Comprehensive analysis of high-throughput experiments for investigating transcription and transcriptional regulation

Joern Michael Toedling

Jesus College

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

European Molecular Biology Laboratory,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SD,
United Kingdom.

Email: toedling@ebi.ac.uk

12 January 2009

Meinen Eltern

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

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

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

12 January 2009 Joern Michael Toedling

Comprehensive analysis of high-throughput experiments for investigating transcription and transcriptional regulation

Summary

Joern Michael Toedling
12 January 2009 Jesus College

As the number of fully sequenced genomes grows, efforts are shifted towards investigation of functional aspects. One research focus is the transcriptome, the set of all transcribed genomic features. We aspire to understand what features constitute the transcriptome, in which context these are transcribed and how their transcription is regulated. Studies that aim to answer these questions frequently make use of high-throughput technologies that allow for investigation of multiple genomic regions, or transcribed copies of genomic regions, in parallel.

In this dissertation, I present three high-throughput studies I have been involved in, in which data gained from oligo-nucleotide tiling microarrays or large-scale cDNA sequencing provided insights into the transcriptome and transcriptional regulation in the model organisms *Saccharomyces cerevisiae* and *Mus musculus*. Interpretation of such high-throughput data poses two major computational tasks. The primary statistical analysis includes quality assessment, data normalisation and identification of significantly affected targets, i.e. regions of the genome deemed transcribed or involved in transcriptional regulation. Second, in an integrative bioinformatic analysis, the identified targets need to be interpreted in context of the current genome annotation and related experimental results. I provide details of these individual steps as they were conducted in the three studies.

For both primary and integrative analysis, functional, extensible and well-documented software is required, which implements individual analysis steps, allows for concise visualisation of intermittent and final results and facilitates the construction of automated, programmed workflows. Ideally such software is optimised with respect to scalability, reproducibility and methodical scope of the analyses. This dissertation contains details of two such software packages in the Bioconductor project, which I (co-)developed.

Preface

This dissertation describes work carried out at the European Bioinformatics Institute (EBI) in Cambridge, UK, between March 2005 and June 2008. The EBI is an outstation of the European Molecular Biology Laboratory (EMBL). I was the recipient of an EMBL predoctoral fellowship to work together with Wolfgang Huber.

In the course of collaborations, I also worked together with Lars Steinmetz, Zhenyu Xu, Marina Granovskaia, Antje Purmann and Tammo Krueger, and I would like to thank them for many constructive discussions, opinions and ideas. Credit is also due to Silke Sperling and Jenny Fischer for the successful collaborations.

The following people helped me to make my time at the EBI interesting, instructional, exciting, and enjoyable:

Wolfgang, vielen Dank dafür, dass Du mir die Möglichkeit gegeben hast, von Dir zu lernen und in Deiner Gruppe zu arbeiten. Es war nicht immer einfach, aber immer spannend, und ich kann sagen, dass ich am Ende viel aus der Zusammenarbeit mit Dir mitnehmen kann. Ich habe viel gelernt von Dir, über die Analyse von high-throughput Daten im Allgemeinen, über statistische Herangehensweisen und natürlich über R. Für Deinen Beistand in der anstrengenden Zeit des HeartRepair-Dilemmas möche ich Dir besonders danken. Und vielleicht bin ich mittlerweile sogar bereit einzusehen, dass auch Diskussionen über Aspect Ratios eine geringe Daseinsberechtigung haben könnten.

Many thanks to Sarah Teichmann, Alvis Brazma, and Lars Steinmetz for their expert advice in and after my thesis advisory committee meetings.

Ich möchte an dieser Stelle auch Rainer Spang und Martin Vingron danken dafür, dass sie mein Interesse an der Bioinformatik geweckt und gefördert haben.

Katherine, thank you so much for carefully proofreading this manuscript and for your wonderful suggestions for improving it. I owe you, and I promise that I won't carry about with putting the preposition "on" everywhere. Also, many thanks for your friendship over the last four years.

Special thanks to all members of the Huber group, many as they were in such a dynamic group. Elin, thanks for countless coffees and for being such a good friend. I hope that you will find happiness and that someday you can also look back and think that it was not too bad a time at EBI. Oleg, thanks for always being very helpful. I knew I could always talk to you about all sorts of things, work-wise and otherwise. It was great to have you there and a sad day when you decided to move on. Tine, many thanks for the shared time, short as it was. Your excitement about all aspects of your work, and be it only the poplar genome sequence, was very contagious, and you were even more fun when you could let go. Greg, your remarks were always very entertaining, first because I could not understand them, later on because they were genuinely true and funny. Richard, thanks for patiently putting up with lots of questions about statistics and ChIP-chip. Olá Lígia, muito obrigado for the Portuguese classes. Kristen, thank you very much for proofreading the two chapters and for providing elegant English phrases. Audrey, Raeka, Ramona, Alessandro, Bernd, David, Florian, Matt, Mike, Paul, Remy, Simon, Thomas, and Tony, thanks for the nice and interesting chats and the good time.

