiently induced or transiently repressed gene clusters in this dataset also exhibit similar transcription rate and mRNA abundance profiles (e.g. Figure 4.13: Clusters 11, 19).

Approach to a new steady state

Many highly induced genes appear to reach a final steady state by the end of this stress response timecourse. For example, the predicted profiles of Clusters 1, 3, 5

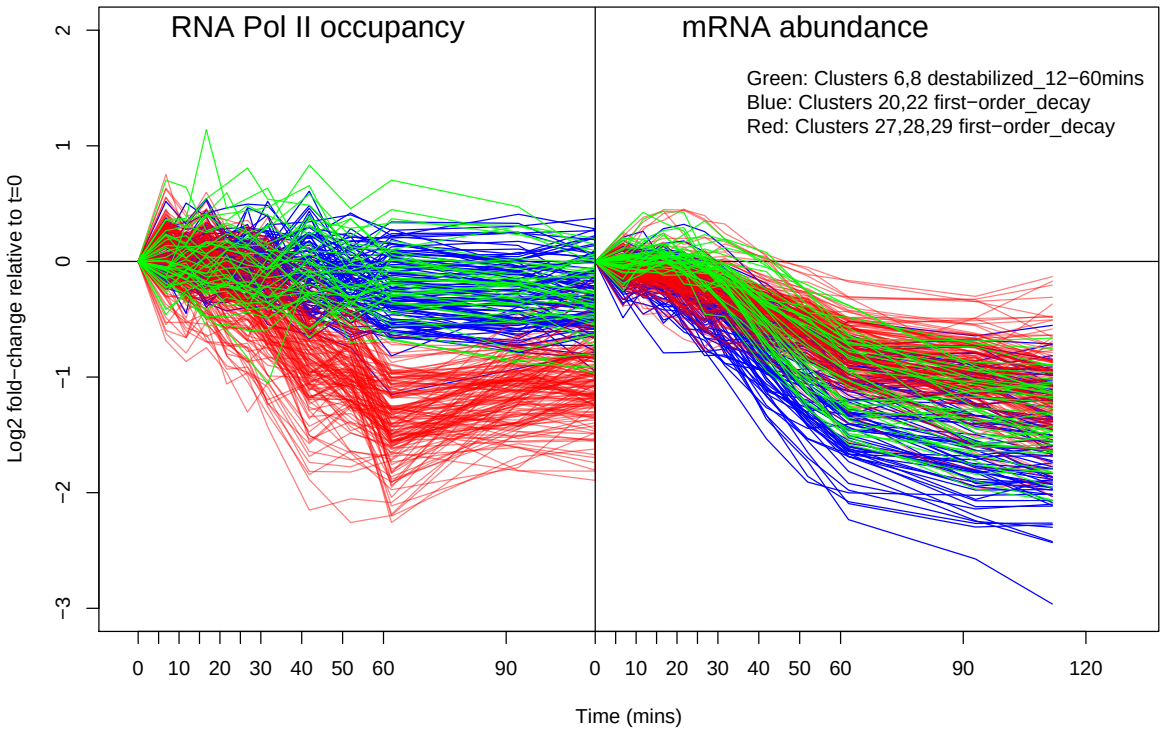


Figure 4.12: Clusters of genes show a loss of mRNA in response to oxidative stress but have distinct temporal transcription and mRNA abundance profiles. **Green** (destabilized) **/ Blue** (first-order decay)**:** Transcription rate remains at a constant level; enriched for GO (biological process) ribosome biogenesis and assembly. **Red:** Decrease in both transcription rate and mRNA abundance; enriched for GO (cellular component) ribosomal proteins.

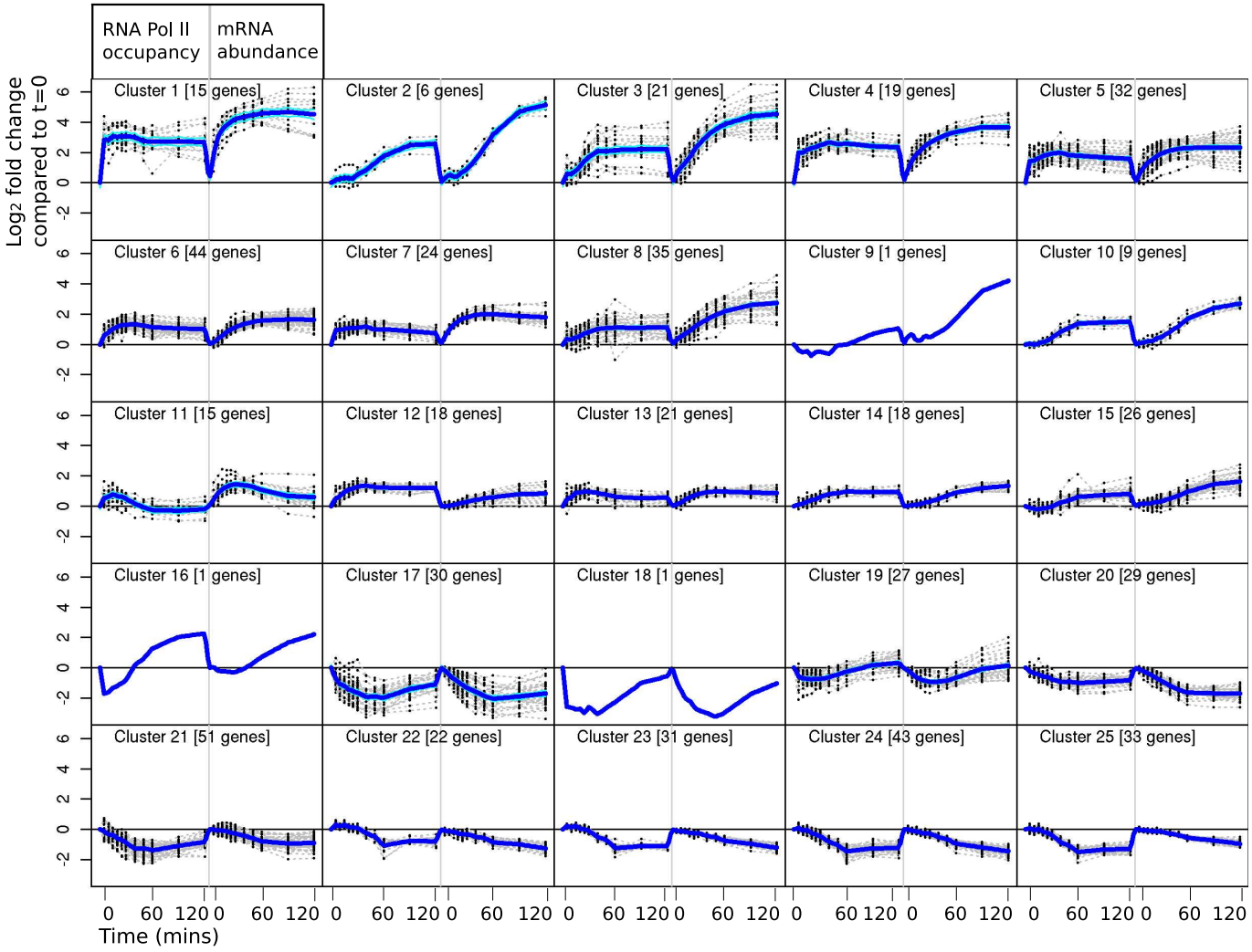


Figure 4.13: Clusters of transcriptionally induced or repressed genes. Clusters from Figure 4.11 with median fold-change > 2 in transcription rate were reclustered. Cluster analysis revealed coherent transcription rate and mRNA abundance profiles amongst transcriptionally induced genes (*priorprecision* = 10^{-7}, *numberofclusters* = *25*).

and 6 in Figure 4.13 show approximately constant transcription rate and mRNA abundance for t = 60, 90, 120 minutes. Assuming that highly induced genes are initially at a steady state of transcription rate and mRNA abundance and that a final steady state is reached by the end of the timecourse, candidate genes for mRNA (de-)stabilization can be identified by comparing the fold change in both transcription rate and mRNA abundance between initial and final steady states. Assuming initial (I) and final (F) steady states, then transcription rate R, mRNA abundance E, and decay rate constant k satisfy:

$$\frac{dE}{dt} = R - kE = 0 \implies k_I = \frac{R_I}{E_I}; \quad k_F = \frac{R_F}{E_F}; \quad \frac{k_I}{k_F} = \frac{R_I/R_F}{E_I/E_F} = \frac{\tau_{\frac{1}{2}F}}{\tau_{\frac{1}{2}I}} \tag{4.4}$$

where mRNA halflife $\tau_{\frac{1}{2}}$ is given by $\tau_{\frac{1}{2}} = ln(2)/k$. Therefore changes in mRNA halflife can be identified by looking at the ratio of fold-change in transcription rate between steady states to fold-change in mRNA abundance between steady states.

Cluster 20 (Figure 4.13) has approximately two-fold lower final ('steady-state') transcription rate compared with the initial ('steady-state') transcription rate, but approximately four-fold lower mRNA abundance compared with the initial mRNA abundance. If Cluster 20 is initially at steady state and reaches a final steady state then this suggests that mRNA is destabilized in response to stress, with the halflife approximately halved at the final steady state compared to the initial steady state. Further experimental evidence is required to demonstrate that the mRNA of these genes is at steady state in exponentially growing cells immediately prior to stress induction. Cluster 20 is enriched for genes previously reported to be downregulated in response to oxidative stress, but also enriched for reported periodic genes: (i) core environmental stress response (57% of cluster (29 genes); 7% rest of genome; $p < 10^{-8}$ (FDR $\alpha = 0.05$); (ii) downregulated in response to oxidative stress [110] (90% of cluster; 8% rest of genome; $p < 10^{-8}$ (FDR $\alpha = 0.05$); (iii) periodic genes [131] (20% of cluster; 2% rest of genome; $p = 0.003$ (FDR $\alpha = 0.05$).

Evidence of rapid early mRNA stabilization

Exponential approach of mRNA abundance to a new steady state is a solution for the first-order mRNA decay model (Eqn 3.4) in response to an instantaneous change in transcription rate or decay rate constant, or both the transcription rate and the decay rate constant, at the onset of the stress response. Assuming that the mRNA species is at steady state before stress induction, an instantaneous change in either the transcription rate or decay rate parameter to a new constant value results in an exponential approach to a new steady state:

1. An instantaneous change in transcription rate from R_0 to R_f changes the abundance to approach E_F from E_I exponentially at a rate determined by k:

$$E(t) = \frac{1}{k}\left[R_F + (R_I - R_F)e^{-kt}\right] \tag{4.5}$$

2. An instantaneous change in mRNA stability from k_I to k_F changes the abundance to approach E_F from E_I exponentially at a rate determined by k_F:

$$E(t) = R\left[\frac{1}{k_F} + \left(\frac{1}{k_I} - \frac{1}{k_F}\right)e^{-k_F t}\right] \tag{4.6}$$

3. Simultaneous instantaneous changes in both stability ($k_I \rightarrow k_F$) and transcription rate ($R_I \rightarrow R_F$)

$$E(t) = \frac{R_F}{k_F} + \left(\frac{R_I}{k_I} - \frac{R_F}{k_F}\right)e^{-k_F t} \tag{4.7}$$

Transforming these expressions for use with microarray measurements to account for arbitrary scaling and shifting between transcription arrays and expression arrays (Chapter 3, page 48) does not alter the form of the exponential approach solution. In this timecourse, we identified genes with an exponential approach of mRNA abundance to a new steady state, as a special case of observed mRNA abundance profiles.

These genes were considered to be candidate genes for rapid initial mRNA (de-)stabilization and/or change in transcription rate. Observed fold-changes in transcription rate and mRNA abundance between initial and final steady states were then used to discriminate between candidates for rapidly modulated mRNA stability or a solely transcriptional response (Equation 4.4).

The model of exponential approach of mRNA abundance to a new steady state is identical to the first-order degradation model (Equation 3.4 on page 41) with *A=0*. Amongst all genes which were not allocated to the piecewise decay model, 910 genes were found to be best explained (greatest adjusted-R^2) by exponential approach to a new steady state. The mRNA abundance profiles of these genes were clustered in an attempt to discriminate between biologically informative exponential approach profiles and genes showing weak trends in mRNA abundance which were seen amongst the least responsive genes (Figures 4.10 and 4.11). Most of the resulting clusters were discarded due to low fold-change of mRNA abundance. Two clusters with a combined size of 78 genes had cluster median fold-change greater than two-fold and were enriched for genes previously reported to be upregulated as part of the core environmental stress response (89% of gene list; 8% rest of genome; $p < 10^{-8}$ (FDR $\alpha = 0.05$)). Cluster analysis of the transcription rate profiles of these 78 genes revealed three distinct transcription rate behaviours (Figure 4.14):

1. constant transcription rate;
2. rapid moderate increase to new steady transcription rate, approximately two-fold within 5 minutes of stress induction;
3. rapid large increase to new steady transcription rate, approximately 3- to 16-fold within 5 minutes of stress induction

To estimate the magnitude of rapid changes in mRNA stability for these 78 genes, exponential approach solutions were re-fitted[4] to provide point estimates of $\frac{E_I}{E_F}$ and

[4]Scaling the transcription rate and mRNA abundance profiles in the original fits means that parameter estimates other than the decay rate constant are not comparable between genes. For the

$\frac{R_I}{R_F}$. A comparison of estimated ratios of initial steady-state to final steady-state conditions for transcription rate and mRNA abundance suggests that there may be widespread rapid mRNA stabilization amongst genes which are rapidly, persistently and highly induced in response to oxidative stress (Figure 4.14C). Rapid stabilization of mRNA may explain the larger observed fold-changes in mRNA abundance than in transcription rate. Atf1 is amongst the 78 selected genes (labelled in Figure 4.14C); aft1 mRNA has been reported to stabilize in response to oxidative stress, contributing to rapid mRNA accumulation [119]. Quantitative estimates of a change in mRNA stability must be treated with caution, however, until the observed fold-changes in expression have been validated: for example, fold-change in mRNA abundance for persistently induced genes could be validated using low-throughput techniques. In addition, estimates of fold-change between the first timepoint and any combination of subsequent timepoints must be treated with caution because the first timepoint is not replicated in this study.

4.4.5 Post-transcriptional regulation: searching for transcript sequence motifs

Transcript sequences of putatively stabilized or destabilized mRNA were searched for overrepresentation of short words which may indicate targeted binding by small regulatory RNAs. Gene lists of putative stabilized or destabilized mRNA were defined by combining previously identified clusters, listed in Table 4.4.

The nucleotide sequences of putative stabilized and destabilized mRNAs were searched for enrichment of short words which may indicate targeted binding by seed regions of small regulatory RNAs. An enrichment and depletion search was performed for all possible 6-mer words (Markov correction = 4) [130] amongst putative stabilized or destabilized mRNA sequences. The search did not identify any significant en-

special case of exponential approach of mRNA abundance it is of interest to estimate the asymptotic 'steady-state' of mRNA abundance and the constant final transcription rate in order to estimate the contribution of rapid mRNA (de-)stabilization. The model $y(t) = B + De^{-kt}$ was therefore re-fitted to unscaled observed fold-change profiles, fold-change relative to $t = 0$.

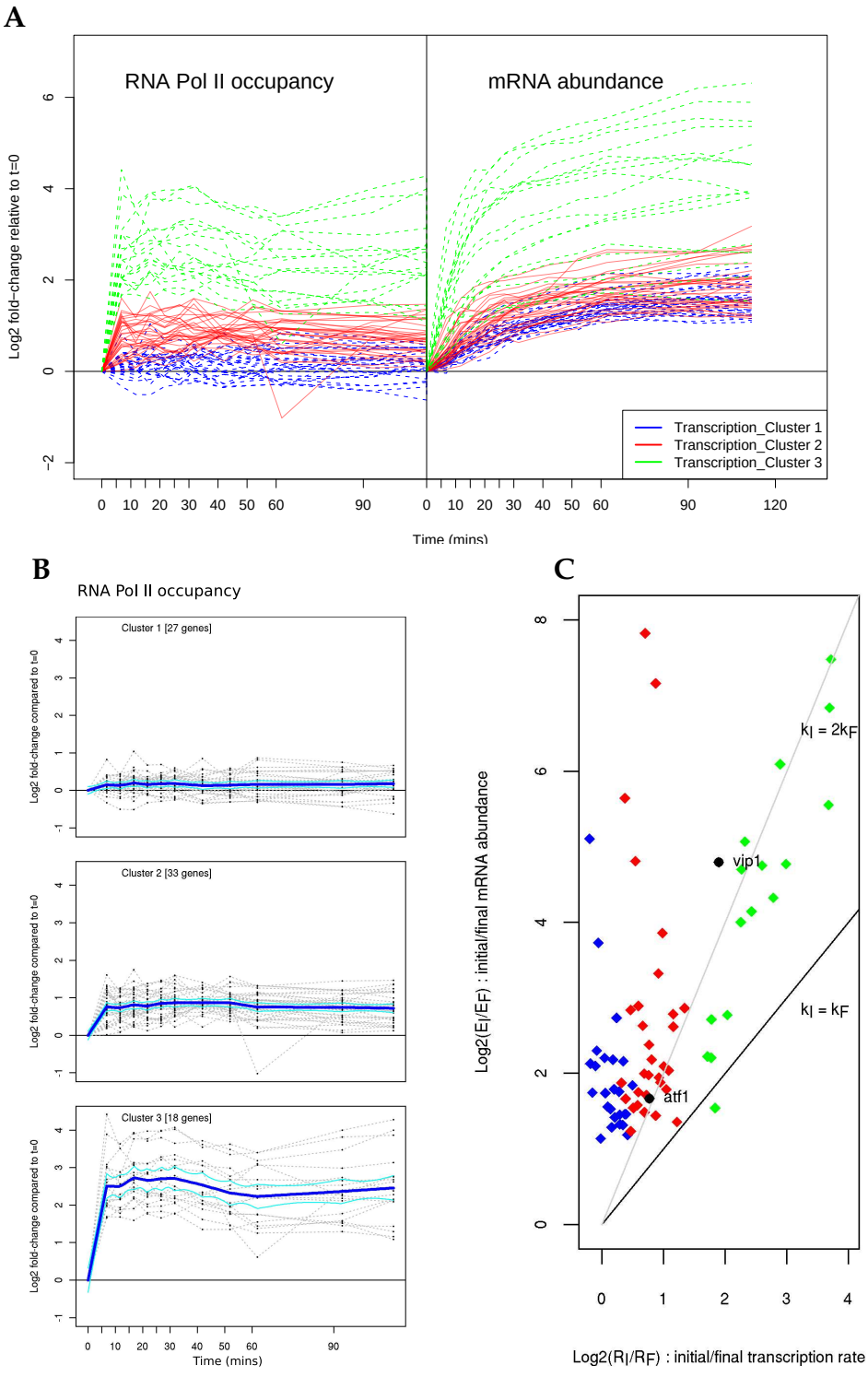


Figure 4.14: Transcription rate and mRNA abundance profiles of 78 genes with exponential approach of mRNA abundance. **A.** Transcription rate (left) and mRNA abundance. Colours correspond to transcription rate profile clustering (**B**) as labelled. **C.** Point estimates of fold-change in transcription rate (*x-axis, log*$_2$) and mRNA abundance (*y-axis, log*$_2$) for each gene. $\frac{E_F}{E_I}$ estimated as the median of the transcription rate profile, excluding the first timepoint. $\frac{R_F}{R_I}$ estimated as $B + D$ where mRNA abundance is modelled as $y(t) = B + De^{-k_F t}$. Colours correspond to transcription profile clusters (identified in figures A, B).

Table 4.4: Gene lists of putative stabilized and destabilized mRNA, selected for use in sequence searches for short words and sequence motifs.

Gene list	Combined clusters	# genes
Putative stabilized	Cluster 32 (Figure 4.10 Clusters 1, 2, 3 (Figure 4.14)	115 genes
Putative destabilized	Clusters 6, 8 (Figure 4.7) Clusters 17, 20, 21, 22 (Figure 4.10)	229 genes
Putative destabilized (delayed)	Clusters 6, 8 (Figure 4.7)	38 genes

richment or depletion which could not be explained by the presence of duplicated genomic sequences in the selected gene list (Figure 4.15). Similar results were obtained for 5-mer and 7-mer words (Markov correction = 3, 5, respectively; Table 4.5). Amongst putative stabilized mRNAs, several 6-mers were reported to have a sharp approach to a broad peak (Figure 4.15) but this enrichment was found to correspond to a group of identical Tf2 transposable elements in the selected gene list. Tf2 transposable elements exist in several copies in the *S. pombe* genome and the overrepresentation of Tf2 sequences in the selected gene list was detected as sharp increases to broad peaks in the enrichment p-value landscape for approximately ten 6-mers. The detected 6-mer 'ACCTAG', for example, is present in 5 copies in each of the four Tf2 GeneDB transcripts. All Tf2 transposable elements were subsequently removed from the gene list (Tf2-1, Tf2-5, Tf2-6, Tf2-7) and the analysis was repeated. No other enrichment or depletion of short words was found amongst putative stabilized or destabilized mRNA.