I am also very grateful to all members of the EBI predoc community and associates for providing entertainment and distraction in the frequent times of need, especially Michael, Daniela, Jacky, John, Daniel, and Melanie. Without you guys, it would have been four very hard and boring years indeed.

Vielen Dank auch meinen Freunden in Berlin, Oxford und Bremen, auf die ich trotz der Entfernung doch immer zählen konnte, speziell meiner Schwester Hanna, Peer, Andreas, und Janine.

Der größte Dank geht an meine Eltern. Danke für alles, und speziell dafür, dass Ihr mich in all den Jahren des Studierens in Berlin und Cambridge immer bedingungslos unterstützt habt! Ich habe es Euch bestimmt nicht immer leicht gemacht, meinen Weg gutzuheißen, aber ich hoffe und denke, daß Ihr mit dem Ergebnis zufrieden seid.

Contents

Summary **i**

Preface **ii**

Contents **iv**

List of Figures **vii**

List of Tables **ix**

List of Acronyms **x**

Glossary **xii**

Concerning gene and protein names xiii

1 Introduction **1**

1.1 Transcription 2

1.2 Regulation of transcription 3

1.2.1 Histone modifications 4

1.2.2 Transcription factors 5

1.3 Transcript orientation 6

1.4 DNA microarrays 6

1.4.1 ChIP-chip 13

1.5 Microarray-based investigation of transcription throughout the cell cycle 14

1.5.1 Budding yeast cell cycle 14

1.5.2 Microarrays and the cell cycle 16

1.6 Gene annotation 18

2 Coexpression of adjacent genes 20

2.1 Introduction 20

2.1.1 Gene clusters 20

2.1.2 Measures of gene coexpression among clusters 21

2.2 Material and Methods 22

2.3 Results 25

2.3.1 Permutation scheme for estimating the null distribution of coexpression scores 25

2.3.2 Chromosomal clustering of transcriptomes 25

2.3.3 Evaluating coexpression across tissues 26

2.3.4 Highly coexpressed clusters 33

2.3.5 HCPs and housekeeping functionality 33

2.3.6 Location, orientation, and dimension of HCCs 33

2.3.7 Functionality, paralogy and transcriptional regulation of highly coexpressed gene clusters 34

2.3.8 *H. sapiens* microarray data 36

2.4 Discussion 38

3 Transcriptional regulation in developing cardiomyocytes 42

3.1 Histone modifications 43

3.1.1 Introduction 43

3.1.2 Material and methods 46

3.1.3 Results 55

3.1.4 Discussion 62

3.2 Transcription factor binding events 67

3.2.1 Introduction 67

3.2.2 Material and methods 68

3.2.3 Results 70

3.2.4 Discussion 74

4 A high-resolution view of transcription in budding yeast 77

4.1 Introduction 77

4.2 Redefining the transcriptome 78

4.2.1 Material and methods 79

4.2.2 Results 85

4.2.3 Discussion 89

4.3 Cell-cycle data 91
4.3.1 Methods 92
4.3.2 Results 97
4.3.3 Discussion 112

5 Software for reproducible research 118

5.1 *tilingArray* 119
5.1.1 Introduction 119
5.1.2 Reporter annotation 120
5.1.3 Data preprocessing 121
5.1.4 Segmentation 123
5.1.5 Visualisation of segments 129
5.1.6 Discussion 129
5.2 *Ringo* 131
5.2.1 Introduction 131
5.2.2 Implementation 133
5.2.3 Functionality 133
5.2.4 Discussion 140

Conclusions 142

A Publications 144

Bibliography 145

List of Figures

1.1 Possible orientation of adjacent transcript pairs 7
1.2 Reporter level versus GC content 11
1.3 *S. cerevisiae* cell cycle 15

2.1 Fantom3: bar plot of Hamming distances 27
2.2 Fantom3: bar plot of scaled Hamming distances 29
2.3 Fantom3: empirical *p*-values for gene pair coexpression . . . 32
2.4 Shared tissue expression of highly coexpressed gene pairs in *M. musculus* 34
2.5 Size of gene coexpression clusters in *M. musculus* 35
2.6 Symatlas *H. sapiens*: empirical *p*-values for gene pair coexpression 37

3.1 Reporters per transcript on the expression microarray 51
3.2 Surface of a ChIP-chip microarray showing hybridisation artifacts 56
3.3 Histone modifications at the TSS of *Hand2* 57
3.4 Combinations of histone modifications in HL1 cells 58
3.5 Relationship between histone modifications and gene expression 60
3.6 Heart transcription factor ChIP enrichment at the TSSs of *Tbx20* and *Tbx3* 72
3.7 Heart transcription factor network from ChIP-chip data . . . 73
3.8 Histone modifications and transcription factor binding sites (TFBSs) at the TSS of *Hand2* 74

4.1 High resolution *S. cerevisiae* tiling array 78
4.2 *S. cerevisiae* tiling array: Normalisation using a direct DNA hybridisation 80

4.3 *S. cerevisiae* tiling array: Threshold for transcribed segments 82
4.4 UTR lengths by GO category 87
4.5 *S. cerevisiae* tiling array: unannotated transcripts 88
4.6 Regularised correlation coefficient 94
4.7 Histone expression along the cell cycle 99
4.8 Correlated sense/antisense pairs 100
4.9 Hierarchical clustering dendrogram of periodic transcripts . 103
4.10 Heatmap of periodic transcripts 104
4.11 Bidirectional promoter 108
4.12 Distance between adjacent transcripts by orientation 109
4.13 *S. cerevisiae*: Relation between transcript orientation and correlation of expression 110

5.1 *tilingArray* segmentation algorithm compared with sliding average thresholding 127
5.2 *tilingArray*: visualisation of a genomic region 130
5.3 Diagram: ChIP-chip workflow facilitated by *Ringo* 134
5.4 *Ringo*: example autocorrelation plot 135
5.5 *Ringo*: Visualising reporter levels 136
5.6 *Ringo*: Smoothing of reporter levels 138
5.7 *Ringo*: Finding ChIP-enriched regions 140

List of Tables

1.1 Gene Ontology evidence codes 19

2.1 Fantom3: Numbers and lengths of HCCs 33
2.2 Functional and transcriptional properties of genomic neighbours in *M. musculus* 36
2.3 Functional and transcriptional properties of genomic neighbours in *H. sapiens* 38

3.1 Codes of combined ChIP-enriched regions 54
3.2 ChIP-enriched regions for histone modifications in three cell types 57
3.3 Linear model: transcript expression related to histone modifications 61
3.4 Heart TFs: ChIP-enriched regions 71

4.1 Yeast exponential growth transcripts 86
4.2 Yeast cell cycle transcripts 98
4.3 *S. cerevisiae*: 37 periodic antisense transcripts 101
4.4 *S. cerevisiae*: TFBSs in clusters 105
4.5 *S. cerevisiae*: GO terms in clusters 106

List of Acronyms

AS	antisense transcript
BIC	Bayesian Information Criterion
bp	base pair
CAGE	cap analysis of gene expression
CC	correlation coefficient
cDNA	complementary DNA
CGH	comparative genomic hybridisation
ChIP	chromatin immunoprecipitation
EST	expressed sequence tag
FDR	false discovery rate
GC content	guanine-cytosine content
GIS	gene identification signature
glog	generalised logarithm: $\mathrm{glog}_\Delta(x) = \log\left(x + \sqrt{x^2 + \Delta}\right)$
GO	Gene Ontology
GSC	gene signature cloning
H3ac	histone H3, acetylated at lysine residues 9 and/or 13
H3K4me2	histone H3, di-methylated at lysine residue 4
H3K4me3	histone H3, tri-methylated at lysine residue 4
H4ac	histone H4, acetylated at lysine residues 5, 8, 12, and/or 16
HCC	highly co-expressed cluster of genes
HCP	highly co-expressed pair of genes
HD	Hamming distance
kb	1 kilobase $\equiv$ $1,000$ base pairs
HMM	Hidden Markov model

i.i.d.	independently identically distributed
Mb	1 Megabase $\equiv 1,000,000$ base pairs
MM	mismatch
ncRNA	non-(protein-)coding RNA
NFR	nucleosome-free region
ORF	open reading frame
PCR	polymerase chain reaction
PM	perfect match
Pol II	RNA Polymerase II
PSSM	position specific score matrix
RMA	robust multiarray analysis
RMP	reporter match position
RNAi	RNA interference
SAT	sliding average thresholding
SGD	Saccharomyces Genome Database
siRNA	small interfering RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site
TU	transcriptional unit
UCP	uncorrelated pair of genes
UCC	uncorrelated consecutive genes
UI	unannotated intergenic
uORF	upstream open reading frame
UTR	untranslated region (of an mRNA)
vsn	variance-stabilising normalisation
25mer	polymer of 25 nucleotides linked by phosphodiester bonds
60mer	polymer of 60 nucleotides linked by phosphodiester bonds

Glossary

The following terms are used throughout this dissertation.

The terms *microarray* and *array* are used interchangeably.

Reporters are the DNA sequences affixed to a microarray for measuring the abundance of complementary *mRNA* fragments with expression microarrays or complementary sonicated DNA fragments in ChIP-chip experiments. Each reporter is assumed to have a unique identifier and a unique sequence, but can appear as multiple *features* on the array surface. Note that in the majority of the microarray-related literature, a reporter is called a *probe*, but since that term does not differentiate between reporter and feature, the term *reporter* is preferable.