4.5 Discussion

In this study, I investigated whether there is evidence of mRNA stabilization or destabilization in response to an oxidative stress (0.5mM hydrogen peroxide) either immediately at the point of stress or later in the stress response. Simultaneous timecourse data were obtained from ChIP-chip arrays and expression arrays at 12 timepoints during the first two hours of oxidative stress response in *S. pombe*. This experimental design specifically excludes the blocking of transcription, for example

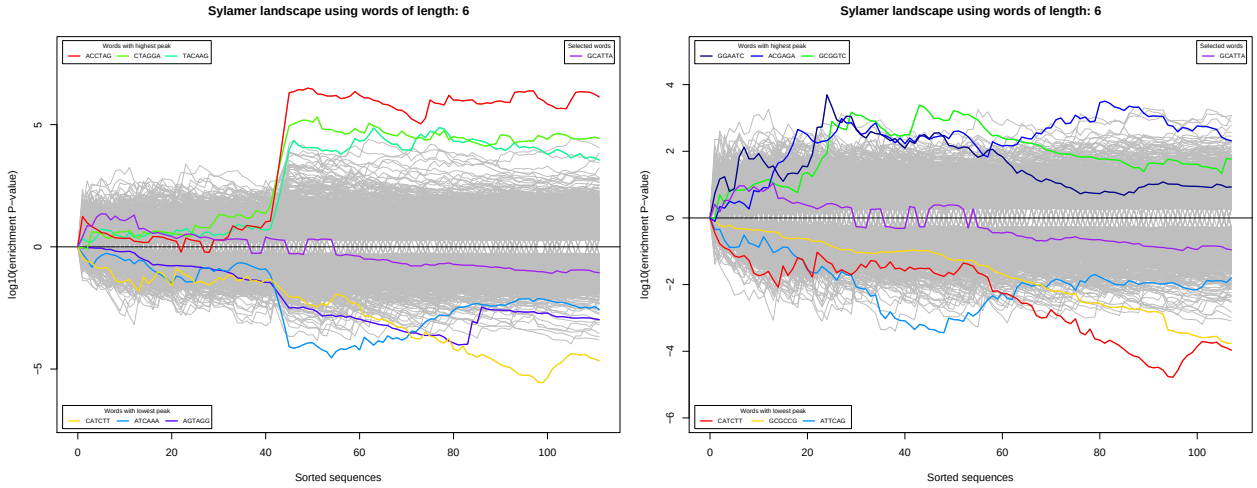


Figure 4.15: Sylamer landscapes for list of putative stabilized genes, searching for enrichment or depletion of all 6-mers, Markov correction = 4, compared to a background of annotated complete transcripts (GeneDB coding sequences flanked by GeneDB annotated 5′/3′ UTRs, if available). Putative stabilized genes are defined as Cluster 32 (Figure 4.10; 52 genes) and Clusters 1, 2, 3 (Figure 4.14; 27 genes, 33 genes) (duplicate entries removed). Words with the 3 highest and 3 lowest peaks are highlighted; purple: arbitrary 6-mer word. **Left:** a sharp increase to a broad peak of enrichment *p*-values indicates enrichment of 6-mers in the lower portion of the gene list. **Right:** enrichment is lost when Tf2 transposable elements are removed from the list of putative stabilized genes. The shape and size of the depletion peak for the word 'CATCTT' is typical of enrichment/depletion peaks seen in random gene lists [130]

Table 4.5: Sylamer 6-mer enrichment/depletion search amongst putative stabilized and destabilized mRNA transcripts, Markov correction = 4. **a.** UTR lengths taken to be (1) annotated UTRs (GeneDB), (2) per-gene median UTR length over all conditions (Wilhelm *et al.* 2008 [50]), (3) 500 base pairs. **b.** Gene lists (see Table 4.4). **c.** Peak log10(enrichment *p*-value). **d.** Empirical quantile: proportion of 100 randomly selected gene lists with higher peak enrichment / lower peak depletion than the observed peak value.

Version: UTR length[a]	Gene list[b]	Peak score[c] (high/low)	Peak quantile[d] (high/low)
Annotated (GeneDB)(1)	Putative stabilized	6.49 / -5.56	0.14 / 0.05
	– ″ – (Tf2 removed)	3.69 / -4.78	0.88 / 0.16
	Putative destabilized	4.54 / -4.58	0.48 / 0.26
	Destabilized (delayed)	3.94 / -3.48	0.41 / 0.57
Median (conditions)(2)	Putative stabilized	6.82 / -5.98	0.11 / 0.03
	– ″ – (Tf2 removed)	3.73 / -4.67	0.86 / 0.13
	Putative destabilized	4.69 / -4.26	0.40 / 0.55
	Destabilized (delayed)	4.32 / -3.79	0.25 / 0.34
500 base pairs(3)	Putative stabilized	5.77 / -4.99	0.12 / 0.08
	– ″ – (Tf2 removed)	3.61 / -4.28	0.95 / 0.39
	Putative destabilized	4.28 / -4.37	0.68 / 0.63
	Destabilized (delayed)	3.55 / -3.70	0.82 / 0.57

by 1,10-phenanthroline or using a heat-shock mutant. Transcription-block studies allow direct estimation of mRNA decay rates during the period of transcriptional inhibition, but post-transcriptional processes are not blocked and transcriptional inhibition itself causes a compounding stress response [102].

The kinetics of transcription and mRNA decay during the stress response were investigated by combining two approaches. Firstly, two models of mRNA degradation were fitted to the observed transcription rate and mRNA abundance profiles: constant mRNA decay after stress induction (first-order decay), and a single change in the mRNA decay rate between 12 mins and 60 mins after stress induction (piecewise first-order decay).

A conservative list of genes was identified as being better explained by a delayed change in mRNA stability than by a constant mRNA decay rate after stress induction, while a model of constant decay rate was found to have good fit to most remaining genes. Secondly, genes assigned to each degradation model were clustered (either by mRNA abundance only, or by concatenated transcription rate and mRNA abundance) to reveal coherent gene clusters with various magnitudes of response in transcription rate or mRNA abundance. Most genes are assumed to be at transcriptional and mRNA abundance steady-state in exponential growth before stress induction[5]. Gene clusters which displayed coherent accumulation or loss of transcripts together with small or incoherent changes in observed transcription rate (compared to all other clusters) were therefore considered to be candidates for rapid immediate mRNA stabilization or destabilization.

Alternative models of mRNA degradation are biologically plausible. In the absence of measurements of absolute quantities of mRNA abundance and transcription rate, however, models with more parameters become unidentifiable. The models considered here are not nested so it is not straightforward to use a likelihood ratio test for model selection [109]. Adjusted-R^2 values were used to compare models while

[5]Exceptions are periodically expressed genes associated with the cell cycle

penalising models which have more parameters.

Through model selection and subsequent clustering, I obtained a dataset of genes with coherent transcription rate and mRNA abundance behaviour in reponse to oxidative stress. Gene clusters were identified which are highly repressed at the level of mRNA abundance but which display no change in transcription rate. These genes are therefore candidates for putative mRNA destabilization either rapidly after stress induction (0-12 mins) or delayed (12-60 mins). Four such clusters are enriched for GO term ribosome biogenesis and assembly. In contrast, gene clusters which showed both highly reduced transcription rate and mRNA abundance are enriched for ribosomal proteins. This indicates that loss of mRNA of ribosomal proteins is explained in part by a reduction in transcription rate, whereas loss of mRNA of ribosome biogenesis and assembly factors may be primarily due to rapid (0-12 mins) or delayed (approximately 30 mins) mRNA destabilization. This is consistent with findings by Grigull *et al.* [102] in *S. cerevisiae* that mRNA in the functional categories of ribosomal RNA biogenesis and ribosome assembly may be destabilized in response to heat shock, mediated by the mRNA deadenylase component Ccr4.

In this particular stress response timecourse, many genes appear to approach a new steady state towards the end of the timecourse (120 minutes). Few genes display a transient response on this timescale. Cluster analysis of genes which are relatively highly induced in both transcription rate and mRNA abundance revealed clusters displaying a rapid upregulation of transcription rate to a new steady transcription rate, and clusters with a delayed (approximately 30 mins) gradual upregulation of both transcription rate and mRNA abundance. Genes with an exponential approach of mRNA abundance to an apparent final steady state were analysed as a special case of transcriptional upregulation in order to estimate the contribution of regulated transcription rate to the mRNA abundance response. Initial and final transcription rate and mRNA abundance were compared. The observed discrepancy between changes in transcription rate and changes in mRNA abundance amongst genes which appear to reach a final steady state indicates possible mRNA stabi-

lization amongst genes which are rapidly and persistently upregulated during the stress response. This finding is concordant with the findings of Shalem *et al.* [98] that genes which are persistently induced are stabilized compared to exponential growth conditions. However, further studies with a replicated initial timepoint would be required in order to gain more accurate estimates of transcription rate fold-change when the transcription rate is upregulated within the first few minutes (and therefore captured by a single datapoint in this timecourse), and to verify that initial and final steady states are in fact reached.

Selected gene clusters of putatively stabilized and putatively destabilized genes were investigated for evidence of targeted post-transcriptional regulation through complementary short non-coding RNAs. However, no enrichment of any short nucleotide motifs was found to be overrepresented amongst the sequences of the gene clusters compared to background sequences. The presence of sequence and structure motifs in targets of RNA-binding proteins can partly explain the differential half-lives in a population of mRNA species [100, 101]. It is possible that targeted binding by RNA-binding proteins also regulates mRNA degradation in response to environmental conditions. *De novo* identification of combined sequence and secondary structure motifs is a challenging computational problem even amongst gene lists which are know to be targeted by specific RNA-binding proteins [125, 132, 133]. Further work would look at the presence of sequence and secondary structure motifs and the ability to discriminate selected gene clusters on the basis of secondary structure motifs, potentially indicating targeted binding.

There could be alternative degradation models that better explain the observed transcription rate and abundance profiles. The complexity of the models which can be considered is limited by the number of timepoints and the noise in the dataset. In particular, a rapid change in mRNA stability coupled with an asynchronous stress response in the cell population may result in observing a progressive change in the decay rate constant from one steady rate constant to another over a period of several minutes. Such behaviour may be modelled using a generalized logistic function, for

example (Figure 3.1). Here I chose to use simple models with a small number of parameters to avoid overfitting.

A timecourse which includes a full recovery to a pre-stress state may reveal further phases of the stress response, defined by several changes in transcription rate and mRNA stability for each mRNA species. This implies a further model of mRNA degradation in which there are two or more instantaneous changes in mRNA stability at different (unknown) times during the stress response. As the number of parameters in the governing equation for mRNA abundance increases, we encounter difficulties with parameter inference when attempting to fit a highly parameterized model to a small number of timepoints.

An assumption of this study is that RNA polymerase II ChIP-chip measurements are a useful estimate of changes in the RNA polymerase II occupancy of transcribed regions and therefore of changes in transcription rate. The use of microarray probes targetting the 3' end of transcripts and the use of cell populations are assumed to minimize the impact of polymerase stalling and early transcript termination on the observed changes in transcription rate. An additional assay, such as the presence of H3K36me3 histone modifications [134], could serve an independent indicator of transcriptionally active regions and provide further evidence that the observed changes in transcription rate were detected in transcriptionally active regions.

This study has identified candidate genes for rapid or delayed mRNA destabilization in response to oxidative stress in *S. pombe*, including clusters of genes with coherent functional annotation. Genes contained in the clusters highlighted in this chapter are listed in the Appendix.

Chapter 5

Integration of global gene expression studies in *Fusarium graminearum*

This chapter presents an integrative study of gene expression patterns in the fungal crop pathogen *Fusarium graminearum*. In contrast to the yeasts *S. cerevisiae* and *S. pombe* studied in previous chapters, *F. graminearum* is not a model organism and relatively little is known about the regulation of developmental stages and pathogen-host pathways on a genomic scale. Following the release of the *F. graminearum* genome sequence in 2003, a number of transcriptomics datasets are now available permitting an early integrative study of *F. graminearum* gene expression patterns.

5.1 Introduction

The *Fusarium* genus contains a diverse collection of plant-pathogenic fungi which cause disease in a wide range of plants and opportunistic infections in humans. The *Fusarium graminearum* species complex is a major cause of disease in cereal crops worldwide, causing blights, rots and wilts in wheat and barley, reducing crop yield and producing mycotoxins which are harmful to human health. Despite the economic impact of *F. graminearum* as a crop pathogen, the *F. graminearum* species complex was only recently described and relatively little is known about the function and regulation of *F. graminearum* genes on a genomic scale.

5.1.1 *F. graminearum* sequencing and gene calls

F. graminearum (strain PH-1) was the first complete *Fusarium* genome to be sequenced [135]. Two automatic gene call sets were produced from the first draft genome released in 2003: the Broad *FG1* gene call set, and the MIPS draft gene call set [135, 136, 137]. Gene calls are manually corrected and curated at the Broad Institute Fusarium Comparative Project and the MIPS *F. graminearum* Genome Database. An oligonucleotide DNA microarray, Affymetrix GeneChip Fusariuma520094, was designed based on a combined gene call set containing manually processed gene calls and predicted genes from both the MIPS draft gene calls and the Broad *FG1* gene calls [138]. Transcriptomics datasets from studies based on the Fusariuma520094 GeneChip are deposited at the PLEXdb plant expression database [139].

5.1.2 Study aims

This aims of this study were twofold. Firstly, all publicly available *F. graminearum* transcriptomics studies were analysed in order to identify genes which are differentially expressed during stages of the *F. graminearum* lifecycle and crop infection. This analysis provides a dataset of summarized expression patterns and groups of coexpressed – and potentially coregulated – *F. graminearum* genes. Secondly, the summarized coexpression groups were investigated for clues about the transcriptional regulation of identified coexpressed genes: (i) the presence and protein domain composition of predicted DNA-binding transcription factors within each coexpression group; and (ii) constraints on the chromosomal location of coexpressed genes and transcription-associated proteins. Localized genomic clusters of coexpressed genes were identified and the putative function of these gene clusters was investigated.

5.2 Datasets

5.2.1 Gene expression datasets

There are eight publicly available *F. graminearum* GeneChip gene expression experiments deposited at the PLEXdb plant expression database [139], at February 2009. Experiments are listed by PLEXdb experiment identifier (FG1-FG7, FG12) in Table 5.1. All eight experiments are based on the Fusariuma520094 Affymetrix GeneChip. Two experiments, FG3 and FG4, are excluded from this study because they had been designed to test the hybridization efficiency and cross-species hybridization properties of the array: FG3 is an RNA dilution experiment to test hybridization efficiency over a range of sample dilutions, and FG4 uses samples containing RNA from three closely related Fusarium species to test for cross-hybridization of probesets by mRNA originating from other species. Five of the remaining six experiments measure changes in gene expression during various stages of the *F. graminearum* lifecycle and crop infection [138, 140, 141, 142], and one compares carbon- and nitrogen-starvation to complete media growth conditions [138] (Figure 5.1).

5.2.2 Gene calls and genome annotation

The *F. graminearum* gene complement was considered to be the 14,100 protein entries which had been used to design the Fusariuma520094 GeneChip [138]. Genetic recombination rates and chromosomal location of protein entries were obtained from the Omnimap FgraMap project [143]: genetic recombination rates had been estimated based on a cross between the sequenced strain and a field strain [144]; gene start and end positions and strand had been defined for 14,044 protein entries based on a BLAT alignment of open reading frames against chromosomal sequences, performed by John Antoniw[1] [143]. Gene Ontology (GO) annotation of the 14,100 gene entries was performed by Richard Coulson[2]. GO identifiers were assigned to *F.*

[1]Rothamsted Research
[2]Microarray Group, EMBL-EBI

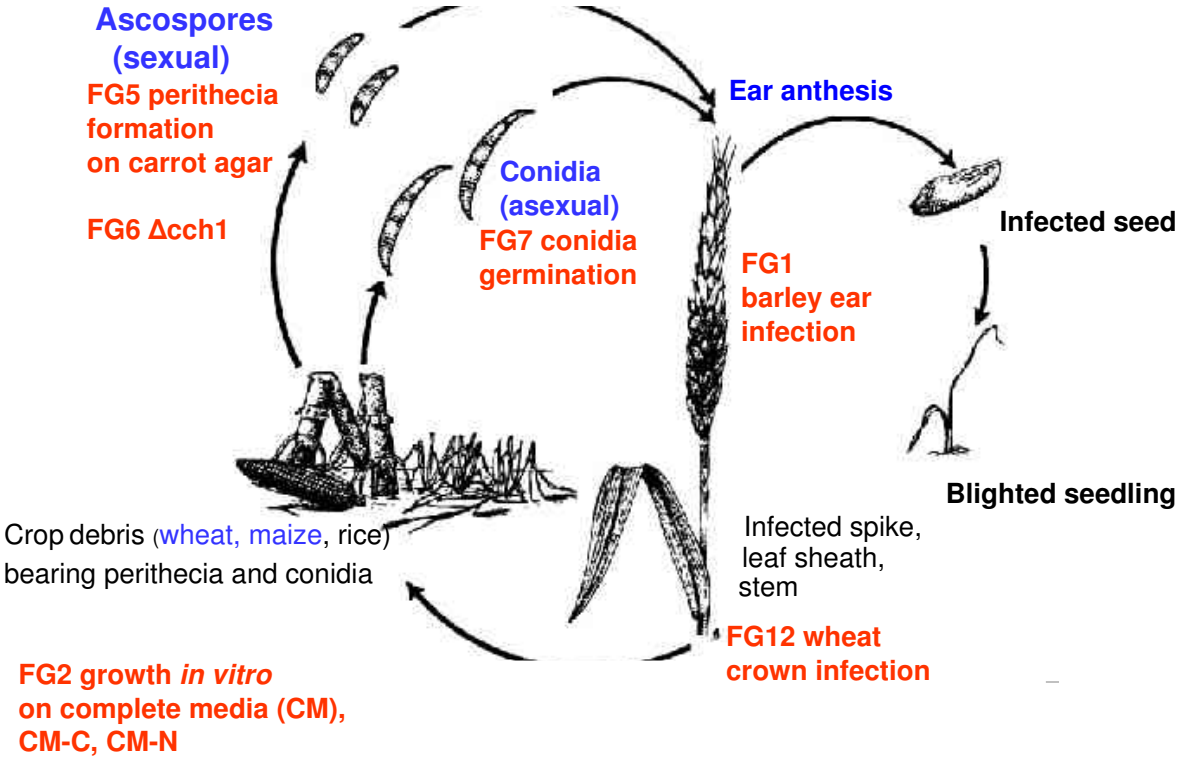


Figure 5.1: *F. graminearum* transcriptomics datasets represent stages of the *F. graminearum* lifecycle and crop infection. Transcriptomics datasets were downloaded from PLEXdb (Table 5.1). (Figure adapted with permission from Kim Hammond-Kosack, personal comm.)

Expt	Description	# hybs	Experimental factors (# replicates)	Experiment Name	Reference
FG1	Barley head infection timecourse	18	**Time (hours post infection):** 24h (3 reps) 48h (3 reps) 72h (3 reps) 96h (3 reps) 144h (3 reps) water control (3 reps)	'Fusarium transcript detection on Morex barley spikes using Fusarium Affy GeneChips'	Güldener et al. 2006 [138]
FG2	Carbon and nitrogen starvation conditions	9	**Growth conditions:** Complete media (3 reps) Carbon starvation (3 reps) Nitrogen starvation (3 reps)	'Expression Profiles in Carbon and Nitrogen Starvation Conditions'	Güldener et al. 2006 [138]
FG3/FG4	RNA dilution experiment / cross-species hybridization experiment (excluded)				Güldener et al. 2006 [138]
FG5†	*In vitro* sexual development	23	**Developmental stages*:** 0h - *vegetative hyphae* (5 reps) 24h - *dikaryotic hyphae* (3 reps) 48h - *perithecium initiation* (4 reps) 72h - *paraphysis development* (3 reps) 96h - *ascus development* (5 reps) 144h - *ascopore formation* (3 reps)	'Fusarium transcript detection during *in vitro* sexual development using Fusarium Affy GeneChips'	Hallen and Trail, 2007 [140]
FG6†	*In vitro* sexual development, Δ*cch1*	9	**Developmental stages*:** 0h - *vegetative hyphae* (3 reps) 96h - *ascus development* (3 reps) 144h - *ascopore formation* (3 reps)	'Transcript detection during *in vitro* sexual development of Fusarium Cch1 calcium channel deletion mutant using Fusariuma Affy GeneChips'	Hallen and Trail, 2008 [141]
FG7	Spore germination	12	**Time (hours, conidia development):** 0h (3 reps) 2h (3 reps) 8h (3 reps) 24h (3 reps)	'Fusarium gene expression profiles during conidia germination stages'	Seong et al. [142]
FG12	Wheat stem base infection timecourse and mycelium culture	15	**Time (days post infection):** 2 dpi (4 reps) 14 dpi (4 reps) 35 dpi (3 reps) mycelia culture (4 reps)	'Fusarium graminearum gene expression during crown rot of wheat'	Submitted by A. Stevens, J. Manners, CSIRO

Table 5.1: PLEXdb *F. graminearum* gene expression datasets used in this study. # *hybs*: number of arrays. All datasets are based on the Affymetrix Fusariuma520094 GeneChip. Raw (CEL) datasets were obtained from the PLEXdb plant expression database [139]. †FG5, FG6 are two *in vitro* sexual development studies performed by the same authors: FG5 uses a wild type, FG6 uses a Δ*cch1* mutant. *Times indicated for FG5, FG6 are approximate and are used as labels for the developmental stages shown [141].

graminearum protein entries based on the presence of protein domains with existing automated GO annotation reported in InterPro [58, 145]. 5,024 of the 14,100 protein entries were annotated with one or more GO identifiers.

5.2.3 Mapping probesets to genes

GeneChip probesets were mapped to 14,100 protein entries using a mapping provided by Ulrich Güldener[3]. Probesets which map to more than one protein entry were discarded (129 probesets discarded). Of the 14,100 protein entries, 13,830 are represented on the array by one or more probesets which are not reported to cross-hybridize to any other gene. Where more than one probeset matched a single gene and the probeset expression profiles differed, the gene was reported once for each expression profile represented amongst matching probesets. In total, 2,317 genes matched 2-5 probesets; all other genes matched exactly one probeset.

5.2.4 Gene expression data selection and preprocessing

Quality assessment and appropriate normalization of microarray datasets are crucial for the interpretation of genome-wide gene expression studies. Raw CEL files for each experiment were obtained from PLEXdb, and quality assessment and normalization steps were carried out for each experiment prior to differential expression analysis.

5.2.4.1 Quality assessment of CEL files

Raw CEL files from each experiment were quality assessed in order to check for integrity of data format, absence of spatial artifacts, and standard oligonucleotide array diagnostics. Quality assessment of CEL files was performed using the R/Bioconductor package `arrayQualityMetrics`[4] [146]. A chip description file (CDF) Bioconductor

[3]MIPS. http://mips.gsf.de/projects/fungi/Fgraminearum.html
[4]version 1.6.1; R 2.7.2

package was built for the Fusariuma520094 GeneChip[5] using the R/Bioconductor package `makecdfenv` [147]. The CDF describes the layout of features on the array, and the corresponding CDF package is required by `arrayQualityMetrics` in order to detect spatial artifacts and to identify control probesets for use in standard diagnostic tests.

An `arrayQualityMetrics` report was generated for each experiment. Several arrays were flagged as potential outlier arrays in each `arrayQualityMetrics` report based one or more diagnostic tests. However, most of the flagged arrays were judged not to be clear outliers on inspection of all diagnostic plots. After careful inspection of the generated diagnostic plots, only one experiment was deemed to contain potential outlier arrays. On inspection, three potential outliers were identified in experiment FG5 (Figure 5.2). As the experiment datasets had been prefiltered by the study authors, we required clear evidence that an array was an outlier before removing it from this study. All arrays were initially retained for normalization within each experiment. Potential outlier arrays were considered for removal only after array normalization.

Experiment FG12 contains samples from a wheat crown infection timecourse and a mycelium culture. The probeset intensity distributions of the infection timecourse arrays are consistently different from the mycelium culture arrays, with modal values at a lower intensity and more probesets in the higher intensity range than seen in the infection timecourse (Figure 5.3). Due to the difference in raw intensity distributions between the infection and mycelium culture arrays, normalization and subsequent differential expression analysis was not attempted between the infection timecourse and mycelium culture samples. Therefore only the arrays from the infection timecourse were used in further analysis.