Genome segments that (are considered to) match the sequence of a reporter are called the *match positions* of that reporter. A *unique reporter match position* (RMP) is an RMP, which is the only match of that reporter in the genome.

A *genomic region* is a segment of the genome.

The *sample* is the aliquot of isolated RNA, or immuno-precipitated or input DNA that is hybridised to the microarray. A genomic region apparently enriched by ChIP is called a *ChIP-enriched region*.

The primary data gained from scanning the microarray and processing the scanned image is called the *raw* data, which in most cases requires preprocessing, or *normalisation*, to improve the signal-to-noise ratio.

An *expression profile* refers to the vector of expression levels of one transcript in a number of conditions or samples.

A transcriptional unit (TU) is a set of (m)RNAs that share at least one transcribed genomic nucleotide and are transcribed from the same segment of the genome in the same orientation.

Concerning gene and protein names I follow the organism-specific conventions for writing gene and protein names in this dissertation. With *Mus musculus*, gene symbols are italicised, with the first letter in upper case and the remaining letters in lower case (*Gata4*). *M. musculus* protein names are the same as the gene symbol, but are not italicised and all letters are in upper case (GATA4)[1]. With *Saccharomyces cerevisiae*, gene symbols are italicised and all letters are in upper case (*STB1*). *S. cerevisiae* protein names are the same as the gene name, but are not italicised, and the first letter is upper case and the remaining letters are lower case (Stb1)[2].

	Gene name	**Protein name**
M. musculus	*Gata4*	GATA4
S. cerevisiae	*STB1*	Stb1

[1] Mouse Genome Informatics (MGI) Nomenclature guide:
`http://www.informatics.jax.org/mgihome/nomen/short_gene.shtml`

[2] Saccharomyces Genome Database (SGD) naming guidelines:
`http://www.yeastgenome.org/gene_guidelines.shtml`

Chapter 1

Introduction

This dissertation describes three high-throughput studies, in which data gained from oligo-nucleotide tiling microarrays or large-scale cDNA sequencing was used to obtain new insights into the transcriptome and transcriptional regulation. I have mostly worked with data from the model organisms *Saccharomyces cerevisiae* and *Mus musculus*, but one study also included an analysis of a data set from *Homo sapiens*.

In this chapter, I will give a general and brief overview of the biological concepts, the microarray technologies, and the relevant annotation data which I have investigated and used in the projects that are described in this dissertation. The following chapters describe each individual study, and each chapter contains a more specific introduction covering the background information that is pertinent to the respective study.

Chapter 2 describes an investigation of coexpression and co-regulation of genes that are adjacent to each other in the genome.

Chapter 3 describes the use of ChIP-chip and expression microarray technology to elucidate the role of four post-translational histone modifications and four transcription factors in the development of heart and muscle cells.

In Chapter 4, I detail the analysis of tiling microarray data to characterise the whole transcriptome of *S. cerevisiae* during exponential growth and in the course of the cell cycle.

Finally, Chapter 5 describes two software packages that I have written or co-authored to facilitate the analyses of the high-throughput studies presented in this dissertation.

1.1 Transcription

Transcription[1] is the process in which a polymerase enzyme produces an RNA complementary to a stretch of genomic DNA. The apparent purpose of transcription is to either translate the information from the DNA into a protein or relay the information in the form of RNA to cellular processes. Here, I will briefly outline the current – and incomplete – understanding of transcription and transcriptional regulation, as these concepts will be referred to extensively in the following chapters (for a more detailed review, please refer to [1] and references therein).

Transcription is a three-step process [2], consisting of:

initiation the polymerase enzyme is recruited to the transcript's transcription start site (TSS) by the *pre-initiation complex* of general transcription factors and starts the complementary RNA with the first nucleotide

elongation the polymerase moves along the transcript's DNA template and consecutively extends the complementary RNA

termination the polymerase stops and detaches from the DNA, and the RNA product is released and post-processed.

The enzymes that are responsible for transcription are DNA-dependent RNA polymerases. Several different polymerases, involved in transcription of different kinds of RNA, have been identified:

- *RNA polymerase I* transcribes a large 45S precursor RNA [3] that splits into three ribosomal RNAs (28S, 18S and 5.8S rRNA).

[1] This dissertation will solely deal with transcription in eukaryotes.

- *RNA polymerase II* is involved in the transcription of all pre-mRNAs [3], snRNAs [4], and miRNAs [5].

- *RNA polymerase III* transcribes all tRNAs, 5S rRNA, and other small ncRNAs [6, and references therein].

- In plants, *RNA polymerase IV* is involved in the synthesis of siRNAs. Other types of RNA polymerase have been found in mitochondria and plant chloroplasts.

The existence of transcribed non-coding RNAs (ncRNAs) was hypothesised as early as 1958 [7], but until the late 1980s, tRNA and rRNA were the only well-known types of ncRNA. Since then, many more forms of ncRNA have been identified, such as miRNA, snRNA, and snoRNA. Over the last few years, using high-throughput microarray and sequencing experiments, eukaryotic cells have been found to contain many more unexpected RNAs not coding for proteins. Most of these non-coding RNAs are currently of unknown function (e.g., [8, 9, 10]). At present, the majority of RNAs produced in the cell are presumed to be non-coding [11].

Chapter 4 describes one tiling microarray study that allowed for an unbiased survey of the complete transcriptome of *S. cerevisiae* at high resolution. In this study, we identified multiple new and unexpected, presumably non-coding, RNAs in the transcriptome of budding yeast.

1.2 Regulation of transcription

Before a segment of the genome can be transcribed, three events have to happen. First, the genomic DNA in the highly condensed chromatin must be made sufficiently accessible to serve as a template for complementary base pairing. Second, the polymerase must be recruited to the transcription start site of the DNA segment to be transcribed. Thirdly, the polymerase must be enabled to move along the DNA of the target to produce the complementary RNA.

1.2.1 Histone modifications

A central structure of chromatin organisation is the nucleosome, a stretch of 146–147 bp of DNA wrapped around a complex of eight histone proteins. The eight histone proteins comprise two copies each of four different proteins (H2A, H2B, H3 and H4). The N-terminal tails of these proteins protrude out of the nucleosome, allowing them to interact with molecules outside their own nucleosome. The histone tails have been shown to be involved in inter-nucleosomal interactions that stabilise higher order structures of the chromatin [12], although the nature of these higher order chromatin structures is still poorly understood. Furthermore, several enzymes have been discovered to covalently attach chemical groups to specific residues in the N-terminal tails of histone proteins, or to remove these attached groups. Among such covalent modifications described are acetylation, methylation, phosphorylation, ubiquitination, and others (see [13] for a recent review).

Many of these histone modifications have been associated with changes in the transcription rate of genes located next to modified nucleosomes (see [14, 15] for examples). *Activating* modifications are those which seem to coincide with increased transcription, while *repressive* modifications are associated with reduced gene expression.

Histone modifications may affect chromatin structure in two ways. First, they may influence the interactions between histone proteins and DNA, as well as the interaction between adjacent nucleosomes. Both types of interactions are important for regulating the chromatin structure. Acetylation of lysine residues in the histone tails has been observed to have an activating effect, which can be explained by a change in charge. The addition of the acetyl group neutralises the positive charge of the lysine residue. The affinity of the histone protein to the negatively charged backbone of the DNA is lowered, and the chromatin structure is opened up [16]. Second, modified histones are known to specifically recruit non-histone proteins, such as chromatin remodelling enzymes that modify the chromatin structure (reviewed in [13]). It has been hypothesised that different combinations of histone modifications may recruit different kinds of target proteins and thus bring about distinct follow-up effects [17]. This “histone code” hypothesis is currently widely debated, and there exist pieces of evidence both in favour and against. Nevertheless, if such a code exists, much further work is required towards understanding how this code is deciphered in the cell.

Chapter 3 describes the analysis and results of ChIP-chip experiments, in which we investigated the role of four histone modifications in transcriptional regulation, and relates our findings to the histone code hypothesis.

Histone modifications have been shown to be at least partially maintained throughout cell divisions by as yet unknown mechanisms [18], and are therefore considered “epigenetic” factors, since they provide heritable information that is beyond the nucleotide sequence of the DNA. The inheritance of histone modifications is similar to, and likely linked to, the inherited methylation of cytosines in CpG islands.

1.2.2 Transcription factors

Certain proteins, collectively known as transcription factors (TFs), bind to the promoter region or other regulatory regions close to the target TSS and help to activate or repress the transcription process. TFs can bind directly to the DNA, to DNA-associated histone proteins, or to other transcription factors already bound to DNA or histone proteins.

General, or *basal*, transcription factors are expressed in a multitude of different cell types, bind to many different kinds of promoter regions, and are involved in transcription of many different RNAs (see [19] for a recent review). Examples of such general transcription factors are the components of the pre-initiation complex that recruits the RNA polymerase II to the TSSs of the genomic region to be transcribed. Specialised transcription factors, on the other hand, regulate the activity of the polymerase at each stage of the transcription process, or regulate the function of the general transcription factors. These specialised TFs are only expressed in certain cell types and bind only to specific regulatory regions that can often be characterised by specific DNA sequence motifs.

1.3 Transcript orientation

Groups of adjacent transcripts have frequently been shown to be coexpressed across tissues and/or time (e.g., [20]). Figure 1.1 depicts the principal types of orientation that two adjacent transcripts can have in relation to each other. The orientation may provide clues as to how the transcription of the two transcripts is coupled. For example, for pairs of *divergent transcripts*, the common regulation of transcription through bidirectional promoter regions has been well characterised in *H. sapiens* [21].