Analysis of experiment FG7 (germination from spores) was restricted to the first two timepoints of the germination timecourse (0h, 2h). Conidia start to die within

[5]The chip description file for the Fusariuma520094 GeneChip was requested from Ulrich Güldener, MIPS

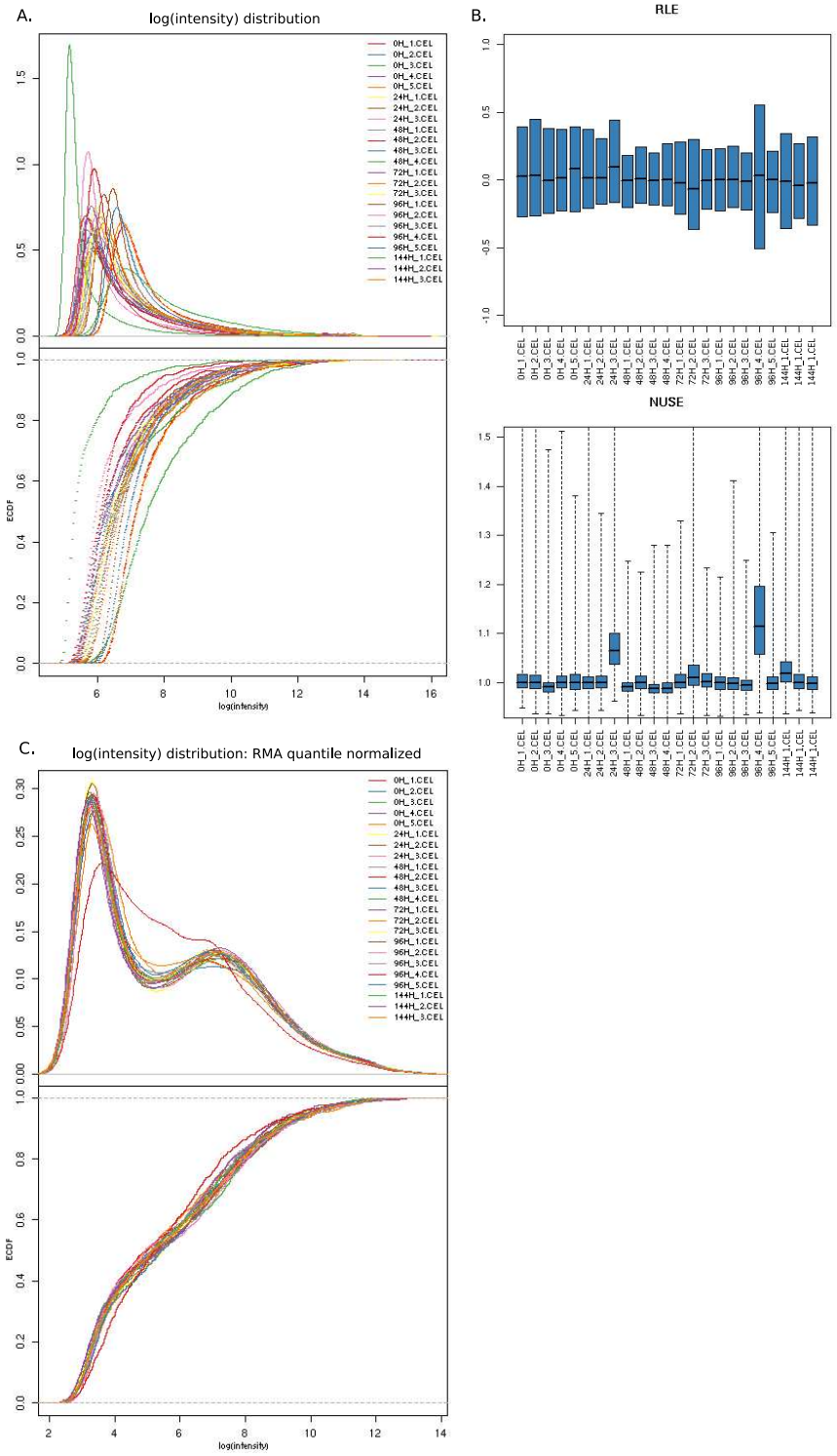


Figure 5.2: Extracts from `arrayQualityMetrics` reports for experiment FG5. **A.** Density (top) and empirical cumulative distribution (bottom) of raw intensities on each array. **B.** RLE (relative log expression; top) and NUSE (normalized unscaled standard errors; bottom) diagnostic plots. Potential outliers arrays are 96H_4, 24H_3 (RLE/NUSE) and 144H_1 (intensity distribution). **C.** Intensity distribution after RMA quantile normalization. Array 96H_4 was discarded from the study and arrays renormalized.

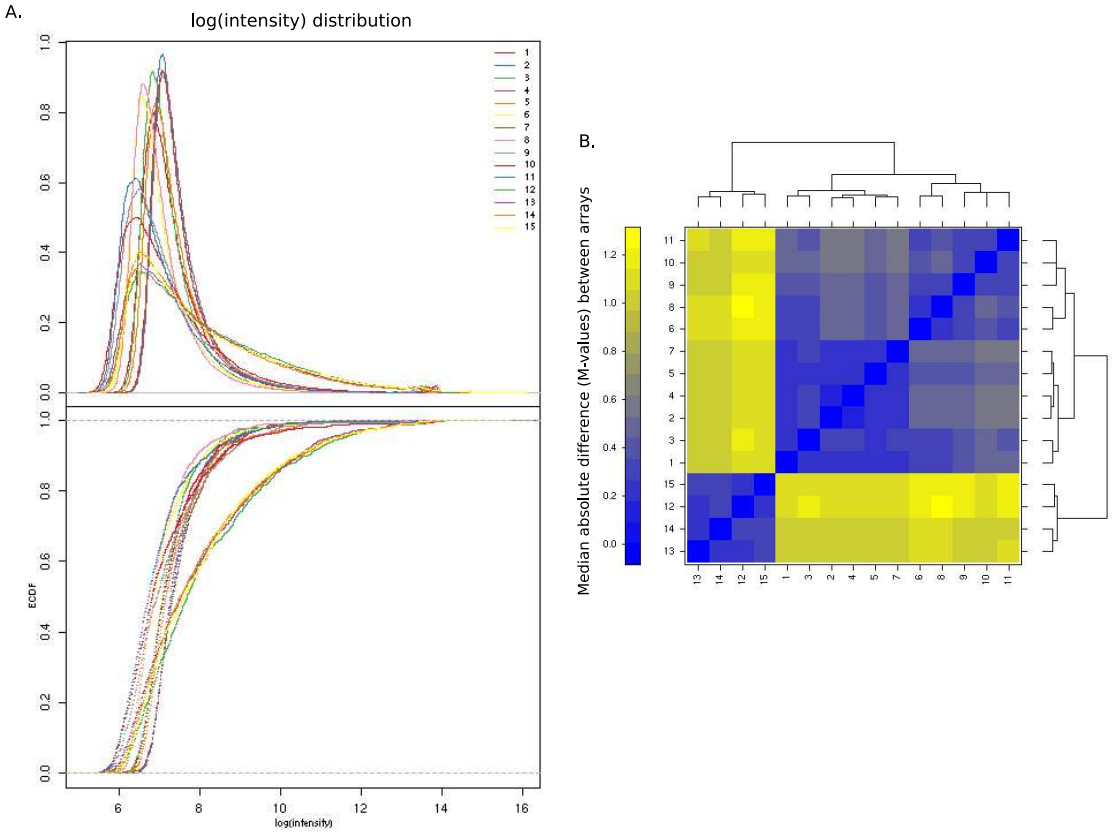


Figure 5.3: Extracts from `arrayQualityMetrics` report for experiment FG12 CEL files. **A.** Density (top) and empirical cumulative distribution (bottom) of intensities on each array. **B.** between-array distances measured as median absolute difference in *M*-values between arrays. Arrays 1-11 are wheat crown infection samples. Arrays 12-15 are mycelium culture samples. The mycelium culture arrays display a distinct intensity distribution from that of the wheat infection arrays.

hours of germination on plates, and the spore germination timecourse in experiment FG7 is therefore dominated by the rapid loss of conidia (Kim Hammond-Kosack, personal communication).

5.2.4.2 Array normalization for differential expression analysis

It has been previously noted that due to a large degree of technical (non-biological) variation between arrays from different datasets that had been generated in different labs at different times, array normalisation and differential expression analysis should first be performed within each individual set, and a comparison of differentially expressed gene sets needs to be determined by meta-analysis [148].

Arrays were normalized within each PLEXdb experiment prior to differential expression analysis. Arrays for experiments FG2, FG5, FG6, FG7 (0h / 2h), and FG12 (infection) were processed by robust multiarray analysis (RMA) using the R/Bioconductor package `RMA` [149]: raw probe intensity values were background-corrected, normalized across arrays by quantile normalization, $\log_2$ transformed, and summarized by the median polish algorithm to produce normalized probeset intensity values. (See Chapter 1 for an overview of preprocessing methods for oligonucleotide arrays.)

It was necessary to consider alternative normalization procedures for experiment FG1. In experiment FG1 (infection of barley heads), a global upward trend in probeset intensities was observed across the timecourse under quantile normalization. This trend was also noted by the authors of the original study [138] and is likely to be caused by the relative dilution of *F. graminearum* RNA in samples taken at early timepoints compared to later timepoints. At later timepoints there is a larger biomass of *F. graminearum* and therefore a higher concentration of *F. graminearum* RNA in the sample. The effect of *F. graminearum* dilution at early timepoints is to overestimate the number of genes which are differentially expressed between early and late timepoints. Arrays in experiment FG1 were therefore renormalized using an alternative method in place of quantile normalization, with the aim of removing the global

upwards trend in probeset intensities along the timecourse. The RNA polymerase subunits were selected as an invariant subset of genes: that is, a set of genes which *a priori* are not expected to be transcriptionally regulated during barley infection. The data were normalized using a variance-stabilizing normalization (VSN) [36] with the parameters of the transformation estimated using the invarient set. The invariant gene set is represented on the array by 27 probesets (Table B.1, Appendix A). A larger invariant set would permit a more robust estimate of the VSN normalization function, but there is considerable uncertainty about which *F. graminearum* genes are expected to maintain a constant mRNA level during the infection timecourse.

It is possible that the wheat infection timecourse in experiment FG12 is similarly affected by an RNA dilution effect at early timepoints compared to later timepoints. However, far fewer probesets were detected at each FG12 infection timepoint than in any other experiment (Table 5.4(a) below). Under quantile normalization, almost all (19 / 24) of the differentially expressed probesets are upregulated late in the timecourse, consistent with relative dilution of *F. graminearum* RNA at early timepoints. Only 4 of the 27 RNA polymerase subunit probesets were detected (*MAS 5.0*) during experiment FG12 and therefore a normalization based on the same invariant set is likely to be misleading. Due to the low rate of detection on the FG12 arrays, the choice of normalization does not have a large impact on conclusions drawn from the FG12 infection timecourse in this study. Quantile normalization was therefore used for FG12, with the caveat that the observed upregulation of 19 genes during wheat infection may be a consequence of early sample dilution rather than transcriptional regulation.

5.3 Methods

5.3.1 Overview

Probeset detection calls were assigned for each array and differentially expressed genes were identified within each experiment. A set of predicted transcription-

associated proteins was identified, including DNA-binding transcription factors. The genomic location of classes of transcription-associated proteins and the genomic clustering of coexpressed genes was assessed. Localized genomic clusters of coexpressed genes were identified, and the coexpressed genes lying in these clusters were annotated with additional predicted protein function.

5.3.2 Probeset detection

Probeset detection calls (*present*, *marginal*, *absent*) were assigned using the *MAS 5.0* algorithm [40] implemented in the function `mas5calls` of the R/Bioconductor package `affy` [147]. This step is complementary to the subsequent differential expression analysis: probesets which are not called as differentially expressed within an experiment may be constitutively expressed (*i.e.* expressed at a similar level in all conditions), or might not be expressed or detected in any condition. Within each experiment, each probeset was classified according to whether it is (i) never detected, or (ii) detected in at least one condition. Probeset detection calls from replicate arrays were combined into a single *detected / not detected* call for each condition. The *present / marginal / absent* flag for each replicate was scored as 1 / 0.5 / 0 respectively. A probeset was called as *detected* in a condition if the mean score of replicates was more than 0.6. This replicate scoring scheme is consistent with the schemes used by Hallen *et al.* [140, 141] in the FG5 and FG6 study publications and by Güldener *et al.* [138] in the FG2 study publication to call probesets detected in each condition.

5.3.3 Defining groups of coexpressed genes

Within each experiment, a discrete expression profile was assigned to each probeset by fitting a linear model to the observed expression levels and testing for differential expression between conditions. The R/Bioconductor package `LIMMA` [150] was used to fit a linear model to normalized data and to identify differentially expressed probesets within each experiment based on selected contrasts. For each timecourse

experiment (FG1, FG5, FG6, FG7, FG12 wheat infection), contrasts were chosen such that the expression level at the first timepoint was compared to the expression level at each subsequent timepoint. Timepoints represent distinct stages of development or infection and have low time resolution compared to the dynamics of induction or repression. For experiment FG7, we report a comparison only between spores (0h) and the first timepoint (2h of development from spores). For experiment FG2, carbon-minimal medium and nitrogen-minimal medium growth conditions were each compared to complete media growth conditions.

For each fitted linear model, an F statistic was calculated for each probeset and p-values were derived using an empirical Bayes approach [151] implemented in `LIMMA`. Probesets were ranked by p-value and a probeset was called as *differentially expressed* if the p-value was below a threshold value. The threshold was defined as the minimum p-value amongst all *AFFX* control probesets on the array: $p\text{-value} \leq \min\{p\text{-values of } AFFX \text{ control probesets}\}$. This method acknowledges that the assumption of an F-distribution null model for non-differentiated probesets may not be an accurate assumption and instead treats the ranking of probesets by p-value as a ranking of confidence in differential expression. For these datasets, a threshold chosen to exclude *AFFX* control probesets from being called as differentially expressed typically identifies fewer differentially expressed probesets than are found by comparing each probeset to an F-distribution null model.

Amongst probesets which were detected and differentially expressed, differential expression patterns were defined using the following contrasts:

- FG1: (48h, 72h, 96h, 144h) vs. 24h
- FG2: (MMC, MMN) vs. CM
- FG5: (24h, 48h, 72h, 96h, 144h) vs. 0h
- FG6: (96h, 144h) vs. 0h
- FG7: spores vs. 2h

- FG12 wheat infection: (14dpi, 35dpi) vs. 2dpi

Symbols 1 / -1 / 0 denote differential expression up / down / no differential expression compared to the reference condition, respectively. There are many possible patterns of differential expression for timecourse experiments with more than three timepoints. Differential expression patterns were summarized into the following *coexpression groups*:

- upregulated and stays upregulated,
- downregulated and stays downregulated,
- transiently upregulated,
- transiently downregulated,
- any other differential expression behaviour.

The majority of differentially expressed probesets within each experiment could be described using the first four coexpression groups. A complete listing of differential expression patterns making up each coexpression group is given in the Appendix.

Probesets were mapped to genes as described above. If more than one probeset mapped to a single gene, the gene was permitted to lie in one or more coexpression groups, but note that almost all genes lie in exactly one coexpression group. A small number of probesets were called as differentially expressed but not detected by *MAS 5.0* (described above); these probesets were classified as 'not detected' and were not included in any of the coexpression groups.

5.3.4 Prediction of transcription-associated proteins (TAPs)

Putative transcription-associated proteins (TAPs) in the *F. graminearum* genome were identified by Richard Coulson[6] using a combination of homology searches. To iden-

[6]Microarray Group, EBI

tify genes involved in transcription and transcriptional regulation, *F. graminearum* protein entries were queried against a reference set of TAPs and protein domains. The method has previously been described in detail for *Plasmodium falciparum*[152] and was used as the basis for comparative genomics studies of proteins involved in transcription in evolutionarily diverse species [153, 154]. Briefly, the sequences of all 14,100 *F. graminearum* protein entries were queried for (i) protein domain homology using PFAM Hidden Markov Models (HMMs) related to transcriptional control in eukaryotes, and (ii) sequence homology to a TAP reference set (Blastp) followed by sequence similarity clustering to identify families of genes with a high degree of sequence similarity. The results of the HMM protein domain and sequence similarity searches were combined to produce a conservative set of putative functional TAPs in *F. graminearum* .

723 of the 14,100 predicted *F. graminearum* protein entries were found to be homologous to a member of the TAP reference set. The classification and prevalance of predicted transcription-associated factors are summarized in Table 5.2. In addition, the clade specificity of putative TAPs was defined by identifying putative functional homologues of the *F. graminearum* TAPs amongst eukaryotic genomes. Putative functional homologues of the *F. graminearum* TAPs were identified by querying 54 eukaryotic genomes using strict similarity criteria of sequence identity (Blastp) and presence of protein domains (PFAM HMM protein domain search). The clade specificity of each TAP was assigned as 'Fusarium', 'Pezizomycotina', 'Fungi', or 'Eukaryotes' according to whether a putative functional homologue is present in the respective clade.

5.3.5 Functional annotation and clade specificity of selected genes

The Broad Institute and MIPS provide manually curated and automated putative functional annotation for predicted *F. graminearum* genes based on sequence homology searches [137]. *F. graminearum* functional annotation is archived by the Broad Institute for the most recent (*FG3*) gene calls. This annotation includes manually

Table 5.2: Categories of *F. graminearum* genes with homology to a reference set of transcription-associated proteins (TAPs) and HMM protein domains.

Category	Description	Count
B	Basal transcription factors and cofactors	63
C	Chromatin remodelling and histone modification	63
D	DNA-binding proteins	546
P	RNA polymerase subunits	27
O	Unclassified† *(CCR-NOT subunits, non-DNA-binding factors)*	24
	Total number of predicted *F. graminearum* TAPs	**723**

† 'Unclassified' genes are homologues of transcription-associated proteins which do not fall into the other categories: more than half of these are homologues of CCR4-NOT complex subunits [155], and the rest are homologues of other transcriptional regulators which are not DNA-binding.

validated gene names, and gene names transferred from other annotated genomes due to strong protein sequence homology.

During this study, a set of 170 genes was identified which were present in coexpressed genomic clusters. To investigate whether the presence of coexpressed gene clusters is related to protein function, a more complete coverage of putative gene function was required for the genes in this set. A comparative analysis of putative protein function was therefore carried out for this gene set. The 170 genes in this set are non-TAP genes and are referred to here as the *query set*.

Putative functional annotation was transferred from eukaryotic protein-coding genes with protein sequence homology to a gene in the query set. Homologues were identified using a Blastp search (`blastp`, NCBI toolkit [62]; E = 10^{-6}; -F F). Low complexity regions of query genes were masked using CAST [156] before performing Blastp searches. The resulting hits were combined with the original query set and a second round of BlastP was run (E = 10^{-6}; -F F) in order to identify further protein sequence homologues. All proteins thus identified as sequence homologues to the query set were partitioned into clusters of homologous proteins using Markov clustering of E values (`mcl` [157]; inflation = 2) [63, 158]. Of the 170 *F. graminearum* proteins in the query set, 139 were assigned to one of 122 protein clusters containing one or more homologous proteins, with the remaining 31 *F. graminearum* proteins hav-

ing no identified homologues (and therefore appearing as singletons in the Markov clustering).

Protein clusters were annotated as follows. For each protein cluster member, the gene name (if any) and Gene Ontology (GO) assignments (if any) annotated in UniProtKB (Uniprot Knowledgebase [61]) were transferred to the protein cluster. The clade of each protein cluster member was determined according to the current NCBI taxonomic description [159] of the respective genome. A short *description* of putative protein function was assigned to each protein cluster by inspection of UniProt gene names, functional annotation, and GO annotation in the cluster. The clade specificity of each cluster was assigned to one of 'Fusarium', 'Pezizomycotina', 'Fungi', 'non-metazoan Eukaryotes', 'Eukaryotes' according to the presence in the cluster of proteins from each clade. The 31 *F. graminearum* proteins with no identified homologues were queried against the UniProtKB database using the UniProt BlastP webservice (E = 10^{-6}) [160] to identify any additional functional information from all functionally annotated organisms.

5.3.6 Testing for chromosomal clustering of coexpressed genes

Genomic clustering of coexpressed genes

Coexpressed genes were tested for evidence of clustering on the genome. For the purposes of this test, *coexpressed genes* are defined as genes which have the same expression pattern within one experiment. Each GeneChip experiment is considered separately, and expression patterns are defined using the combined differential expression patterns described above. The background gene list was taken to be the 13,830 genes, out of the 14,100 *F. graminearum* protein entries, which are mapped onto the array. Genes were ordered according to the genomic ordering of the first base of each gene.

A measure of the degree of chromosomal clustering amongst a set of genes was defined: for a gene list L containing c coexpressed genes, the number of genes N_g

which lie within g genes of a coexpressed gene was counted (Figure 5.4). If N_g (the number of genes with a proximity of g genes or fewer to a coexpressed gene) is significantly higher than expected under the null hypothesis, this is evidence that coexpressed genes exhibit a higher degree of chromosomal clustering than expected by chance. The null hypothesis was that coexpressed genes are drawn uniformly from all annotated genes on the genome, where N_g is sampled from a null distribution n_g. The null distribution of n_g for $1 < g < g_{max}$ was simulated using 1000 samples of c genes, where genes were sampled uniformly from the background gene list without replacement. As the 13,830 genes are distributed on only four chromosomes, chromosome edge effects were ignored and chromosomes 1, 2, 3, 4 were concatenated for both the test case and for resampling. An empirical p-value was defined as the proportion of simulated samples N_g^{sim} for which $N_g \geq n_g$ and used to determine the significance of observing N_g. Within each gene list L, p-values were corrected for multiple testing of proximities $g = 1..200$ using the Benjamini-Yekutieli correction [161].

For each gene list L, a *Z-score* was also calculated for each N_g. The *Z-score* is defined as

$$Z = \frac{N_g - \mu_g}{\sigma_g} \tag{5.1}$$

where μ_g, σ_g are the mean and standard deviation of the empirical null distribution, respectively. The Z-score allows us to use a rapid approximate method to determine whether specific localized clusters identified by other methods can account for observed significant genome-wide clustering. If a genes have been identifed to be part of a localized gene cluster, then using a Z-score of 3 to determine significant genome-wide clustering, the number of genes N_g with proximity g to a coexpressed gene must satisfy $N_g - a > \mu + 3\sigma$ in order for the genome-wide clustering to still be considered significant after discounting the effect of the known gene cluster.

TAP-centric clustering of coexpressed genes

The 723 TAPs were tested for the presence of significant clusters of coexpressed genes within a neighbourhood around each TAP. For the purposes of this test, the *coexpressed genes* for a given TAP are all the genes coexpressed with the TAP, in any experiment. For a gene to be coexpressed with the TAP, it is sufficient for the gene to have the same differential expression pattern as the gene in any one of the GeneChip experiments.

Within a chromosomal region defined by t adjacent genes in a chromosome of length G, we consider the distribution of the number of coexpressed genes, k, present in the region from a list L of c coexpressed genes. The null distribution of k is the hypergeometric distribution, so that the probability of finding exactly k genes from list L of size c in a region of size t is $p(k;\ G, t, c) = \binom{t}{k}\binom{g-t}{c-k} / \binom{G}{c}$. For each TAP, windows of size 2 to 40 genes centred on the TAP were tested for enrichment of genes coexpressed with the TAP using Fisher's exact test with $\alpha = 0.05$. All p-values were corrected for multiple testing using the Benjamini-Hochberg correction [57].