In Chapter 4, I present a study about the orientation of adjacent transcripts that are expressed throughout the cell cycle of *S. cerevisiae*.

1.4 DNA microarrays

Microarrays are devices for measuring the abundance of multiple RNAs or DNAs in a sample in parallel. The most common application of microarrays is to simultaneously measure the mRNA abundance of transcripts of multiple genes [22]. Microarrays produced for this purpose are commonly called expression microarrays. More recent alternative uses of microarrays, such as in array comparative genomic hybridisation (CGH) require a more general description of the potential applications of microarrays.

During the last decade, a number of different microarray platforms have been established. Among these different platforms, oligo-nucleotide microarrays such as the ones manufactured by Affymetrix [23] are probably the most commonly used. Many techniques to reduce and control the inherent noise of microarray gene expression data [24, 25, and others] have been developed, establishing microarray expression data as a reliable source of information about transcript abundance. Statistical meth-

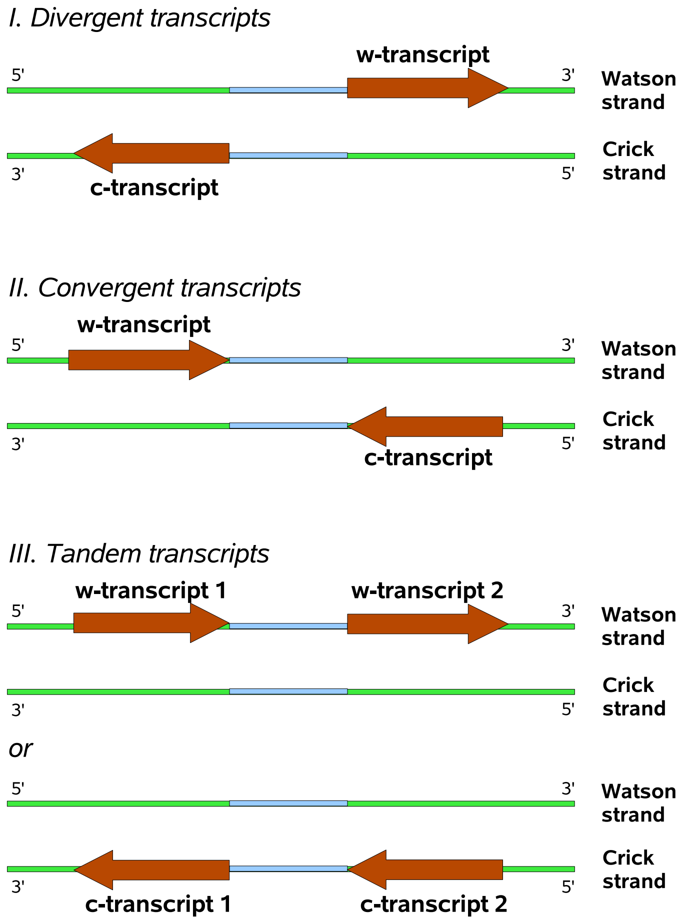


Figure 1.1: *Scheme of possible orientations, in which two adjacent genomic regions can be transcribed.*

ods have been developed for the identification of differentially expressed genes under different conditions [26, 27], as well as methods for other applications for microarrays.

cDNA microarrays The classical type of microarrays uses polymerase chain reaction (PCR) products, hence the name "cDNA" relating to complementary DNA fragments, of about 1 kb length as reporters [22]. The reporters represent transcripts of genes and are usually derived from ESTs and mRNAs. A sample RNA of interest, for example RNA obtained from tumour cells, is labelled with one dye, such as Cy5, and is hybridised to the microarray together with a control RNA sample, which is labelled with another dye, such as Cy3. The RNA from the sample of interest and the control RNA compete for binding to the reporters that are fixed on to the microarray surface.

Because of the use of two dyes, the term two-channel microarrays, or two-colour microarrays, has been used synonymously with cDNA arrays. By now, however, experiments in which two samples labelled with different dyes are hybridised to oligonucleotide microarrays are also common.

The reporter level of two-colour microarrays is usually specified as the ratio or fold change between the two individual channel intensities.

Oligonucleotide expression microarrays With this type of microarray, each transcript is represented by a number n (typically $n \in [11, 20]$) of oligonucleotide reporters of 24 to 60 nucleotides in length [23]. The reporters are usually complementary in sequence to segments at the 3′ end of the transcribed regions of the genes. The expression level of a transcript is derived by summarising over the n individual reporters.