Localized clustering of coexpressed genes

Similar to the test for TAP-centric clustering, the number k of genes from list L of size c which are found in a region of size t follows the hypergeometric distribution under the null distribution of uniform sampling from the genome. In order to identify a conservative list of chromosomal regions which are significantly enriched for coexpressed genes, all possible regions of the genome were tested for enrichment for coexpressed genes and an empirical correction for multiple testing was applied.

The Positional Gene Enrichment (PGE) tool [162] was used to detect regions of the *F. graminearum* genome which are enriched for genes in each coexpression group[7].

[7]Source code for the PGE tool and for applying the *min* p_i multiple testing correction was provided by Roland Barriot (KU Leuven, Bioinformatics Research Group). I adapted the code for use with the *F. graminearum* genome and for the code to be run on a local machine cluster using the *minPi* multiple testing correction.

PGE aims to test almost all possible chromosomal regions, from pairs of neighbouring genes to whole chromosomes, for overrepresentation of a given gene list. As the number of regions of any size located at any position in the genome is massive, the emphasis is on identifying localised clusters of genes. Large regions are discarded if they contain smaller, more significant regions. To reduce the number of possible regions to test, PGE tests only *pertinent regions*, defined by Rules 1-6 according to De Preter and colleagues [162]. Rules 1-4 avoid the testing of redundant regions, and Rules 5-6 cause large regions to be discarded if they contain smaller, more significant regions. According to De Preter and co-authors, a region is *pertinent* if:

Rule 1: it contains at least two genes of interest,

Rule 2: there is no smaller region containing the same number of genes of interest,

Rule 3: there is no bigger region with more genes of interest and the same number of genes not of interest,

Rule 4: there is no larger encompassing region with a higher percentage of genes of interest,

Rule 5: there is no smaller encompassed region with a better P-value,

Rule 6: it does not contain any region having fewer than expected genes of interest.

A large number of chromosomal regions are tested for enrichment of a given gene list of coexpressed genes, so all p-values must be corrected for multiple testing using an appropriate correction method. A *min* p_i correction was used to correct for multiple tests. For each coexpression list, the set of coexpressed genes was resampled 10,000 times from the background gene list uniformly without replacement. For each random gene list, we find the minimum p-value amongst all pertinent regions. A corrected p-value of 0.05 means that 5% of random lists of genes of the same length contained a chromosomal region with a smaller p-value for enrichment of genes in a gene list of the same length.

5.4 Results

5.4.1 Overview

Differentially expressed genes and coexpression groups are summarized for each PLEXdb *F. graminearum* experiment. In particular, the predicted DNA-binding transcription factors identified within each coexpression group are listed. There is evidence of constraints on the chromosomal position of transcription-associated proteins (TAPs) and genomic clustering of coexpressed genes. In several coexpression groups, significantly more genes lie adjacent to or within 2-10 genes of a coexpressed gene than expected by chance under uniform selection from all genes on the array. Moreover, 8 TAP-centric genomic regions were found to be enriched for coexpressed genes and centred on a coexpressed TAP, and 18 genomic regions were found to be significantly dense in coexpressed genes. Combining all TAP-centric and non-TAP-centric coexpressed gene clusters, 20 distinct genomic clusters were identified containing 2 – 22 coexpressed genes. Coexpressed genes in all TAP-centric and non-TAP-centric gene clusters were annotated with predicted protein function. A systematic analysis of predicted protein function based on Gene Ontology annotation did not identify differential function between TAP-centric and non-TAP-centric coexpressed gene clusters. A bias in protein domain composition was identified amongst detected and differentially expressed TAPs, in which classes of protein domain were under- or overrepresented in some of the experiments.

5.4.2 Differential expression within experiments

There is much variation in the number of genes and TAPs detected and differentially expressed during each experiment (Figure 5.5(a, b)) and a high degree of specificity of differentially expressed genes and TAPs to each experiment (Figure 5.5(c)). For this summary, experiments FG5 and FG6 are combined into a single set of genes which are differentially expressed during either of the two sexual develop-

ment timecourses (FG5/6). Experiment FG7 (development from spores) is excluded due to uncertainty about the interpretation of differential expression calls in FG7 (germination from spores) (see Datasets). Most of the differentially expressed genes are specifically differentially expressed during one gene expression response: 80% (3603/4479) of genes differentially expressed during experiments FG1, FG2, FG5/6, and FG12 are differentially expressed in exactly one experiment. Similarly, 86% (134/155) of differentially expressed DNA-binding TAPs are specific to FG1, FG2 or FG5/6. A comparison of growth media (FG2) revealed that the two largest coexpression groups contain genes which are specifically up- or down-regulated in either carbon- or nitrogen- starvation conditions compared to complete media (Table 5.3, below). There is a high degree of overlap, however, between genes differentially expressed during the two crop infection timecourses (FG1 and FG12). Despite the low probeset detection rate in the FG12 wheat crown infection timecourse, almost all genes differentially expressed in FG12 are also differentially expressed in the FG1 barley infection timecourse (FG1: 807, FG12: 22, overlap: 18, n = 13830; $\chi^2 = 231$, $p < 0.0005$). These genes may therefore form part of a core gene expression response involved in the progression of crop infection (Table B.3) [138], although the size of such a core infection response may be significantly underestimated here due to the low detection rate in FG12.

Amongst genes differentially expressed within each experiment, groups of coexpressed genes were defined (Table 5.3). The DNA-binding TAPs in each coexpression group are listed here in full (Table B.2). Our results describe the expression patterns of putative *F. graminearum* DNA-binding transcription factors across diverse developmental and environmental conditions, and provide a basis for further work on the transcriptional regulation of *F. graminearum* gene expression programmes.

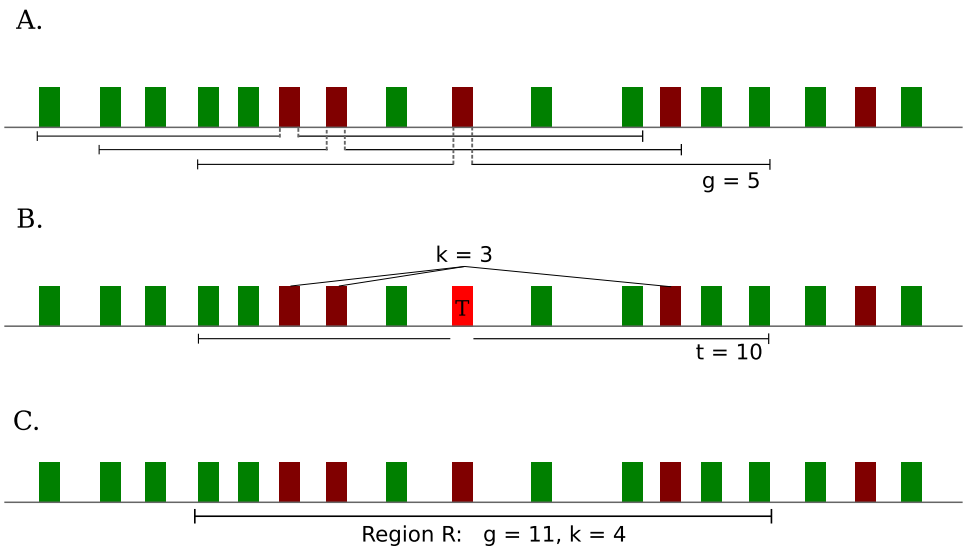


Figure 5.4: Testing for chromosomal clustering of coexpressed genes. **A. Genomic clustering of coexpressed genes:** for a gene list of L genes (red), the number N_g of genes within g genes of a coexpressed gene is counted. **B. TAP-centric clusters of coexpressed genes:** for each TAP (T), the number of genes k coexpressed with the TAP and lying within a TAP-centred window spanning t adjacent genes is counted. **C. Localized clustering of coexpressed genes:** for each genomic region R containing g genes, the number of coexpressed genes k is counted. See text for details of each test.

(a) Number of detected and differentially expressed (DE) genes within each experiment

	Detected (MAS5)		**Detected (MAS5) Not DE (limma)**		**Detected (MAS5) DE (limma)**	
	# genes	# TAPs	# genes	# TAPs	# genes	# TAPs
FG1	6795	348	5988	333	807	15
FG2	10055	658	8305	570	1750	88
FG5	11475	669	9605	614	1870	55
FG6	12990	678	11287	620	1703	58
FG12 (infection timecourse)	3096	106	3074	105	22	1
FG7 (0h & 2h arrays only)	8496	589	6041	480	2455	109

(b) Number of detected / detected and differentially expressed TAPs, displayed by TAP class. *B: basal transcription factors and cofactors; C: chromatin remodelling and histone modification; D: DNA-binding proteins; P: RNA polymerase subunits; O: unclassified.*

# TAPs detected / detected & DE	**B**	**C**	**D**	**O**	**P**
on array	**61**	**63**	**536**	**23**	**27**
FG1	44 / 2	42 / 2	227 / 11	16 / 0	19 / 0
FG2	59 / 0	61 / 3	490 / 85	22 / 0	26 / 0
FG5	60 / 6	62 / 0	497 / 47	23 / 1	27 / 1
FG6	59 / 6	62 / 5	508 / 46	23 / 1	26 / 0
FG12 (infection timecourse)	13 / 0	12 / 1	72 / 0	5 / 0	4 / 0
FG7 (0h & 2h arrays only)	59 / 9	60 / 8	424 / 76	20 / 4	26 / 12

(c) Summary of all genes (top) and DNA-binding TAPs (bottom) differentially expressed in one or more experiments

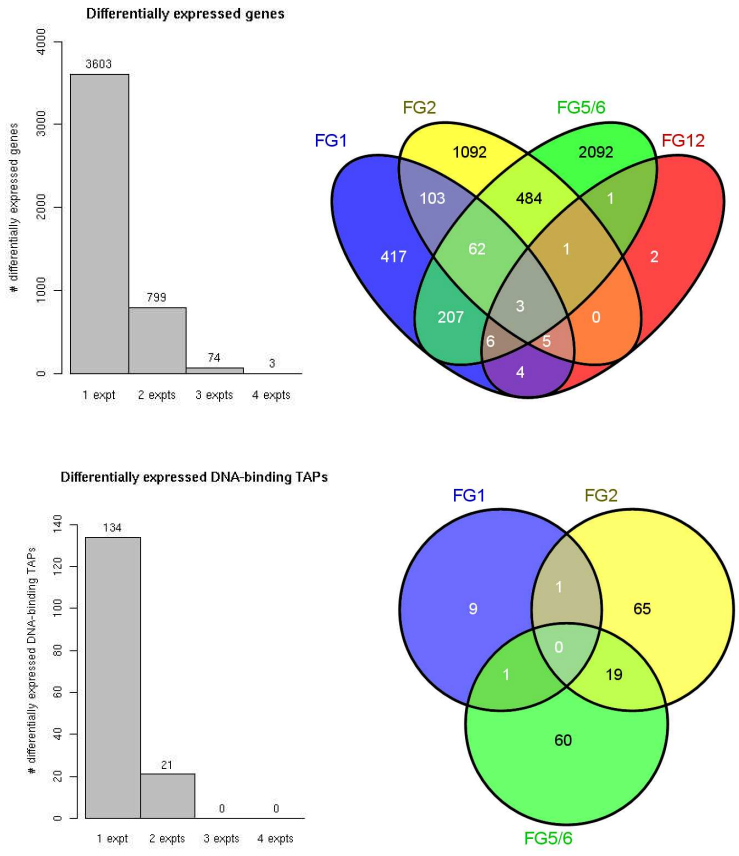


Figure 5.5: Summary of genes and DNA-binding TAPs which are differentially expressed in one or more experiments. [Venn diagrams were drawn using the Venny tool [163]]

Table 5.3: Definition of coexpression groups and symbols. Non-empty coexpression groups are listed. †See Appendix A for a complete list of limma patterns defining each coexpression group.

Experiment	Coexpression group†	Symbol	# genes	# TAPs	# DNA-binding TAPs
FG1	Downregulated and stays downregulated during infection	FG1↓	4	-	-
	Upregulated and stays upregulated during infection	FG1↑	781	14	10
	Transiently upregulated	FG1↑↓	13	-	-
	Other behaviour	FG1~	10	1	1
FG12 infection	Downregulated early and stays downregulated during infection	FG12.-1-1	1	-	-
	Transiently downregulated	FG12.-11	1	-	-
	Downregulated late	FG12.0-1	1	-	-
	Upregulated late	FG12.01	19	1	-
FG2	Downregulated in carbon-minimal media (CMM) and in nitrogen-minimal media (NMM) compared to complete media	FG2.-1-1	206	11	11
	Downregulated in CMM but not in NMM	FG2.-10	321	7	6
	Down-regulated in CMM; upregulated in NMM	FG2.-11	69	1	1
	Downregulated in NMM but not in CMM	FG2.0-1	131	5	5
	Upregulated in NMM but not in CMM	FG2.01	348	35	35
	Upregulated in CMM; downregulated in NMM	FG2.1-1	23	1	1
	Upregulated in CMM but not in NMM	FG2.10	484	13	11
	Upregulated in CMM and in NMM	FG2.11	178	15	15
FG5	Downregulated and stays downregulated during sexual development	FG5↓	426	8	8
	Upregulated and stays upregulated	FG5↑	806	23	18
	Transiently downregulated	FG5↓↑	247	1	1
	Transiently upregulated	FG5↑↓	316	17	16
	Other behaviour	FG5~	101	6	4
FG6	Downregulated and stays downregulated during sexual development/Δ cch1	FG6↓	781	28	26
	Upregulated and stays upregulated	FG6↑	668	24	15
	Transiently downregulated	FG6.-11	38	1	1
	Transiently upregulated	FG6.1-1	1	-	-
	Transiently downregulated	FG6↓↑	126	1	1
	Transiently upregulated	FG6↑↓	98	5	4
FG7 (2 hrs vs. spores)	Down in spores compared to 2 hrs	FG7_2h_spores.-1	1222	40	16
	Up in spores compared to 2 hrs	FG7_2h_spores.1	1234	69	60

5.4.3 Positional constraints on TAPs and coexpressed genes

Current evidence points to several levels of gene regulation which affect gene order in eukaryotic genomes, including chromosomal clustering of coexpressed genes and transcription factor target genes [164, 165], chromatin remodelling which acts to promote the transcriptional regulation of genes lying within open chromatin regions [166], and the spatial location of chromosomal regions within the nucleus [167]. Gene location and gene order also need to be considered in light of meiotic recombination, which is a major source of heterogeneity across the eukaryotic genome. The *F. graminearum* genetic map is unusual in that long regions of high or low recombination rates have been observed spanning several hundred kilobases, whereas regions of high or low recombination rate (so-called recombination *hotspots* and *coldspots*) are more commonly found to span only a few kilobases in other eukaryotes [144]. The mechanisms and consequences of recombination for the evolution of gene location and gene order are not yet well understood (see [168], for example, for a recent high-resolution study of recombination in *S. cerevisiae*) but it seems likely that further classes of genes may be associated with differential recombination rates.

There is growing evidence that transcriptionally coregulated genes are not randomly distributed across the genome. Using *S. cerevisiae* transcription factor ChIP-chip datasets to identify *in vivo* binding to the promotor regions of target genes, Janga *et al.* [165] found that the target genes of most transcription factors have a higher degree of proximity than is seen for whole genome permutations of the transcription factor-targets network. More limited studies have similarly reported non-random distribution of putatively coregulated genes across the *S. cereviciae* genome [169, 170].

Genomic clusters of adjacent coexpressed genes have been observed in many eukaryotic genomes, ranging from frequent pairs and triplets of adjacent coexpressed genes in human and *Mus musculus* to stretches of up to 30 adjacent or near-adjacent genes in *Drosophila* (*e.g.* [171, 172, 173]; reviewed in [164]). In the *F. graminearum* genome

there are well-characterized examples of localized clusters of coexpressed genes. Mutant knock-outs of many genes involved in secondary metabolite biosynthesis cause reduced pathogenicity in *F. graminearum* [174] and some of these genes lie in highly localized genomic clusters of coexpressed genes involved in the biosynthesis of mycotoxins and other secondary metabolites. The aurofusarin biosynthesis gene cluster contains 12 genes involved in the aurofusarin biosynthesis pathway and lying within a 30kb genomic region [175]. The gene cluster contains a transcriptional activator *aurR1/GIP2* (*fg02320*; DNA-binding) which is required for aurofusarin production [176, 177]. The cluster includes the aurofusarin polyketide synthase *PKS12* (*fg12040*) which is one of 15 type I polyketide synthases in *F. graminearum* [178]. Gene clusters of 4 – 9 adjacent genes located within 17kb genomic regions are involved in the biosynthesis of the polyketide mycotoxins zearalenone and butenolide, respectively [179, 180]. It is not yet clear whether functionally related clusters of coexpressed genes occur frequently in the *F. graminearum* genome, whether such gene clusters are typically transcriptionally regulated by adjacent genes, and whether such clusters are primarily involved in the biosynthesis of mycotoxins and other secondary metabolites or are functionally diverse.

Are coexpressed genes and gene encoding transcription-associated proteins randomly distributed on the *F. graminearum* genome, or is there evidence of widespread structured genomic organisation? I investigated the genomic organization of (i) genes encoding transcription-associated proteins (TAPs), and (ii) all genes which are coexpressed within each gene expression dataset. TAP genes were tested for evidence of non-random genomic location. Coexpressed genes were investigated for genomic clustering using three methods. First, the degree of genomic clustering of coexpressed genes was tested across the genome. Second, each differentially expressed TAP was tested for the presence of coexpressed genes in a surrounding region. Third, the presence of localized clusters of coexpressed genes was assessed, regardless of whether a coexpressed TAP lies within the gene cluster.

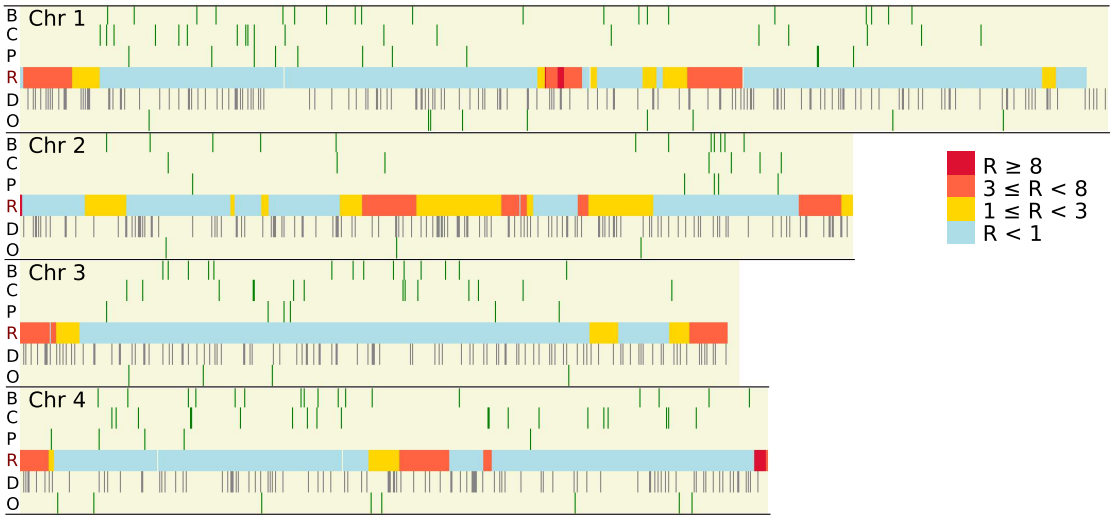
5.4.3.1 Constrained chromosomal distribution of TAPs

The chromosomal location of genes coding for transcription-associated proteins are subject to positional constraints (Figure 5.6). The recombination rate of chromosomal regions appears to act as a constraint on the presence of some classes of TAPs. Basal transcription factors and cofactors, RNA polymerase subunits, and chromatin remodelling / histone modification factors tend not to be located in chromosomal regions of high recombination rate. In contrast, DNA-binding transcription factors and 'unclassified' (CCR4-NOT subunits and other non-DNA-binding transcription-associated factors) are not subject to the same constraint and are not over- or under-represented in regions of high or low recombination rate.

5.4.3.2 Evidence of chromosomal clustering of coexpressed genes

Each coexpression group was tested for chromosomal clustering on the genome. The degree of chromosomal clustering of each group of coexpressed genes was summarised by counting the number of genes in the group which lie within g genes of a coexpressed gene. The proximity g was varied from 1 to 200 genes. For 11 of the 25 coexpression groups, there is some evidence of a higher degree of clustering than would be expected by chance from a gene list drawn uniformly from all 13,830 genes on the array. The coexpression groups with significant chromosomal clustering are shown in Table 5.4. Where there is significant clustering with proximity of $g \geq 5$, the significance of the clustering may be accounted for if a small number of localized coexpressed clusters is identified. For $g > 4$, it is typically sufficient for approximately 20 genes to lie in localized clusters on the genome in order to explain the significance of the genome-wide clustering: the clustering of the remaining coexpressed genes would not deviate significantly from the clustering expected under the null hypothesis of uniform sampling. There are significantly many more genes lying adjacent to a coexpressed gene ($g = 1$) than expected from a uniform null distribution. For example, coexpression group FG2.10 (genes upregulated in

A



B

		$R \geq 8$	$3 \leq R < 8$	$1 \leq R < 3$	$R < 1$	X^2; p value	
B	*Obs*	0	0	5	58	15.54; $p = 0.007$	*
	Exp	0	9	9	44		
C	*Obs*	0	3	1	59	16.56; $p = 0.005$	*
	Exp	0	9	9	44		
P	*Obs*	0	0	1	26	8.80; $p = 0.04$	*
	Exp	0	4	4	19		
D	*Obs*	3	85	81	370	1.38; $p = 0.7$	
	Exp	4	77	79	380		
O	*Obs*	0	2	4	18	0.91; $p = 0.8$	
	Exp	0	3	4	17		

Figure 5.6: **A.** The *F. graminearum* genome is divided into four blocks according to the recombination rate, *R* (cM/27kb). The genomic locations of TAP classes *B, C, P,* (above) and *D, O* (below) are shown (for TAP class descriptions see Table 5.2). **B.** Observed (*Obs*) and expected (*Exp*) gene counts are shown for each TAP class (*B, C, P, D, O*) in each recombination block under a null model of uniform distribution over *F. graminearum* gene positions. *P*-values result from a χ^2 test for each TAP class (* : $p < 0.05$).

Table 5.4: Testing for genomic clustering of coexpressed genes. The 9 groups of coexpressed genes which show significant chromosomal clustering are shown (see Methods). ***p*-value:** the proportion of simulated gene lists with the same or greater number of genes within the indicated gene proximity of a coexpressed gene, after correction for multiple testing. ***Z*:** the observed *Z-score*. ***obs*:** the observed number of genes within the indicated proximity of a coexpressed gene. $\boldsymbol{exp}_{Z=3}$**:** the number of genes which must be observed in order to achieve significance at $Z = 3$. All groups of coexpressed genes in experiments FG1, FG2, FG5, FG6 and FG12 were tested; the groups not shown had no significant chromosomal clustering.

Group	proximity:	1	2	3	4	5
FG1↑	*p-value (Z)*		0 (5.80)	0 (4.55)	0 (4.60)	0 (3.62)
	obs ($exp_{Z=3}$)		245 (203)	299 (274)	362 (335)	398 (388)
FG2.01	*p-value*	0 (6.17)	0 (5.68)	0 (4.81)	1e-04 (4.12)	
	obs ($exp_{Z=3}$)	52 (34)	76 (56)	91 (75)	103 (92)	
FG2.10	*p-value*	0 (10.40)	0 (8.40)	0 (7.28)	0 (5.60)	0 (4.94)
	obs ($exp_{Z=3}$)	112 (56)	148 (94)	175 (127)	188 (156)	207 (182)
FG2.-1-1	*p-value*	0 (5.10)	0 (5.02)	1e-04 (4.27)	0 (4.23)	2e-04 (3.69)
	obs ($exp_{Z=3}$)	23 (16)	35 (26)	41 (34)	49 (41)	53 (48)
FG2.11	*p-value*	0 (7.07)	1e-04 (5.02)	0 (4.94)		
	obs ($exp_{Z=3}$)	25 (13)	29 (21)	37 (28)		
FG5↓	*p-value*	0 (11.00)	0 (8.24)	0 (6.34)	0 (5.36)	2e-04 (3.81)
	obs ($exp_{Z=3}$)	101 (46)	123 (76)	138 (104)	154 (128)	159 (149)
FG5↑	*p-value*	0 (6.40)	0 (6.24)	0 (4.41)	0 (3.86)	
	obs ($exp_{Z=3}$)	170 (128)	268 (218)	317 (294)	373 (358)	
FG5↑↓	*p-value*	1e-04 (4.24)	0 (4.48)	0 (5.30)	0 (4.51)	0 (4.42)
	obs ($exp_{Z=3}$)	36 (30)	58 (48)	83 (65)	93 (80)	107 (93)
FG6↓	*p-value*	0 (10.39)	0 (8.47)	0 (7.99)	0 (6.64)	0 (5.82)
	obs ($exp_{Z=3}$)	209 (121)	286 (205)	357 (277)	399 (339)	439 (392)

Group	proximity:	6	7	8	9	10
FG1↑	*p-value (Z)*	1e-04 (3.55)	1e-04 (4.34)	2e-04 (3.77)		
	obs ($exp_{Z=3}$)	443 (434)	495 (474)	521 (509)		
FG2.01	*p-value*					
	obs ($exp_{Z=3}$)					
FG2.10	*p-value*	0 (4.75)	0 (4.38)	0 (4.46)		
	obs ($exp_{Z=3}$)	229 (206)	246 (228)	267 (248)		
FG2.-1-1	*p-value*					
	obs ($exp_{Z=3}$)					
FG2.11	*p-value*					
	obs ($exp_{Z=3}$)					
FG5↓	*p-value*	0 (4.39)	0 (4.44)	0 (4.55)	0 (4.37)	4e-04 (3.67)
	obs ($exp_{Z=3}$)	186 (169)	205 (187)	223 (204)	236 (219)	242 (234)
FG5↑	*p-value*					
	obs ($exp_{Z=3}$)					
FG5↑↓	*p-value*	0 (4.20)	0 (4.10)	0 (4.03)	1e-04 (3.75)	
	obs ($exp_{Z=3}$)	118 (106)	129 (118)	139 (128)	146 (138)	
FG6↓	*p-value*	0 (4.89)	0 (5.36)	0 (4.55)	0 (4.47)	1e-04 (3.88)
	obs ($exp_{Z=3}$)	470 (439)	517 (480)	539 (516)	568 (547)	587 (575)

carbon-minimal media, but not in nitrogen-minimal media, compared to complete media growth conditions; 484 genes) includes 121 genes which are adjacent to a co-expressed gene ($g = 1$), whereas only 56 such genes must be observed in order to be significant at $Z = 3$. There is probably a high false positive rate associated with this observation. It is important to note that in the most recently released gene calls (version FG3, [137]) there were 26 merges of adjacent gene calls from version 1 (FG1) into a single gene call in version 3 (FG3). We may speculate, however, that a proportion of coexpressed adjacent genes have constrained chromosomal positions which may be due to shared promoter regions, transcriptional run-through of adjacent transcripts, or the presence of a transcriptionally active chromosomal region. There is no evidence of chromosomal clustering in the remaining coexpression groups, nor any evidence of significant clustering with a proximity of 10 genes or more in any coexpression group.

5.4.3.3 TAP-centric clustering of coexpressed genes

Eight transcription-associated proteins (TAPs) were identified as having significant enrichment of coexpressed genes in a chromosomal neighbourhood around the TAP gene, at a corrected p-value of $p < 0.05$ (Fisher's exact test with Bejamini-Hochberg correction [57]). Seven of the eight identified TAPs are DNA-binding. Two of the eight identified TAP-centric clusters overlap the aurofusarin gene cluster, and there are seven disjoint TAP-centric gene clusters (Figure 5.7).

Three of the seven TAP-centric clusters overlap known gene clusters of gene which have previously been shown to be coexpressed. Clusters overlapping the aurofusarin and butenolide biosynthesis gene clusters are downregulated during sexual development, while a cluster overlapping the mating type locus is upregulated in carbon-minimal conditions (Figure 5.7).

A TAP-centric cluster upregulated during sexual development contains the polyketide synthase gene involved in the biosynthesis of black perithecia pigment biosyn-

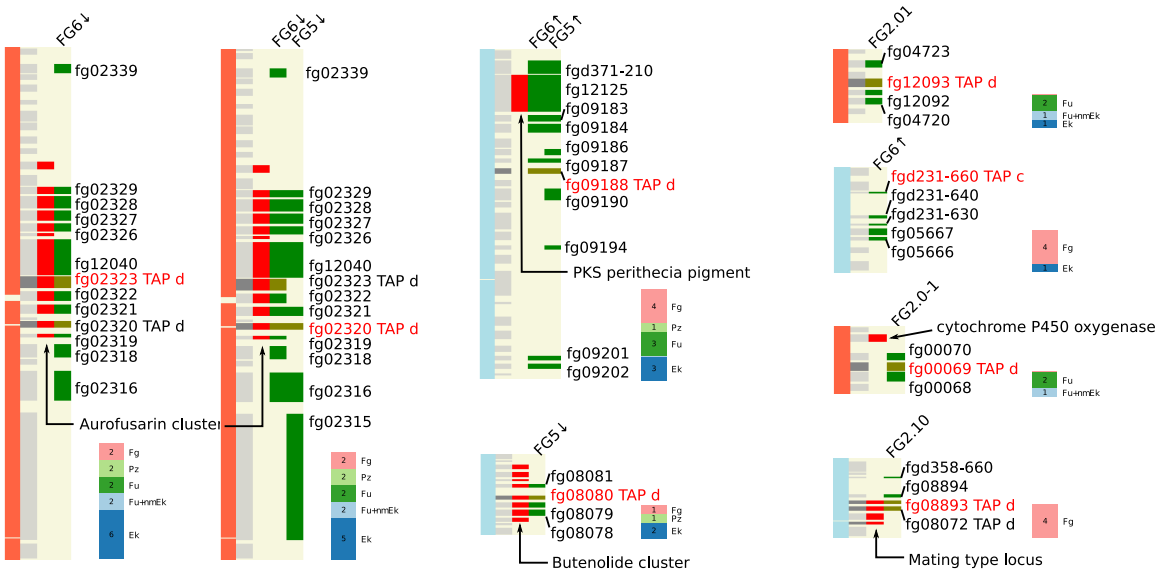


Figure 5.7: The eight TAPs located within a window significantly enriched for genes coexpressed with the TAP are shown ($p < 0.05$). For each region the columns are (left to right) (i) recombination rate, R (*red:* $3 \leq R < 8$; *yellow:* $1 \leq R < 3$; *blue:* $R < 1$ *cM/27kb*); (ii) all 14,100 protein entries (*dark grey:* DNA-binding TAPs; *light grey:* all other genes); (iii) (where shown) previously annotated genes of interest are shown in red and labelled; (iv) coexpressed genes with the expression pattern indicated at the head of the column (*olive green:* DNA-binding TAPs; *green:* all other genes). Barplots show the clade specificity of coexpressed genes in each cluster (*Fg:* Fusarium; *Pz:* Pezizomycotina; *Fu:* Fungi; *Fu+nmEuk:* Fungi and non-metazoan Eukaryotes; *Ek:* Eukaryotes).

thesis (*fg12125*). The gene cluster contains 11 genes which are coexpressed with the polyketide synthase gene. The identifying TAP is DNA-binding (*fg09188*). Coexpressed genes in the cluster may be part of a coordinately regulated perithecium pigment gene cluster and are candidates for further study.

The remaining three TAP-centric clusters do not overlap known or putative genomic clusters. These clusters contain 3–5 adjacent genes and therefore have a similar size to the detected gene cluster overlapping the butenolide synthesis gene cluster. One of the clusters (*fg00068 - fg00070*) is adjacent to the p450 oxygenase, Tri1 (*fg00071*). Tri1 is involved in the biosynthesis of trichothecene mycotoxins but lies distal to the trichothecene gene cluster [181]. However, the Tri1 gene was not detected as coexpressed with the adjacent gene cluster and the proximity of the detected cluster to Tri1 may not be functionally relevant.

Four of the seven disjoint TAP-centric clusters overlap known secondary metabolite gene clusters, or contain a polyketide synthase gene (and are therefore putative members of a coregulated secondary metabolite gene cluster). The lack of further TAP-centric clusters of coexpressed genes suggests that gene clusters of coexpressed genes containing a coexpressed TAP do not occur frequently in *F. graminearum* under the conditions considered here. The putative function of the three unidentified TAP-centric gene clusters (TAPs: *fg12093*, *fgd231-660*, *fg00069*) is investigated below.

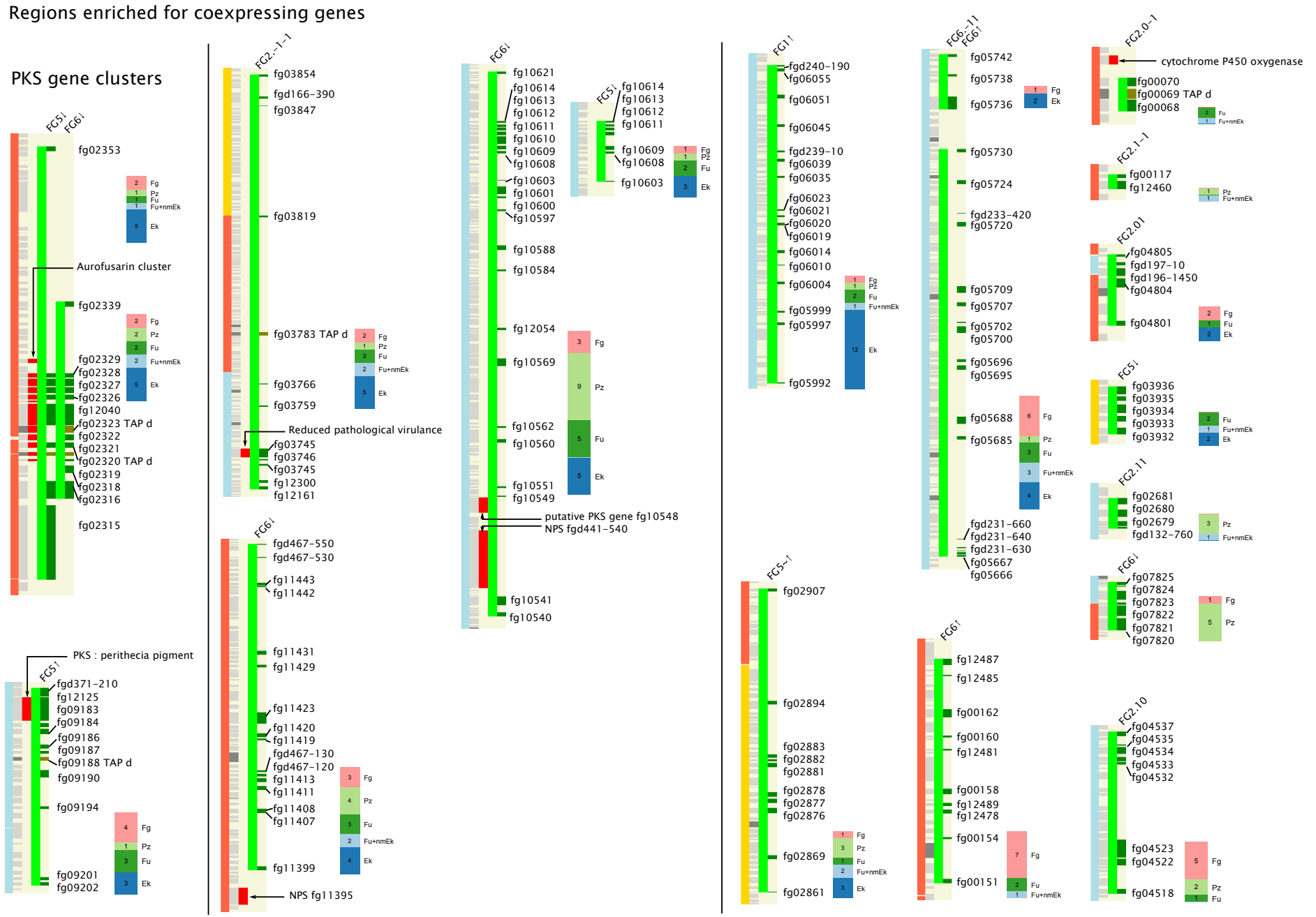


Figure 5.8: All summarized chromosomal regions enriched for coexpressed genes are shown (see Methods). Columns and barplots are as for Figure 5.7, with light green columns showing the maximal extent of nested or overlapping significantly enriched regions.

5.4.3.4 Localized chromosomal clusters of coexpressed genes

Eighteen disjoint chromosomal regions are significantly enriched for coexpressed genes (Figure 5.8). These regions are significantly enriched *pertinent regions* ($p < 0.05$ after $min\ p_i$ multiple testing correction; see Methods).

The 18 disjoint regions identified here as significantly enriched for coexpressed genes are considered to be high-confidence regions with respect to a shared functional and evolutionary significance, the exact nature of which is investigated below.

A further pair of adjacent genes (*fg01079*, *fg01080*) was found to define a significantly enriched region upregulated during wheat infection (experiment FG12). These two genes were merged into a single gene call in the most recent gene calls release (FG3) [137] and this cluster was therefore not considered further.

A strict significance criterion was used to identify chromosomal regions which are enriched for coexpressed genes. Clearly the density of coexpressed genes required in a given region in order to achieve the threshold for significant enrichment varies with the number of coexpressed genes. A conservative multiple testing method was used in order to reduce the false positive rate, so that regions identified as enriched for coexpressed genes are unlikely to have occured by chance due to the crowding of coexpressed genes into the genome (see Methods). There is likely to be a high false negative rate for the detection functional or evolutionarily selected clustering of coexpressed genes. It is difficult to estimate the false negative rate, however, given the incomplete current understanding of the function and evolution of coexpressed chromosomal clusters. For example, a genomic region spanning five adjacent coexpressed genes (*fg03932* - *fg03936*) is significantly enriched for genes which are persistently downregulated during experiment FG5 (*in vitro* sexual development; Table 5.1). On inspection, the same five adjacent genes are also coexpressed and persistently downregulated during experiment FG6 (*in vitro* sexual development, Δcch1). However this region does not meet the criterion for enrichment for coexpressed genes amongst the persistently downregulated genes in FG6: genomic re-

gions which are similarly dense in coexpressed genes are seen in more than 5% of random draws containing the same number of genes (see Methods). This discrepancy is due to the higher number of genes identified as persistently downregulated in experiment FG6 compared to experiment FG5 (781 genes in FG6; 426 genes in FG5; Table B.2): the occurence of five adjacent coexpressed genes is more likely by chance amongst the 721 genes differentially expressed in FG6 than amongst the 426 genes differentially expressed in FG5, and this region does not meet the conservative threshold for significant enrichment amongst the 721 genes persistently downregulated during FG6.

5.4.3.5 Functional annotation of localized coexpressed genes

Putative functional annotation was assigned to the coexpressed genes contained in TAP-centric gene cluster and localized coexpressed gene cluster (Figures 5.9, 5.10). A total of 170 genes were investigated for putative function as described (see Methods; page 115). To investigate systematically whether TAP-centric and non-TAP-centric gene clusters are associated with differential protein function, all identified TAP-centric and localized gene clusters were further combined into 20 distinct genomic regions (see Methods; page 119). Each combined region was identified as TAP-centric/non-TAP-centric according to the presence/absence of a coexpressed DNA-binding TAP in the region, and each region was annotated with the Gene Ontology identifiers associated with one or more coexpressed genes contained in the region. Regions were clustered using an asymmetric binary distance measure, such that the distance between regions A and B is the proportion of GO identifiers present in exactly one of A or B amongst GO identifiers present in A or B or both. Figure 5.11 shows a clustering of regions using the generic GO slim mapping [182], for example: the manual classification of TAP-centric and non-TAP-centric regions is not recovered, and there is therefore no evidence from this analysis that TAP-centric and non-TAP-centric clusters are associated with differential protein function.

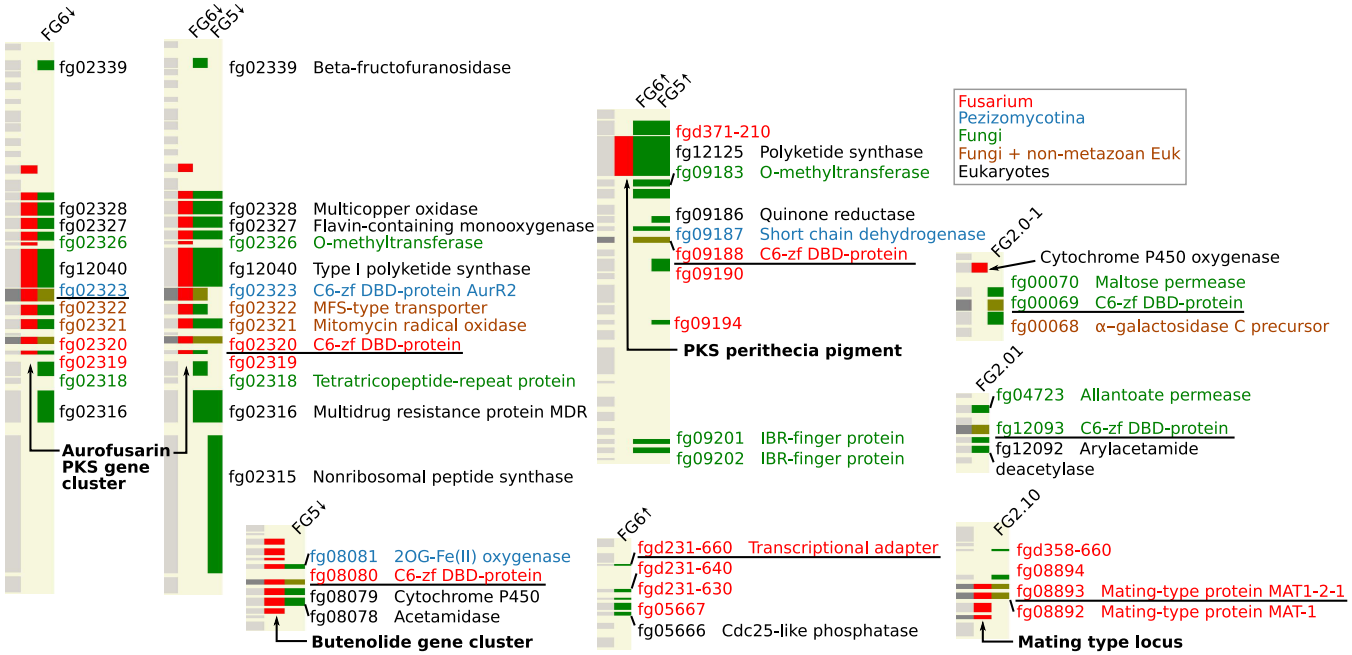


Figure 5.9: Predicted functional annotation and clade specificity of coexpressed genes are shown for all significant regions consisting of coexpressed genes flanking a coexpressed TAP. Clade specificity is colour coded as shown in the key. Functional annotation was assigned using a Blastp search against 54 eukaryotic genomes followed by Markov clustering of similarity scores (Blastp E-values). Functional annotation was inferred for each protein cluster by inspection of all annotated genes in the cluster. Clade specificity was determined according to the presence of clades within each protein cluster.

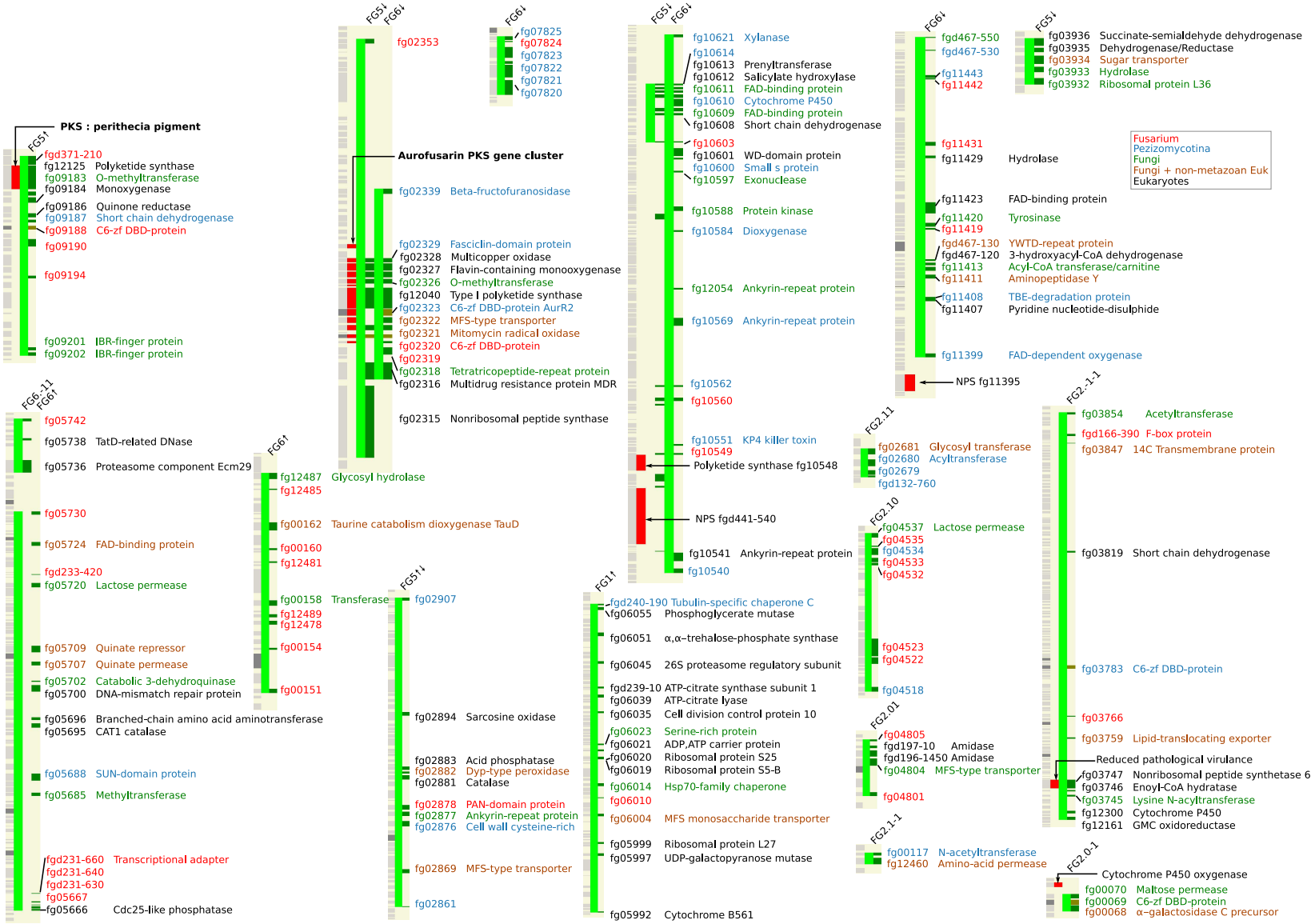


Figure 5.10: Predicted functional annotation and clade specificity of coexpressed genes are shown for all regions enriched for coexpressed genes. Clade specificity is colour coded as shown in the key. Functional annotation and clade specificity were assigned as for Figure 5.9.