In the traditional design used by Affymetrix (Santa Clara, California, USA), a transcript is represented by a "probe set" of $2 \cdot n$ reporters of length 25. Of these, n are perfect match (PM) reporters that match the sequence of the transcript to 100%. In addition, there are an equal number of mismatch (MM) reporters, which have the same sequences as the PM reporters except for the nucleotide in the centre of each reporter, which is replaced by its complementary base. The MM reporters were originally intended to provide reliable estimates of the reporter background levels. The usefulness of mismatch reporters for this purpose, however, has been doubted, and many approaches to the analysis of oligonucleotide microarray data now disregard MM reporters [24].

Tiling microarrays Improvements in techniques for the production of microarrays have resulted in an extension of the microarray types described above. A key development was the reduction of the size of each feature, that is the set of copies of one reporter affixed at one position on the microarray surface. As a result, genomic tiling microarrays can include target regions covering a whole genome, not just the coding sequences of genes [28]. A specialised (and older) type of tiling microarray has only selected regions of the genome represented by reporters. Promoter tiling arrays are a common example of microarrays of this design.

With arrays that should represent the whole genome, reporter selection of an equal standard to commercial or self-spotted expression microarrays is not feasible. The inclusion of intergenic regions into the target sequences, to be represented by reporters on the array, results in reporters showing sequence characteristics of these regions, which are less homogeneous than the characteristics of annotated open reading frame (ORF) sequences. The reporters on tiling arrays thus show substantial variation in their physical characteristics regarding staining and hybridisation of these reporters and their targets. An obvious source of variance in reporter hybridisation is the varying GC content of the reporter sequences [29]. Because of repetitive elements in the genome, tiling array reporters also differ largely in the specificity of reporter matches to the genome.

The analysis of tiling-microarray data poses specific challenges, due to the high density of reporters and the massive number of genomic regions to be represented by reporters. One of these challenges is dealing with reporter specificity and differential reporter response due to sequence characteristics. Another challenge is to merge the reporter levels from single reporter match positions (RMPs) into segments that correspond to genomic regions of interest, such as transcribed regions or ChIP-enriched regions [30].

A major application of whole-genome tiling microarrays is unbiased transcriptome analysis, meaning the detection of transcribed regions beyond previously annotated coding sequences (e.g., [8, 9]). From the first transcriptome studies of this kind, it has become obvious that in many higher organisms there is a large amount of transcription beyond what was known and expected from annotated genome elements and gene predictions. Many of these new transcripts do not contain an ORF of more than a few residues in length and are therefore considered to be non-protein-coding RNAs (ncRNAs). The exact mappings, architectures, functions, and significance of the majority of these non-coding transcripts are yet to be determined [28].

Other applications of whole-genome tiling microarrays are ChIP-chip (see Section 1.4.1 and, e.g., [31]), and assaying DNA copy number variations (array-CGH, e.g., [32]). Tiling microarrays can also be used to get a high-resolution view on other cellular processes that affect the DNA or RNA, such as recombination events [33].

Exon microarrays The commonly used oligonucleotide expression microarrays contain reporters that are complementary in sequence to segments at the 3′ end of the translated regions of the genes [23]. The reason for this placement of reporters is that the reverse transcription of mRNA into cDNA (or cRNA), which is then hybridised to the microarray, starts from the poly-A tail of the mRNA. The reverse-transcribing polymerase generates complementary copies to the 3′ end of the translated region of an average length of 1000 bp [23]. Hence, common expression microarrays are only able to distinguish transcripts of the same gene if the transcripts differ in their last few included exons (627 bp is the average length of exons in *M. musculus*, as annotated in the Ensembl database [34], release 50, July 2008). An intermittent step between such traditional expression arrays and whole-genome tiling microarrays are exon arrays. The reporters on this type of microarray represent all annotated individual exons of transcripts. For example, on the exon arrays that are produced by Affymetrix, each exon is represented by four reporters. Exon microarrays provide summarised expression levels for individual exons and allow for hypotheses about the expression of distinct, alternatively spliced, transcripts of a gene [35].

BeadArrays Another type of microarray are the BeadArrays manufactured by Illumina Inc. [36]. With this microarray design, each transcript is only represented by one or two reporters of 50 nucleotides length, but