5.4.4 Bias in protein domain composition of detected and differentially expressed TAPs

A bias was observed in the protein domain composition of TAPs which are detected, or detected and differentially expressed (Figure 5.12, broken down by clade specificity). In particular, in all GeneChip experiments, TAPs containing a zinc cluster domain were found to be underrepresented amongst detected TAPs. In crop infection experiments (FG1, barley head infection and FG12, wheat crown infection), TAPs containing $bZIP_1/bZIP_2$, GATA and C_2H_2 zinc finger domains were found to be overrepresented amongst all detected TAPs. While this observation is based on a small number of experiments and a sparse matrix of protein domain counts, the concordance of domain overrepresentation between the two infection timecourses and the consistent absence of zinc cluster domains amongst detected TAPs is striking. It remains to be seen whether these observations continue to hold as additional datasets are made available.

5.5 Discussion

Groups of detected and coordinately expressed genes were identified within six transcriptomics experiments, representing various stages of the *F. graminearum* lifecycle and crop infection. Furthermore, coexpressed predicted DNA-binding TAPs were identified and may be considered as candidates for transcriptional regulators of coexpressed genes.

A dataset of *F. graminearum* coexpressed gene clusters is presented together with putative functional annotation. Twenty distinct genomic regions were found to be significantly dense in coexpressed genes. Eight gene clusters were found to contain coexpressing genes for DNA-binding TAPs; almost all such clusters are already well-characterized and there is no evidence from this study that TAP-centric coexpressed gene clusters are widespread in the *F. graminearum* genome. The aurofusarin and butenolide biosynthesis gene clusters [176, 179] were identified amongst the 20

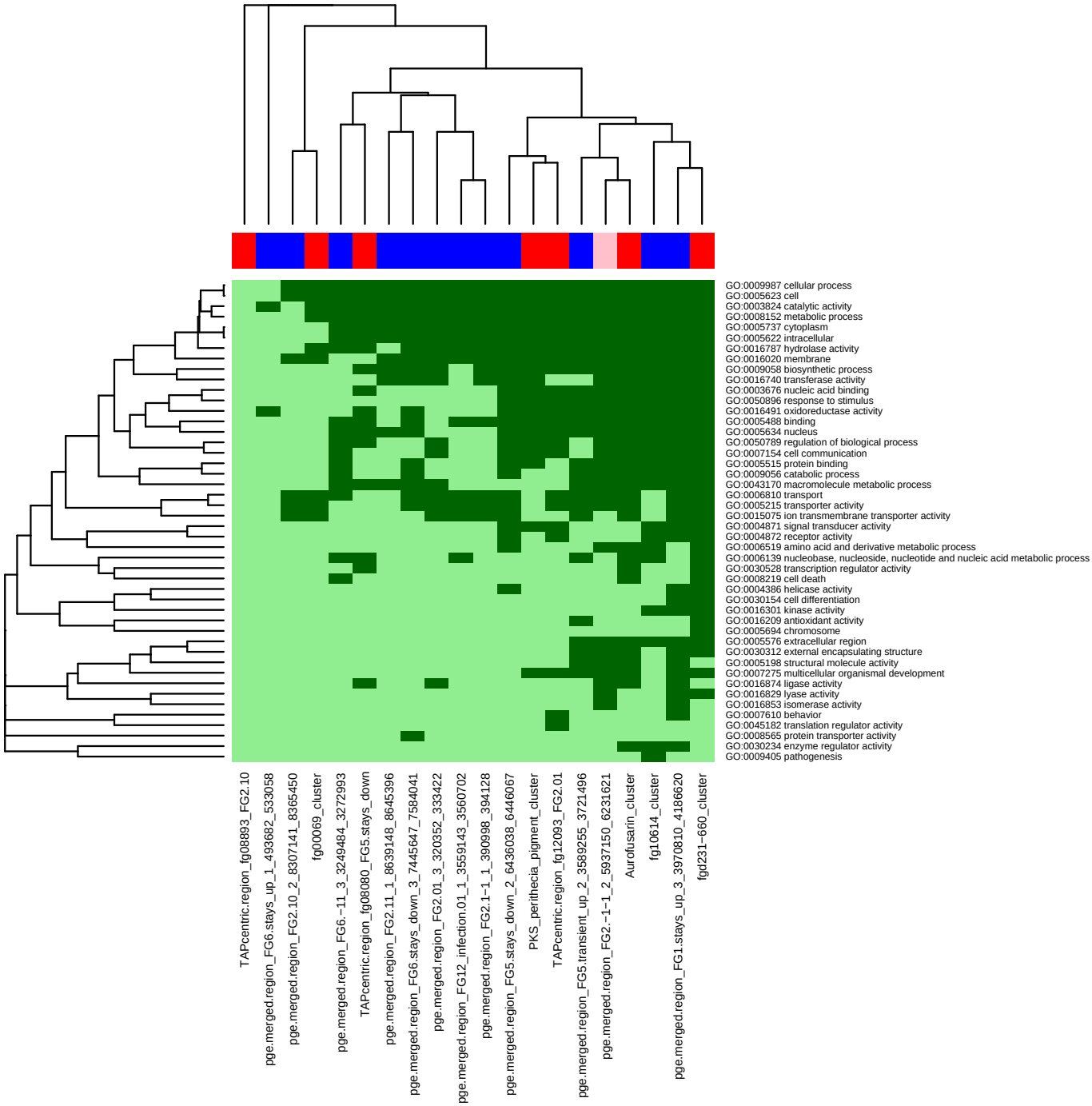


Figure 5.11: Heatmap representation of generic GO slim identifiers associated with coexpressed gene clusters. *Dark green*: annotated. *Light green*: not annotated. Rows (gene clusters) and columns (GO IDs) are clustered by complete linkage using an asymmetric binary distance function. Row and column dendrograms are shown. The horizontal colour bar (top) identifies gene clusters which contain a coexpressed TAP gene. *Red*: TAP-centric gene cluster (localized around a TAP gene). *Pink*: contains a coexpressed gene for a DNA-binding TAP but not identified as a TAP-centric gene cluster. *Blue*: does not contain a coexpressed DNA-binding TAP gene. *Rows*: generic GO slim identifiers associated with one or more gene clusters. *Columns*: distinct genomic clusters of coexpressed genes.

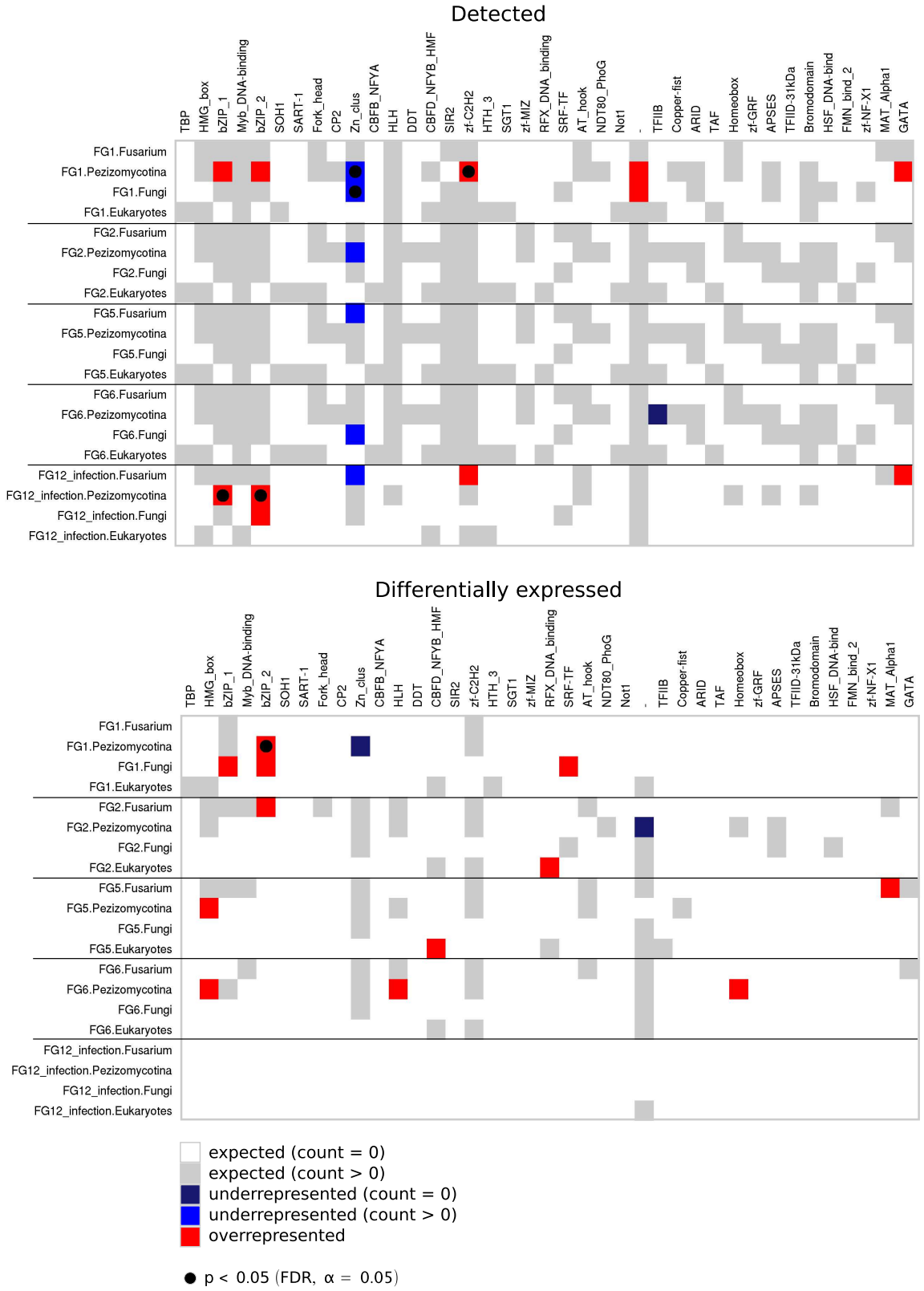


Figure 5.12: Bias in protein domains amongst detected (top) and differentially expressed (bottom) DNA-binding TAPs. x-axis: protein domain HMM classes. y-axis: *F. graminearum* experiments, further broken down by clade specificity (*Fusarium*, *Pezizomycotina*, *Fungi*, *Eukaryotes*). For each *experiment.clade*, under/overrepresentation of each HMM class is shown (χ^2 test; $p < 0.05$). In addition, FDR-adjusted $p < 0.05$ is shown after correcting for multiple testing within each row (*experiment.clade*).

coexpressed gene clusters. A previously uncharacterized coexpressed gene cluster was identified which contains the black perithecium pigment polyketide synthase gene [178], and this may partially define a perithecium pigment biosynthesis gene cluster. A systematic analysis of the putative function of genes coexpressed in localized gene clusters did not identify differential protein function between genes located in TAP-centric and non-TAP-centric coexpressed gene clusters.

Predicted basal transcription factors, chromatin and histone remodelling factors, and RNA polymerase subunits were found to be underrepresented in *F. graminearum* regions with relatively high recombination rate. Pál and Hurst [183] showed that clusters of essential genes in *S. cerevisiae* lie in regions of low recombination rate, and that the low recombination rate around clusters of essential genes may be driven by selection for genetic proximity of essential genes rather than by a bias in mutation rate. Although recombination is thought to be mutagenic, so that a higher recombination rate is associated with a higher mutation rate [184], this is not consistent with Pál and Hurst's observation that unclustered essential genes do not appear to have low recombination rates [183].

Subsets of genes in each expression group may be coregulated by DNA-binding transcription factors including the putative DNA-binding TAPs identified in this study. However, there has been no attempt here to identify such subsets of coregulated genes. To identify putative transcriptionally coregulated genes, ChIP-chip [46] or ChIP-Seq [185] experiments could help to identify *in vivo* binding sites of certain transcription factors. In the absence of such experimental data, shared sequence motifs could be sought in the promotors of coexpressed genes. This was not investigated at the present time, however, because a search for *de novo* DNA-binding motifs is likely to have a high false positive rate due to the typically short length of binding motifs in promotor regions [186]. An approach by Fraenkel and others [76] in *S. cerevisiae* used closely related genome sequences to identify *de novo* sequence motifs which are conserved between organisms, and this will become feasible for the *F. graminearum* genome with the release of whole-genome sequences for related

Fusarium species [187].

There are many possible approaches for defining groups of coexpressed genes within one or more experiments. The methods used in this study to call differential expression (limma with significance cutoffs determined by *AFFX* control probesets) and probeset detection (*MAS 5.0*) were selected to produce a consistent, conservative set of differentially expressed genes within each experiment. Alternative approaches include cluster analysis of individual or concatenated experiments, and differential expression analysis based on linear modelling of conditions but using alternative differential expression criteria or including minimum fold-change constraints. Furthermore, the summary of limma groups into larger coexpression groups is in general an arbitrary choice, made here to consolidate a large number of similar expression profiles into a smaller number of coexpression groups. On inspection of the expression profiles of genes in each limma group, division into more fine-grained coexpression groups did not appear to be justified and the chosen coexpression groups were judged to be useful coexpression summaries for these datasets. If additional datasets are included in this analysis as part of future work then the set of summarized coexpression groups may need to be extended, for example if there are distinct groups of genes upregulated at different times during a timecourse.

Chapter 6

Discussion

With the arrival of post-genomic datasets, the genome-wide regulation of transcriptional and post-transcriptional processes can be modelled at different levels and with varying levels of complexity. This thesis has presented three studies of transcriptional regulation based on genome-wide gene expression datasets from three fungal species. Each of the three studies investigated distinct aspects of a gene expression response:

- transcriptional control of a gene expression timecourse by linear combinations of binding transcription factors, in the model organism *S. cerevisiae*;
- the contribution of regulated mRNA stability to shaping a gene expression response, in the model organism *S. pombe*;
- differential expression between steady-state conditions and the coexpression of predicted transcription factors, in the crop pathogen *Fusarium graminearum*.

In the first study, a high resolution stationary phase timecourse was modelled as the sum of interactions between active (for example, post-transcriptionally modified) DNA-binding proteins and coregulated target genes. Two related models and inference methods were applied in order to quantify hidden variables – interpreted as concentrations of active transcription factors and their control of target genes – which can explain the observed gene expression profiles of target genes. While such

models are vastly simplified representations of transcriptional control, they are a useful exploratory tool for identifying putative functional transcription factor-target interactions.

The second study considered the contribution of regulated mRNA stability to shaping a gene expression response. A dynamic model of mRNA abundance was used to identify genes which are better explained by a change in mRNA stability after stress induction than by constant mRNA stability. This finding has implications for dynamic models of transcriptional regulation: excluding the effect of regulated mRNA stability from a model of transcriptional regulation may erroneously force observed mRNA abundance profiles to be explained by a regulated transcription rate.

Finally, the third study presented an integrative study of gene expression patterns in the fungal crop pathogen *F. graminearum*. Unlike in the first two studies, in which the datasets were from the model organisms *S. cerevisiae* and *S. pombe*, comparatively little is known about the transcriptional regulation of *F. graminearum* and there is relatively little existing genome annotation. All available gene expression datasets were analysed in order to identify genes and putative transcriptional regulators which are differentially expressed between different near-steady-state conditions. This has provided a first genome-wide survey of coexpressed genes and predicted transcription factors in *F. graminearum* which will contribute to the understanding of gene expression programmes and transcriptional regulation in this economically important pathogen. Additionally, it was shown that groups of coexpressed genes are found in localized regions of the genome, and a comparative genomics approach was used to annotate these genes with putative protein function.

The technologies used to study transcription and transcription-related processed on a genome-wide scale are rapidly developing. High-throughput sequencing technologies can be used to map all transcripts which are present in a population of cells, while recent developments survey populations of RNA-binding proteins, *in*

vivo protein-protein interactions and subcellular locations, and chromatin modifications, amongst other regulatory processes in the cell. Our ability to use genome-scale datasets to construct quantitative or semi-quantitative models of transcriptional and post-transcriptional regulation will be aided by the continued development of methods which are able to model genome-scale datasets and to quantify technical and biological errors inherent in the data.

Appendix A

Supplementary Tables for Chapter 4

Clusters are listed in the order in which they are mentioned in Chapter 4.

Genes are listed using *S. pombe* systematic gene IDs, or converted to *S. pombe* gene names where these exist. (For further details of *S. pombe* systematic IDs and gene names, see the *S. pombe* GeneDB database [188]).

Table A.1: Putative destabilized (delayed) [Figure 4.7]

Cluster 6	Cluster 7	Cluster 8
SPBC19F5.02c	SPAC24H6.01c	SPBC1604.06c
SPBC428.15	SPCC63.07	SPAC664.08c
SPAC1687.16c	SPBC660.14	SPAC1F7.02c
SPAC977.11	SPCC613.07	SPAC222.06
SPCC1393.06c	SPAC31G5.10	SPAC16E8.06c
SPBC359.06	SPBC1709.01	SPBC83.15
SPAPB1A10.15	SPBC4F6.07c	SPBC17D1.02
SPBC16C6.10	SPAC1039.02	SPBC1604.09c
SPBC36B7.04	SPAPB15E9.01c	SPAC22A12.05
rpl12-1: rpl12.1		pmt2
SPBC365.04c		SPBC244.02c
SPAC890.04c		SPBC16C6.12c
SPBC1A4.07c		SPAC144.01
SPAC23C4.15		SPBC947.07
SPAP7G5.02c		SPCC24B10.18
SPAC1B3.13		SPBC19C2.13c
SPAC1D4.04		SPAC926.08c
SPBP16F5.05c		SPBC800.06
		SPBC29A3.06
		SPCC330.09

Table A.2: Putative destabilized (rapid) and putative stabilized (rapid) [Figure 4.10]

Destabilized				Stabilized
Cluster 17	Cluster 20	Cluster 21	Cluster 22	Cluster 32
SPAC13G6.10c	SPBC11G11.03	SPAC1556.03	SPAC1002.06c	SPBC15D4.07c
SPAC14C4.12c	SPBC11G11.05	SPCC1020.08	SPCC1020.09	SPAC1006.01
SPAC1687.10	SPAC12G12.02	SPAC1093.05	SPAC10F6.03c	SPBC106.13
SPCC16C4.05	SPCC1450.10c	SPCC1183.07	SPAC1142.04	SPCC1223.10c
SPAC19D5.10c	SPAC1486.09	SPCC1235.07	SPBC13G1.09	SPBC1347.01c
SPBC25B2.07c	SPBC16D10.01c	SPAC12G12.06c	SPAC144.12	SPBC1683.06c
SPAC25B8.15c	SPBC1711.07	SPCC1450.03	SPAC1565.05	SPBC1734.08
SPAC26A3.03c	SPCC1739.05	SPAC1527.03	SPCC16C4.07	SPAC18G6.01c
SPAC3F10.15c	SPAC17H9.05	SPAC16.04	SPBC16E9.10c	SPAC1D4.02c
SPAC3G6.11	SPCC1827.05c	SPAC16.05c	SPBC16G5.10	SPBC215.07c
SPAC4D7.04c	SPAC18B11.06	SPCC1682.12c	SPBC1703.05	SPAC22H12.03
SPBC56F2.07c	SPAC19A8.07c	SPAC1687.19c	SPBC1709.02c	SPAC23C4.08
SPBC839.07	SPAC19B12.01	SPCC16C4.06c	SPAC1834.10c	SPAC25A8.02
SPAC1006.08	SPAC19B12.11c	SPAC16C9.03	SPBC18H10.20c	SPAC27E2.06c
SPAC27D7.03c	SPAC19G12.16c	SPBC1703.03c	SPCC1919.13c	SPAC2C4.07c
SPAC25H1.05	SPAC20G8.09c	SPBC1711.05	SPAC19D5.05c	SPAC2C4.15c
SPAPB15E9.02c	SPBC215.06c	SPAC1751.04	SPAC22G7.05	SPBC2F12.15c
SPCC320.13c	SPAC22F8.09	SPBC17D1.05	SPAC139.06	SPAC2G11.13
SPNCRNA.132	SPAC23H4.15	SPCC18.01c	SPBC23G7.07c	SPBC365.09c
	SPBC19F5.05c	SPCC1827.01c	SPAC24C9.11	SPAC3A12.08
	SPBC26H8.08c	SPBC18E5.03c	SPAC27D7.12c	SPAC3H1.11
	SPAC2C4.11c	SPCC191.02c	SPAC30C2.04	SPAC4G9.19
	SPBC2G5.03	SPCC191.08	SPBC31A8.02	SPBC530.05
	SPCC320.11c	SPAC23C11.03	SPBC3F6.04c	SPAC5D6.04
	SPAC3G9.10c	SPAC23C4.17	SPBC409.15	SPCC61.05
	SPBC4F6.13c	SPBC23E6.05	SPBC428.19c	SPCC622.15c
	SPBC4F6.14	SPAC23H3.07c	SPAC4F10.06	SPAC630.05
	SPAC4F8.04	SPAC25B8.05	SPAC4G8.02c	SPCC74.03c
	SPBC651.01c	SPAC26A3.06	SPCC569.02c	SPAC750.05c
	SPAC823.04	SPBC27B12.09c	SPAC56F8.03	SPCC962.01
	SPAC890.05	SPBC2A9.13	SPAC56F8.09	SPBC14C8.01c
	SPAC8F11.04	SPAC2C4.06c	SPAC57A7.06	gna1:spgna1
	SPBC9B6.07	SPAC2C4.12c	SPCC31H12.08c	mhk1:pmk1
	SPAC140.02	SPAC2G11.02	SPAC607.02c	SPBP35G2.12
	SPBP8B7.20c	SPAC31A2.07c	SPCC613.08	SPBP4H10.12
	SPCPB16A4.04c	SPAC31G5.02	SPCC613.12c	SPBP4H10.16c
	SPCC1259.03	SPCC320.12	SPCC622.14	SPBP4H10.19c
		SPCC364.01	SPBC31E1.06	SPAP8A3.12c
		SPBC3B8.05	SPBC800.08	SPBP8B7.13
		SPBC11C11.10	SPCC825.04c	SPAPB24D3.03
		SPAC3F10.16c	SPBC83.18c	SPAC57A10.05c
		SPBC28E12.05	SPBC839.14c	SPBC16H5.07c
		SPAC4F10.05c	SPCC895.06	SPBC649.03
		SPBC543.06c	SPAC26A3.17c	SPAC688.13
		SPAC56F8.10	SPAC1420.02c	SPBC365.05c
		SPCC576.01c	CTOKYO_453.18	SPAC10F6.09c
		SPCC584.01c	SPACUNK4.09	SPAC9.04
		SPAC6G9.02c	SPAC2F7.11	SPAPB15E9.03c
		SPAC6G9.10c	SPBC646.14c	SPAC27E2.08
		SPCC736.02	SPCP1E11.08	SPAC13D1.01c
		SPBC776.08c	SPBC3B9.07c	SPCC777.10c
		SPBC776.17	rpb6:rpo15	SPBC16A3.09c
		SPAC823.08c	SPBC14C8.12	
		SPBC839.11c	rpc19:rpa17	
		SPBC8D2.10c	rpc40:rpa42	

continued on next page ...

... continued				
Cluster 17	Cluster 20	Cluster 21	Cluster 22	Cluster 32
		SPAC9G1.12	SPBC776.01	
		dim1:C336.02	SPAC15E1.03	
		gpm1	rps30-1	
		hsk1	sfc6	
		kap123	snu13	
		lps1	sso1	
		misc_RNA_1.1.46.RC	SPCC584.04	
		nuc1: rpa1	sup45	
		P1E11.11	uvi22	
		P22H7.10c		
		pfk1		
		pi031		
		prh1		
		rpa49		
		rpl30: rpl30-1		
		sum3: ded1: slh3: moc2		

Table A.3: Reduced transcription rate and mRNA abundance [Figure 4.10]

Cluster 27	Cluster 28	Cluster 29
SPCC1393.08	SPAC6B12.15	SPAC4F10.14c
SPBC25B2.09c	SPAC9.09	SPBC106.14c
SPBC25H2.05	SPAC513.01c	SPAC10F6.01c
SPBC29A3.07c	SPAPB1E7.12	SPBC12C2.07c
SPBC56F2.12	SPCC31H12.04c	SPAC110.01
SPAC589.10c	rpl1-2:rpl10a-2	SPBC14F5.06
SPBC839.15c	SPAC1805.13	SPBC15D4.02
SPBC19C2.07	SPCC576.11	SPCC162.01c
SPBC2D10.10c	SPBC2F12.04	SPCC16A11.10c
pab1:pabp	rpl18-1:rpl18	SPBC16A3.08c
SPAP7G5.05	SPAPB17E12.13	SPAC1782.11
SPAC26A3.07c	SPCC1322.11	SPAC17G8.05
SPBC17G9.10	rpl27-2:rpl27b	SPAC17G8.08c
SPAC664.05	rpl28-1:rpl27a	SPBC18E5.07
rpl13a-1:rpl16-1	SPCC5E4.07	SPAC18G6.11c
SPAC1783.08c	SPAC17A5.03	SPCC1450.02
SPCC1682.14	SPBC1711.06	SPBC19G7.06
SPAC3A12.10	SPBP8B7.03c	SPAC20G8.04c
SPAC26A3.04	SPAC3H5.12c	SPBC211.05
SPAC959.08	SPBC11C11.09c	SPAC24C9.06c
SPAC11E3.15	SPCC622.18	SPBC29A3.08
SPAC3G9.03	SPAC3H5.07	SPBC29A3.13
SPCC330.14c	SPBC18H10.12c	SPBC29A3.18
SPBC29B5.03c	SPBC29A3.04	SPAC2F3.13c
rpl27-1:rpl27a	rpl8-2:rpk37:rpk5b	SPCC330.13
SPAC890.08	SPAC4G9.16c	SPBC36.01c
SPBC16C6.11	SPBP8B7.06	SPCC548.06c
SPAC3H5.10	SPCC1393.03	SPBC685.06
SPAC23A1.08c	rps15a-2:rps22-2	SPBC713.12
SPCC1322.15	SPBC16D10.11c	SPCC74.04
rpl35a:rpl33	SPCC1259.01c	SPCC1529.01
SPCC970.05	SPBC21C3.13	SPBC839.12
SPBC405.07	SPAC17G6.06	SPBC8D2.16c
SPAPB17E12.05	rps3a-1:rps1-1	SPCC965.14c
rpl37a-1:rpl43-1	rps3a-2:rps1-2	SPAC977.12
rpl37a-2: rpl43-2	rps4-2	cyc1
rpl38-1	rps5	eft2-2
rpl41-2	rps6-1	eno1
rpp1-2	rps9-1: rps9a	gaf2
rps11-1	tif1	hos3
rps11-2	ubi3	ilv3
rps12-1		leu2
rps15-2		lys7
SPCC576.08c		SPNCRNA.101
rps24-2		P4H10.15
rps26-2		pac2
rps29		PB10D8.01
rps30-2		PB1A10.14
rps4-3		PB1E7.07
rps7		PB2B2.09c
rps8-2		pma1
rps9-2: rps9b		pol5
rpsa-2: rps0-2		prl35
tif51		rpl13a-2: rpl16-2
uep1: ubi2		rpl17-2

continued on next page ...

... continued		
Cluster 27	Cluster 28	Cluster 29
		rpl24 rpl25a rpl30-2 rpl35 rpl3-b rpl44: rpl28 rpl8-1: rpk5a: rpl2-1: rpk5 rpl8-3: rpk5-b: rpkd4 rpp1-3 rpp2-3: rla6 rps10-1 rps10-2: pi023 rps14-1 rps17-1 rps19-2 rps23-2 rps26-1 rps28-1 rps28-2 rps3 rps4-1 sce3 tif512 ups ura1

Table A.4: Exponential approach to increased mRNA abundance level [Figure 4.14]

Cluster 1	Cluster 2	Cluster 3
SPAC1006.01	SPAC11E3.14	SPBC11C11.06c
SPCC1494.03	SPAC823.03	C1685.13.RC
SPBC1683.06c	SPBC21H7.06c	SPBC16E9.16c
SPAC1687.14c	SPAC22E12.03c	SPBC21C3.19
SPAC16E8.16	SPBC23G7.10c	SPAC23H3.15c
SPBC16E9.11c	SPAC607.08c	SPBC365.12c
SPBC25B2.10	SPAC824.07	SPAC637.03
SPAC29B12.11c	NC133b	SPAC9E9.04
SPAC2C4.07c	SPAC3C7.14c	SPCC757.07c
SPBC2F12.15c	SPCC330.06c	SPACUNK4.15
SPAC3A12.06c	atf1:mts1:sss1:gad7.B	SPBC32F12.03c
SPBC530.05	SPBC106.03	SPAP8A3.04c
SPAC57A7.09	SPAC18G6.09c	I23_C660.16
SPCC594.06c	SPAC1687.22c	SPNCRNA.44
SPCC622.15c	SPCC23B6.01c	SPAC343.12
SPCC736.15	SPBC25B2.03	SPAC328.03
SPCC594.01	SPAC25H1.03	SPAC10F6.06
SPAC23C4.12	SPAC26H5.04	SPAC57A10.09c
SPBC4F6.06	SPAC29A4.17c	
mcs1:res2:pct1	SPAC31A2.12	
SPBP4H10.16c	SPAC328.04	
SPBC16H5.07c	SPAC328.07c	
SPAC4A8.03c	SPCC4G3.13c	
SPAC688.13	SPCC553.10	
SPBC365.05c	SPCC569.01c	
SPCC777.10c	SPAC607.09c	
SPBC16A3.09c	SPCC63.13	
	SPAC688.03c	
	gti1	
	P7G5.01	
	PB24D3.09c	
	trx2	
	uvi15	

Appendix B

Supplementary Tables for Chapter 5

Table B.1: Probesets mapping to RNA polymerase subunits. Considered to be an invariant set in order to fit a variance-stabilizing normalization function for experiment FG1 (barley infection timecourse)

Invariant set: probesets mapping to RNA polymerase subunits
fgd254-570_at
fgd228-190_at
fgd185-1310_at
fgd347-50_at
fg12242_at
fgd185-1110_at
fgd61-20_at
fgd230-70_at
fgd30-30_at
fg06937_s_at
fgd41-90_at
fgd35-880_at
fgd231-150_at
fgd37-350_at
fgd192-280_at
fgd451-80_at
fgd132-550_s_at
fg02659_s_at
fgd74-90_at
fgd183-140_at
fg06683_s_at
fgd266-550_at
fgd13-370_at
fgd276-280_s_at
fg12319_s_at
fgd259-1330_at
fgd56-110_at

Table B.2: Coexpression groups identified from each *F. graminearum* GeneChip experiments. DNA-binding TAPs (dTAPs) are listed, colour-coded according to clade specificity (*black: Eukaryotes; green: Fungi; blue: Pezizomycotina, red; Fusarium*)

Coexpression group	#genes	#TAPs (#dTAPs)	DNA-binding TAPs
FG1~	10	1 (1)	fg11623
FG1↓	4	0 (0)	
FG1↑	781	14 (10)	fg10179, fg11627, fg08696, fg10142, fg07863, fgd108-320, fg07075, fg09286, fg07052, fg09715
FG1↓↑	0	0 (0)	
FG1↑↓	13	0 (0)	
FG12.-1-1	1	0 (0)	
FG12.-11	1	0 (0)	
FG12.0-1	1	0 (0)	
FG12.01	19	1 (0)	
FG2.-1-1	206	11 (11)	fg07079, fg09333, fg03566, fg10350, fg09001, fg10429, fg01350, fg09177, fg00813, fg03783, fg04683
FG2.-10	321	7 (6)	fg11561, fg07431, fg00240, fg02939, fgd304-860, fg03201
FG2.-11	69	1 (1)	fg10674
FG2.0-1	131	5 (5)	fg09217, fg00069, fg04035, fg12190, fg10470
FG2.01	348	35 (35)	fg03881, fg03159, fg10639, fg12093, fg04786, fg09524, fg08696, fg03727, fg03487, fg03606, fg03390, fg07638, fg03327, fg12416, fg00153, fg04671, fg00678, fg10627, fg03214, fg09064, fg04643, fg10129, fg08321, fg03786, fg12454, fg03415, fg09047, fg04170, fg04974, fg03649, fg06481, fg11462, fg01760, fg01378, fg06934
FG2.1-1	23	1 (1)	fg10266
FG2.10	484	13 (11)	fg06359, fg09884, fg08892, fg00307, fgd146-300, fg08893, fg00795, fg09832, fgd122-200, fg05151, fg01307
FG2.11	178	15 (15)	fg07420, fg05283, fg12061, fg06231, fg10731, fg02750, fg05682, fg01172, fg08397, fg05381, fg04803, fg01173, fg07368, fg08434, fg08791
FG5~	101	6 (4)	fg08972, fgd150-720, fg00342, fg03786
FG5↓	426	8 (8)	fg08080, fg03390, fg12575, fg04932, fg02320, fg11301, fg04170, fg10812
FG5↑	806	23 (18)	fg05304, fg07420, fg09217, fg08626, fg03727, fg01327, fg01214, fg02969, fgd185-1060, fg07187, fg00696, fg05242, fg08890, fg05904, fg09188, fg04480, fg12757, fg06160

Continued on next page

Table B.2 – **continued from previous page**

Coexpression group	#genes	#TAPs (#dTAPs)	DNA-binding TAPs
FG5↓↑	247	1 (1)	fg03783
FG5↑↓	316	17 (16)	fg00725, fg08892, fg08064, fg08411, fg03040, fg08431, fg05682, fg05041, fg08321, fg03912, fg01576, fg00800, fg01366, fg01173, fg05151, fg08434
FG6.-11	38	1 (1)	fg01724
FG6.1-1	1	0 (0)	
FG6↓	781	28 (26)	fg03292, fg07079, fg02696, fg02068, fg07638, fg04932, fg04083, fg06713, fg04035, fg02320, fg12462, fgd150-720, fg05588, fg02676, fg06714, fg07052, fg02323, fg01341, fg10057, fg06382, fg03861, fg01176, fg06421, fg03663, fg03783, fg07097
FG6↑	668	24 (15)	fg05304, fg09217, fg03881, fg12654, fg00573, fg08972, fg01214, fg00678, fg09188, fg06262, fg01100, fg09047, fg12757, fg07504, fg01139
FG6↓↑	126	1 (1)	fg02676
FG6↑↓	98	5 (4)	fg01327, fg04191, fg01366, fg05151
FG7_2h_spores.-1	1222	40 (16)	fg05304, fg06427, fg09857, fg07076, fg08892, fg01488, fg09868, fg04557, fgd108-320, fg07187, fg01201, fg07052, fg06684, fg09410, fg10944, fg12528
FG7_2h_spores.1	1234	69 (60)	fg04293, fg01915, fg09921, fg10851, fg03292, fg09770, fg00713, fg01936, fg09884, fg12406, fg01562, fg09111, fg07863, fg02874, fg10440, fg08064, fg09178, fg10429, fg02633, fgd213-220, fg02320, fg12658, fgd467-70, fgd210-20, fg01507, fg05926, fg11271, fgd101-60, fg08431, fg02787, fg02323, fg04680, fg08397, fg04888, fg08321, fg06262, fg05402, fgd224-590, fg06311, fg12287, fg09014, fg08380, fg01753, fg11301, fg08617, fg04974, fgd246-10, fg08403, fg06160, fg05068, fg05503, fg12183, fg06324, fg05151, fg07097, fg08713, fg08434, fg03219, fg01731, fg10868

Table B.3: Table of protein entries differentially expressed during both crop infection experiments FG1 (barley head infection) and FG12 (wheat crown infection). [1]Broad gene name is shown unless stated (MIPS). [2]fg01079 and fg01080 were later merged in gene calls version *FG3* [187].

		Coexpression group		
Gene ID	Chromosome	FG1	FG12	Gene name (Broad/MIPS)[1]
fg00062	1	FG1↑	FG12.01	[hypothetical]
fg01079	1	FG1↑	FG12.01	ATP synthase subunit alpha, mitochondrial precursor[2]
fg01080	1	FG1↑	FG12.01	ATP synthase subunit alpha, mitochondrial precursor[2]
fg01956	1	FG1↑	FG12.01	[conserved hypothetical]
fg02284	1	FG1↑	FG12.01	[hypothetical]
fg08366	2	FG1↑	FG12.01	[conserved hypothetical] (MIPS)
fgd166-270	2	FG1↑	FG12.-11	Probable glycine-rich RNA-binding protein (MIPS)
fg06021	3	FG1↑	FG12.01	ADP,ATP carrier protein
fg06268	3	FG1↑	FG12.01	Cytochrome c oxidase polypeptide VIb
fg06289	3	FG1↑	FG12.01	60S ribosomal protein L3
fg11368	3	FG1↑	FG12.01	[hypothetical] similar to mannose-binding lectin
fg12313	4	FG1↑	FG12.01	Probable translation initiation factor eIF-4A (MIPS)
fg06931	4	FG1↑	FG12.01	60S ribosomal protein L2
fg07335	4	FG1↑	FG12.01	Actin
fg07530	4	FG1↑	FG12.01	Alcohol oxidase
fg07647	4	FG1↑	FG12.01	[predicted protein]
fg07765	4	FG1↑	FG12.01	[hypothetical] similar to cytochrome P450 monooxygenase
fg09690	4	FG1↑	FG12.01	[conserved hypothetical]

Bibliography

[1] N. J. Proudfoot, A. Furger, and M. J. Dye. Integrating mRNA processing with transcription. *Cell*, 108:501–512, 2002.

[2] R. G. Roeder. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Lett*, 579:909–915, 2005.

[3] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nature Struct Mol Biol*, 14:796–806, 2007.

[4] J. Zeitlinger, A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genet*, 39:1512–1516, 2007.

[5] Radonjic M., Andrau J. C., Lijnzaad P., Kemmeren P., Kockelkorn T. T., van Leenen D., van Berkum N. L., and Holstege F. C. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Molec Cell*, 18:171–83, 2005.

[6] Muse G. W., Gilchrist D. A., Nechaev S., Shah R., Parker J. S., Grissom S. F., Zeitlinger J., and Adelman K. RNA polymerase is poised for activation across the genome. *Nature Genet*, 39:1507–11, 2007.

[7] B. Gebelein, D. J. McKay, and R. S. Mann. Direct integration of Hox and segmentation gene inputs during Drosophila development. *Nature*, 431:653–659, 2004.

[8] J. Smith, C. Theodoris, and E. H. Davidson. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*, 318:794–797, 2007.

[9] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[10] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S.

Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[11] O. Hobert. Gene regulation by transcription factors and microRNAs. *Science*, 319:1785–1786, 2008.

[12] C. A. Grove and A. J. Walhout. Transcription factor functionality and transcription regulatory networks. *Mol Biosyst*, 4:309–314, 2008.

[13] M. Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, 389:349–352, 1997.

[14] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128:693–705, 2007.

[15] C. B. Millar and M. Grunstein. Genome-wide patterns of histone modifications in yeast. *Nature Rev Mol Cell Biol*, 7:657–666, 2006.

[16] C. L. Liu, T. Kaplan, M. Kim, S. Buratowski, S. L. Schreiber, N. Friedman, and O. J. Rando. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol*, 3:e328, 2005.

[17] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309:626–630, 2005.

[18] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, DK Bailey, and et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120:169–181, 2005.

[19] F. Ozsolak, J. S. Song, X. S. Liu, and D. E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nature Biotech*, 25:244–248, 2007.

[20] J. J. Fischer, J. Toedling, T. Krueger, M. Schueler, W. Huber, and S. Sperling. Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics*, 91:41–51, 2008.

[21] B Meyer. Sex in the worm; counting and compensating X-chromosome dose. *Trends Genet*, 16:247–253, 2000.

[22] A. P. Wolffe and M. A. Matzke. Epigenetics: Regulation through repression. *Science*, 286:481–486, 1999.

[23] P. A. Jones and P. W Laird. Cancer epigenetics comes of age. *Nature Genet*, 21:163–167, 1999.

[24] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295:1306–1311, 2002.

[25] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genet*, 38:1348–1354, 2006.

[26] J. Dekker. Gene regulation in the third dimension. *Science*, 319:1793–1794, 2008.

[27] V. N. Kim. Small RNAs: classification, biogenesis, and function. *Mol Cells*, 19:1–15, 2005.

[28] D. H. Lackner and J. Bähler. Translational control of gene expression from transcripts to transcriptomes. *Int Rev Cell Mol Biol*, 271:199–251, 2008.

[29] K. Kapur, Y. Xing, Z. Ouyang, and W. H. Wong. Exon arrays provide accurate assessments of gene expression. *Genome Biol*, 8:R82, 2007.

[30] T. C. Mockler, S. Chan, A. Sundaresan, H. Chen, S. E. Jacobsen, and J. R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85:1–15, 2005.

[31] J. B. Fan, K. L. Gunderson, M. Bibikova, J. M. Yeakley, J. Chen, E. Wickham Garcia, L. L. Lebruska, M. Laurent, R. Shen, and D. Barker. Illumina universal bead arrays. *Meth Enzymol*, 410:57–73, 2006.

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

[33] R Lyne, G Burns, J Mata, CJ Penkett, G Rustici, D Chen, C Langford, D Vetrie, and J Bähler. Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, 4, 2003.

[34] G. G. Leparc, T. Tuchler, G. Striedner, K. Bayer, P. Sykacek, I. L. Hofacker, and D. P. Kreil. Model-based probe set optimization for high-performance microarrays. *Nucl Acids Res*, 37:12, 2009.

[35] G. K. Smyth and T. P. Speed. Normalization of cDNA microarray data. *Methods*, 31:265–273, 2003.

[36] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–S104, 2002.

[37] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl Acids Res*, 30:e15, 2002.

[38] J. van de Peppel, P. Kemmeren, H. van Bakel, M. Radonjic, D. van Leenen, and F. C. P. Holstege. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports*, 4:387–393, 2003.

[39] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol*, 14:1675–1680, 1996.

[40] Affymetrix Inc., Santa Clara, CA95051, USA (2002) Statistical Algorithm Description Document.

[41] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA*, 98:31–36, 2001.

[42] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucl Acids Res*, 31:e15, 2003.

[43] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

[44] D. M. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *J Comput Biol*, 8:557–569, 2001.

[45] M. J. Zilliox and R. A. Irizarry. A gene expression bar code for microarray data. *Nature Methods*, 4:911–913, 2007.

[46] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306, 2000.

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

[48] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309:626–630, 2005.

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

[50] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453:1239–1243, 2008.

[51] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev Genet*, 10:57–63, 2009.

[52] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*, 25:25–29, May 2000.

[53] M. Aslett and V. Wood. Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%. *Yeast*, 23:913–919, 2006.

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

[55] F. Schubert and J Bähler, 2008. Gene List Analyser 1.0, unpublished.

[56] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257–258, Jan 2007.

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

[58] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*, 13:662–672, Apr 2003.

[59] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl Acids Res*, 26:320–322, Jan 1998.

[60] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucl Acids Res*, 36:D281–288, Jan 2008.

[61] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucl Acids Res*, 37:D169–D174, 2009.

[62] Blastp is implemented as part of the NCBI Toolkit and was downloaded from ftp://ncbi.nlm.nih.gov/blast.

[63] S van Dongen. Graph clustering via a discrete uncoupling process. *Siam J Matrix Anal Appls*, 30:121–141, 2008.

[64] T. Schlitt and A. Brazma. Modelling in molecular biology: describing transcription regulatory networks at different scales. *Philos Trans R Soc Lond B*, 361:483–494, Mar 2006.

[65] B. Lehne and T. Schlitt. Protein-protein interaction databases: keeping up with growing interactomes. *Hum Genom*, 3:291–297, Apr 2009.

[66] J. I. Castrillo, L. A. Zeef, D. C. Hoyle, N. Zhang, A. Hayes, D. C. Gardner, M. J. Cornell, J. Petty, L. Hakes, L. Wardleworth, B. Rash, M. Brown, W. B. Dunn, D. Broadhurst, K. O'Donoghue, S. S. Hester, T. P. Dunkley, S. R. Hart, N. Swainston, P. Li, S. J. Gaskell, N. W. Paton, K. S. Lilley, D. B. Kell, and S. G. Oliver. Growth control of the eukaryote cell: a systems biology study in yeast. *J Biol*, 6:4, 2007.

[67] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863–14868, Dec 1998.

[68] G Sanguinetti, ND Lawrence, and M Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22:2775–2781, 2006.

[69] C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22:739–746, 2006.

[70] M. Barenco, E. Papouli, S. Shah, D. Brewer, C. J. Miller, and M. Hubank. rHVDM: an R package to predict the activity and targets of a transcription factor. *Bioinformatics*, 25:419–420, 2009.

[71] F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, 2004.

[72] J. V. Gray, G. A. Petsko, G. C. Johnston, D. Ringe, R. A. Singer, and M. Werner-Washburne. "Sleeping beauty": quiescence in Saccharomyces cerevisiae. *Microbiol Molec Biol Rev*, 68:187–206, 2004.

[73] J. Wu, N. Zhang, A. Hayes, K. Panoutsopoulou, and S. G. Oliver. Global analysis of nutrient control of gene expression in Saccharomyces cerevisiae during growth and starvation. *Proc Natl Acad Sci USA*, 101:3148–3153, 2004.

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

[75] M. J. Martinez, S. Roy, A. B. Archuletta, P. D. Wentzell, S. S. Anna-Arriola, A. L. Rodriguez, A. D. Aragon, G. A. Quiones, C. Allen, and M. Werner-Washburne. Genomic analysis of stationary-phase and exit in Saccharomyces cerevisiae: gene expression and identification of novel essential genes. *Mol Biol Cell*, 15:5295–5305, 2004.

[76] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7, 2006.

[77] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S. A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucl Acids Res*, 37:D868–872, 2009.

[78] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[79] G. Sanguinetti, M. Rattray, and N. D. Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22:1753–1759, 2006.

[80] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Methods. In *Machine Learning*, pages 183–233. MIT Press, 1998.

[81] B. Papp and S. Oliver. Genome-wide analysis of the context-dependence of regulatory networks. *Genome Biol*, 6:206, 2005.

[82] N. Zhang, J. Wu, and S. G. Oliver. Gis1 is required for transcriptional reprogramming of carbon metabolism and the stress response during transition into stationary phase in yeast. *Microbiology*, 155:1690–1698, 2009.

[83] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20 Suppl 1:i248–i256, 2004.

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

[85] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[86] Y. L. Yang and J. C. Liao. Network component analysis of Saccharamyces cerevisiae stress response. *Conf Proc IEEE Eng Med Biol Soc*, 4:2937–2940, 2004.

[87] A. L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor Biol Med Model*, 2:23, 2005.

[88] S. A. Tenenbaum, C. C. Carson, U. Atasoy, and J. D. Keene. Genome-wide regulatory analysis using en masse nuclear run-ons and ribonomic profiling with autoimmune sera. *Gene*, 317:79–87, 2003.

[89] J. E. Perez-Ortin, P. M. Alepuz, and J. Moreno. Genomics and gene transcription kinetics in yeast. *Trends Genet*, 23:250–257, 2007.

[90] J. Garcia-Martinez, A. Aranda, and J. E. Perez-Ortin. Genomic Run-on Evaluates Transcription Rates for All Yeast Genes and Identifies Gene Regulatory Mechanisms. *Molec Cell*, 15:303–313, 2004.

[91] S.-L. Jiang, D. Samols, D. Rzewnicki, S. S. MacIntyre, I. Greber, J. Sipe, and I. Kushner. Kinetic Modeling and Mathematical Analysis Indicate That Acute Phase Gene Expression in Hep 3B Cells Is Regulated by Both Transcriptional and Posttranscriptional Mechanisms. *J Clin Invest*, 95:1253–1261, 1994.

[92] M. M. Molina-Navarro, L. Castells-Roca, G. Belli, J. Garcia-Martinez, J. Marin-Navarro, J. Moreno, J. E. Perez-Ortin, and E. Herrero. Comprehensive transcriptional analysis of the oxidative response in yeast. *J Biol Chem*, 283:17908–17918, 2008.

[93] D. Cao and R. Parker. Computational modeling of eukaryotic mRNA turnover. *RNA*, 7:1192–1212, 2001.

[94] J. L. Hargrove, M. G. Hulsey, and E. G. Beale. The kinetics of mammalian gene expression. *Bioessays*, 13:667–674, 1991.

[95] J Ross. mRNA stability in mammalian cells. *Microbiol and Molec Biol Rev*, 59:423–450, 1995.

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

[97] E. Yang, E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco, and J. E. Darnell, Jr. Decay rates of human mRNAs: Correlation with functional characteristics and cequence attributes. *Genome Res*, 13:1863–1872, 2003.

[98] O Shalem, O Dahan, M Levo, M Rodriguez Martinez, I Furman, E Segal, and Y Pilpel. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular Systems Biology*, 4, 2008.

[99] C. Molin, A. Jauhiainen, J. Warringer, O. Nerman, and Sunnerhagen P. mRNA stability changes precede changes in steady-state mRNA amounts during hyperosmotic stress. *RNA*, 15:600–614, 2009.

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

[101] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S. Ko. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*, 16:45–58, 2009.

[102] J Grigull, S Mnaimneh, J Pootoolal, MD Robinson, and TR Hughes. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molec Cell Biol*, 24:5534–5547, 2004.

[103] C Cheadle, J. S. Fan, Y. S. Cho-Chung, T. Werner, J. Ray, L. Do, M. Gorospe, and K. G. Becker. Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC Genomics*, 6, 2005.

[104] J. Fan, X. Yang, W. Wang, W. H. Wood III, K. G. Becker, and M. Gorospe. Global analysis of stress-regulated mRNA turnover by using cDNA arrays. *Proc Natl Acad Sci USA*, 99:10611–10616, 2002.

[105] J. Legen, S. Kemp, K. Krause, B. Profanter, R. G. Herrmann, and R. M. Maier. Comparative analysis of plastid transcription profiles of entire plastid chromosomes from tobacco attributed to wild-type and PEP-deficient transcription machineries. *Plant J*, 31:171–88, 2002.

[106] J. W. Lilly, J. E. Maul, and D. B. Stern. The Chlamydomonas reinhardtii organellar genomes respond transcriptionally and post-transcriptionally to abiotic stimuli. *Plant Cell*, 14:2681–2706, 2002.

[107] F. J. Richards. A flexible growth function for empirical use. *J Exp Bot*, 10:290–300, 1959.

[108] S S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat*, 9:60–62, 1938.

[109] N. Goldman and S. Whelan. Statistical Tests of Gamma-Distributed Rate Heterogeneity in Models of Sequence Evolution in Phylogenetics. *Molec Bio Evol*, 17:975–978, 2000.

[110] Chen D., Wilkinson C. R., Watt S., C. J. Penkett, Toone W. M., Jones N., and Bähler J. Multiple pathways differentially regulate global oxidative stress responses in fission yeast. *Mol Biol Cell*, 19:308–317, 2008.

[111] Chen D., Toone W. M., Mata J., Lyne R., Burns G., Kivinen K., Brazma A., Jones N., and Bähler J. Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell*, 14:214–229, 2003.

[112] L. Lopez-Maury, S. Marguerat, and J. Bähler. Gene expression tuning to changing environments: from rapid responses to evolutionary adaptation. *Nature Rev Genet*, 9:583–593, 2008.

[113] D. A. Smith, W. M. Toone, D. Chen, J. Bähler, N. Jones, B. A. Morgan, and J. Quinn. The Srk1 protein kinase is a target for the Sty1 stress-activated MAPK in fission yeast. *J Biol Chem*, 277:33411–33421, 2002.

[114] G. Rustici, H. van Bakel, D. Lackner, F. Holstege, C. Wijmenga, J. Bähler, and A Brazma. Global transcriptional responses of fission and budding yeast to changes in copper and iron levels: a comparative study. *Genome Biol*, 8(R73), 2007.

[115] J. R. Warner. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*, 24:437440, 1999.

[116] M. Holcik and N. Sonenberg. Translational control in stress and apoptosis. *Nature Rev Mol Cell Biol*, 6:318–327, 2005.

[117] M. Brengues, D. Teixeira, and R. Parker. Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science*, 310:486–489, 2005.

[118] T. Kawai, T. Fan, K. Mazan-Mamczarz, and M. Gorospe. Global mRNA Stabilization Preferentially Linked to Translational Repression during the Endoplasmic Reticulum Stress Response. *Molecul Cell Biol*, 24:6773–6787, 2004.

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

[120] C. L. Lawrence, H. Maekawa, J. L. Worthington, W. Reiter, C. R. M. Wilkinson, and N. Jones. Regulation of *Schizosaccharomyces pombe* Atf1 protein levels by Sty1-mediated Phosphorylation and heterodimerization with Pcr1. *J Biol Chem*, pages 5160–5170, 2007.

[121] D. J. Hogan, D. P. Riordan, A. P. Gerber, D. Herschlag, and P. O. Brown. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*, 6(10:e255), 2008.

[122] D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding. Potent effect of target structure on microRNA function. *Nature Struct Molec Biol*, 14:287–294, 2007.

[123] R. A. Martienssen, M. Zaratiegui, and D. B. Goto. RNA interference and heterochromatin in the fission yeast *Schizosaccharomyces pombe*. *Trends Genet*, 21:450–456, Aug 2005.

[124] P. Provost, R. A. Silverstein, D. Dishart, J. Walfridsson, I. Djupedal, B. Kniola, A. Wright, B. Samuelsson, O. Radmark, and K. Ekwell. Dicer is required for chromosome segregation and gene silencing in fission yeast cells. *Proc Natl Acad Sci USA*, 99:16648–16653, 2002.

[125] M. Rabani, M. Kertesz, and E. Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA*, 105:14885–14890, 2008.

[126] R. Piessens, E. de Doncker-Kapenga, C. Uberhuber, and D. Kahaner. *Quadpack: a Subroutine Package for Automatic Integration*. Springer Verlag, 1983.

[127] S. Dorai-Raj. *powell: Powell's UObyQA algorithm*. R package version 1.0-0.

[128] N. A. Heard, C. C. Holmes, D. A. Stephens, D. J. Hand, and G. Dimopoulos. Bayesian Co-clustering of Anopheles Gene Expression Time Series: A Study of Immune Defense Response To Multiple Experimental Challenges. *Proc Natl Acad Sci USA*, 102:16939–16944, 2005.

[129] N. A. Heard, C. C. Holmes, and D. A. Stephens. A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *J Am Stat Ass*, 101:18–29, 2006.

[130] S. van Dongen, C. Abreu-Goodger, and A. J. Enright. Detecting microRNA binding and siRNA off-target effects from expression data. *Nature Methods*, 5:1023–1025, 2008.

[131] G. Rustici, J. Mata, K. Kivinen, P. Lió, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bähler. Periodic gene expression program of the fission yeast cell cycle. *Nature Genet*, 36:809–817, 2004.

[132] G. D. Stormo. An overview of RNA structure prediction and applications to RNA gene prediction and RNAi design. *Curr Protoc Bioinf*, Chapter 12:Unit 12.1, 2006.

[133] B. C. Foat and G. D. Stormo. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Molec Syst Biol*, 5:268, 2009.

[134] P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X. S. Liu, and J. Ahringer. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genet*, 41:376–381, Mar 2009.

[135] C. A. Cuomo et al. The Fusarium graminearum Genome Reveals a Link Between Localized Polymorphism and Pathogen Specialization. *Science*, 317:1400–1402, 2007.

[136] U. Güldener, G. Mannhaupt, M. Münsterkötter, D. Haase, M. Oesterheld, V. Stümpflen, H.-W. Mewes, and G. Adam. FGDB: a comprehensive fungal genome resource on the plant pathogen *Fusarium graminearum*. *Nucl Acids Res*, 34:D456–D458, 2006.

[137] Gene calls and functional annotation for *Fusarium graminearium*, versions FG1 and FG3, are available at the Broad Institute *Fusarium* Comparative Database. http://www.broad.mit.edu/annotation/genome/fusarium_graminearum/. Details of Fusarium gene calls are available at http://www.broad.mit.edu/annotation/fungi/fusarium/gene_finding.html.

[138] U. Güldener, K. Y. Seong, J. Boddu, S. Cho, F. Trail, J. R. Xu, G. Adam, H. W. Mewes, G. J. Muehlbauer, and H. C. Kistler. Development of a *Fusarium graminearum* Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fung Genet Biol*, 43:316–25, 2006.

[139] R. P. Wise, R. A. Caldo, L. Hong, L. Shen, E. K. Cannon, and J. A. Dickerson. *BarleyBase/PLEXdb: A Unified Expression Profiling Database for Plants and Plant Pathogens*, volume 406 of *Methods in Molecular Biology*, pages 347–363. Humana Press, Totowa, NJ., 2007. `http://www.plexdb.org`.

[140] H. E. Hallen, M. Huebner, S. Shin-Han, U. Güldener, and F. Trail. Gene expression shifts during perithecium development in *Gibberella zeae* (anamorph *Fusarium graminearum*), with particular emphasis on ion transport proteins. *Fung Genet and Biol*, 44:1146–1156, 2007.

[141] HE Hallen and F Trail. The l-type calcium channel, Cch1, affects ascospore discharge and mycelial growth in the filamentous fungus *Gibberella zeae* (anamorph *Fusarium graminearum)*. *Eukaryotic Cell*, 7:415–24, 2008.

[142] K. Y. Seong, X. Zhao, J. R. Xu, U. Guldener, and H. C. Kistler. Conidial germination in the filamentous fungus Fusarium graminearum. *Fung Genet and Biol*, 2007.

[143] *Fusarium graminearum* gene calls and associated annotation are available from the Omnimap FgraMap project. J. Antoniw, Kim Hammond-Kosack *et al.*. Rothamsted Research. http://www.omnimapfree.org.

[144] L. R. Gale, J. D. Bryant, S. Calvo, H. Giese, T. Katan, K O'Donnell, H. Suga, M. Taga, T. R. Usgaard, T. J. Ward, and H. C. Kistler. Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations. *Genetics*, 171:985–1001, 2005.

[145] S. Hunter, R. Apweiler, Attwood T. K., A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, Laugraud A., I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist,

M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. InterPro: the integrative protein signature database. *Nucl Acids Res*, 37:D224–228, 2009.

[146] A. Kauffman, R. Gentleman, and W. Huber. arrayQualityMetrics a Bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25:415–416, 2009.

[147] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K Hornik, T. Hothorn, W. Huber, S. Iacus, R Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80, 2004.

[148] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62:4427–33, 2002.

[149] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, and et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[150] G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[151] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Molec Biol*, 3: Iss. 1, Article 3., 2004.

[152] R. M. R. Coulson, N. Hall, and C. A. Ouzounis. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res*, 14:1548–1554, 2004.

[153] J. M. Carlton, J. H. Adams, J. C. Silva, S. L. Bidwell, H. Lorenzi, E. Caler, J. Crabtree, S. V. Angiuoli, E. F. Merino, P. Amedeo, Q. Cheng, R. M. R. Coulson, B. S. Crabb, H. A. del Portillo, K. Essien, T. V. Feldblyum, C. Fernandez-Becerra, P. R. Gilson, A. H. Gueye, X. Guo, S. Kang'a, T. W. A. Kooij, M. Korsinczky, E. V.-S. Meyer, V. Nene, I. Paulsen, O. White, S. A. Ralph, Q. Ren, T. J. Sargeant, S. L. Salzberg, C. J. Stoeckert, S. A. Sullivan, M. M. Yamamoto, S. L. Hoffman, J. R. Wortman, M. J. Gardner, M. R. Galinski, J. W. Barnwell, and C. M. Fraser-Liggett. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455:757–763, 2008.

[154] A. C. Ivens, C. S. Peacock, E. A. Worthey, L. Murphy, G. Aggarwal, and et al. The genome of the kinetoplastid parasite *Leishmania major*. *Science*, 309:436–442, 2005.

[155] Y. L. Bai, C. Salvadore, Y. C. Chiang, M. A. Collart, H. Y. Liu, and C. L. Denis. The CCR4 and CAF1 proteins of the CCR4-NOT complex are physically and functionally separated from NOT2, NOT4, and NOT5. *Mol Cell Biol*, 19:6642–6651, 1999.

[156] V. J. Promponas, A. J. Enright, S. Tsoka, D. P. Kreil, C. Leroy, S. Hamodrakas, C. Sander, and C. A. Ouzounis. CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, 16:915–22, 2000.

[157] An implementation of the Markov clustering algorithm was obtained from http://micans.org/mcl/.

[158] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucl Acids Res*, 30:1575–1584, 2002.

[159] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucl Acids Res*, 37:D5–15, 2009.

[160] The UniProt BLAST webservice is available at http://www.uniprot.org.

[161] Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, 29:1165–1188, 2001.

[162] K. De Preter, R. Barriot, F. Speleman, J. Vandesompele, and Y. Moreau. Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucl Acids Res*, 2008.

[163] J. C. Oliveros. Venny. an interactive tool for comparing lists with venn diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html., 2007.

[164] L. D. Hurst, C. Pál, and M. J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nature Rev Genet*, 5:299–310, 2004.

[165] S. C. Janga, J. Collado-Vides, and M. M. Babu. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc Natl Acad Sci USA*, 105:15761–15766, 2008.

[166] R. van Driel, P. F. Fransz, and P. J. Verschure. The eukaryotic genome: a system regulated at different hierarchical levels. *J Cell Sci*, 116:4067–4075, 2003.

[167] J. M. O'Sullivan, D. M. Sontam, R. Grierson, and B. Jones. Repeated elements coordinate the spacial organization of the yeast genome. *Yeast*, 26:125–138, 2009.

[168] E. Mancera, R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454:479–485, 2008.

[169] F. Kepes. Periodic transcriptional organization of the *E. coli* genome. *J Mol Biol*, 340:957–964, 2004.

[170] R. Hershberg, E. Yeger-Lotem, and H. Margalit. Chromosomal organization is shaped by the transcription regulatory network. *Trends in Genetics*, 21:138–142, 2005.

[171] B. A. Cohen, R. D. Mitra, J. D. Hughes, and G. M. Church. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet*, 26:183–186, 2000.

[172] P. T. Spellman and G. M. Rubin. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol*, 1:5, 2002.

[173] A. Purmann, J. Toedling, M. Schueler, P. Carninci, H. Lehrach, Y. Hayashizaki, W. Huber, and S. Sperling. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics*, 89:580–587, 2007.

[174] T. K. Baldwin, R. Winnenburg, M. Urban, C. Rawlings, J. Köhler, and K. E. Hammond-Kosack. The Pathogen-Host Interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity. *Mol Plant-Microbe Interact*, 19:1451–1462, 2006.

[175] R. J. N. Frandsen, N. J. Nielsen, N. Maolanon, J. C. Sorensen, S. Olsson, J. Nielsen, and H. Giese. The biosynthetic pathway for aurofusarin in *Fusarium graminearum* reveals a close link between the naphthoquinones and naphthopyrones. *Mol Microbiol*, 61:1069–1080, 2006.

[176] S. Malz, M. N. Grell, C. Thrane, F. J. Maier, P. Rosager, A. Felk, K. S. Albertsen, S. Salomon, L. Bohn, W. Schär, and H. Giese. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the *Fusarium graminearum* species complex. *Fung Genet and Biol*, 42:420–433, 2005.

[177] J-E. Kim, J. Jin, H. Kim, J-C. Kim, S-H. Yun, and Y-W. Lee. Gip2, a putative transcription factor that regulates the aurofusarin biosynthetic gene cluster in *Gibberella zeae*. *App Env Microbiol*, 72:1645–1652, 2006.

[178] I. Gaffoor, D. W. Brown, R. Plattner, R. H. Proctor, W. H. Qi, and F. Trail. Functional analysis of the polyketide synthase genes in the filamentous fungus *Gibberella zeae* (anamorph *Fusarium graminearum*). *Eukaryotic Cell*, 4:1926–1933, 2005.

[179] L. J. Harris, N. J. Alexander, A. Saparno, B. Blackwell, S. P. McCormick, A. E. Desjardins, L. S. Robert, N. Tinker, J. Hattori, C Piche, J. P. Schernthaner, R. Watson, and Ouellet T. A novel gene cluster in Fusarium graminearum contains a gene that contributes to butenolide synthesis. *Fung Genet Biol*, 44:293–306, 2007.

[180] E. Lysøe, K. R. Bone, and S. S. Klemsdal. Real-time quantitative expression studies of the zearalenone biosynthetic gene cluster in *Fusarium graminearum*. *Phytopath*, 99:176–184, 2009.

[181] S. P. McCormick, L. J. Harris, N. J. Alexander, T. Ouellet, A. Saparno, S. Allard, and A. E Desjardins. Tri1 in *Fusarium graminearum* encodes a p450 oxygenase. *Appl Env Microbiol*, 70:2044–2051, 2004.

[182] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet*, 25:25–9, 2000.

[183] C. Pál and L. D. Hurst. Evidence for co-evolution of gene order and recombination rate. *Nature Genet*, 33:392–395, 2003.

[184] M. J. Lercher and L. D. Hurst. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*, 18:337–340, 2002.

[185] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, May 2007.

[186] K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, 2:e36, 2006.

[187] L.-J. Ma, H. C. van der Does, K. A. Borkovich, J. J. Coleman, M.-J. Daboussi, A. Di Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrissat, P. M. Houterman, S. Kang, W.-B. Shim, C. Woloshuk, X. Xie, J.-R. Xu, J. Antoniw, S. E. Baker, B. H. Bluhm, A. Breakspear, D. W. Brown, R. A. E. Butchko, S. Chapman, R. M. R. Coulson, P. M. Coutinho, E. G. J. Danchin, A. Diener, L. R. Gale, D. M. Gardiner, S. Goff, K. E. Hammond-Kosack, K. Hilburn, P. M. Houterman, A. Hua-Van, W. Jonkers, K. Kazan, C. D. Kodira, M. Koehrsen, L. Kumar, Y.-H. Lee, L. Li, J. M. Manners, D. Miranda-Saavedra, M. Mukherjee, G. Park, J. Park, S.-Y. Park, R. H. Proctor, A. Regev, M. C. Ruiz-Roldan, D. Sain, S. Sakthikumar, S. Sykes, D. C. Schwartz, B. G. Turgeon, I. Wapinski, O. Yoder, S. Young, Q. Zeng, S. Zhou, J. Galagan, C. A. Cuomo, H. C. Kistler, and M Rep. Fusarium comparative genomics reveals pathogenicity related lineage-specific genome expansion. *Nature*, in press.

[188] C. Hertz-Fowler, C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A. C. Ivens, M. A. Rajandream, and B. Barrell. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucl Acids Res*, 32:D339–D343, 2004.