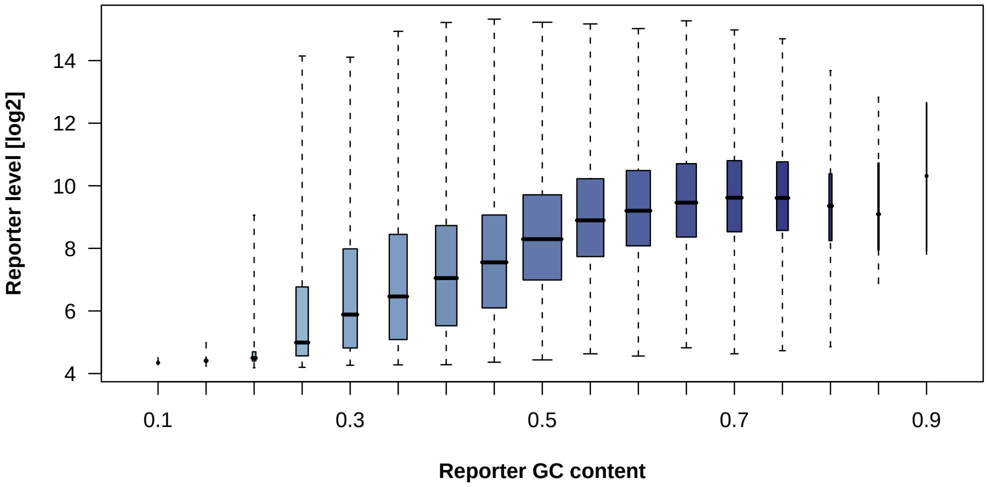


Figure 1.2: *Boxplots showing the typical distribution of normalised reporter levels versus the GC content of reporter. The GC content groups are specified from 10% to 90%, in steps of 5%, with the box label specifying the upper limit of the interval. A box label of "0.5", for example, means that the reporters considered for this box had a GC content of* $0.45 < gc_i \leq 0.5$. *The width of each box is proportional to the number of reporters in the respective group.*

each reporter occurs roughly 30 times on the microarray. The thirty-odd copies of the reporter are placed in individual beads at random positions onto the array surface. Each bead contains an additional oligo-nucleotide stretch that needs to be decoded to identify the reporter at each bead position. The BeadArray platform has shown to yield precise results and to be highly comparable with Affymetrix expression microarrays that contain multiple distinct oligonucleotides for measuring a transcript [37]. However, since each transcript is only represented by one or two reporters, these reporter sequences must be carefully chosen from all 50-nucleotide subsequences of the transcript.

Reporter issues Reporter levels are strongly affected by the GC content of the reporter sequence. Figure 1.2 demonstrates this effect on example data. The reporters, which are 24mers in this example, are grouped by their GC content and the distribution of the reporter levels per group is shown. The level of each reporter is shown as median normalised expression level over six heart and skeletal muscle samples (these samples are the expression data from the study described in Chapter 3). There is clearly a significant relationship between the reporter levels and their GC content. Increases in the GC content coincide with higher median reporter levels. The reason is that there are three hydrogen bonds with complementary guanine and cytosine bases, but only two between adenine and thymine. This effect has been reported before, and normalisation methods for microarray data that specifically take the GC content of the reporters into account have been suggested [38].

Microarray intensity preprocessing Even when only considering data generated on the same microarray platform and from the same biological condition, reporters that are supposed to measure the abundance of the same mRNA typically show a large variation across their raw intensities. Sources of variation can be divided into reporter-specific variation, such as reporter sequence characteristics and reporter spotting efficiency; sample-processing variation, such as purification and amplification of biological material, labelling of the material, hybridisation of the material and scanning of the microarrays; and sample-specific sources of variation. The non-biological variation, namely reporter-specific and sample-processing variation, can be referred to as “obscuring” variation [24]. One usually aims to carefully reduce this obscuring variation while preserving the biological variation by *normalisation*. Most normalisation methods work under the assumption that, on any microarray, the observed intensity measurement y_i of reporter i is a function of $[\text{mRNA}_i]$, the concentration of the transcript that reporter i represents, and a reporter-specific or general background intensity bg_i:

$$y_i = f\left([\text{mRNA}_i], bg_i\right) \ . \tag{1.1}$$

This function usually contains additional error terms that capture additional noise in the data. The concentration $[\text{mRNA}_i]$ is the signal of interest, and the aim of most normalisation methods is to down-weight the influence of the background measurement bg_i of the reporter level. Normalisation methods differ with respect to the type of function f that is assumed for the observed intensity, how the background measurement bg_i is estimated, and how bg_i is consequently treated. Many normalisation methods, such as *vsn* [25], assume that most transcripts (or a specified subset of transcripts) are not differentially expressed between the analysed samples and use replicate measurements of these transcripts to estimate the background levels and error terms. Normalisation methods have frequently been shown to improve the signal-to-noise ratio in expression microarray studies.

One approach to obtaining estimates of the reporter-wise background intensities is to co-hybridise a control RNA or genomic DNA in addition to the RNA samples of interest to the microarray platform. This approach is commonly applied with two-channel expression microarrays, with array CGH experiments, and with ChIP-chip experiments. In Chapter 5, I describe how the readouts from genomic DNA hybridisations can be used to normalise transcription data on whole-genome tiling microarrays.

1.4.1 ChIP-chip

ChIP-chip, chromatin immunoprecipitation combined with tiling microarrays, is a well-established high-throughput assay for DNA-bound proteins [39] and post-translational chromatin/histone modifications.

Briefly, the procedure is as follows (see the supplement to [39] for a more extensive description of each step):

1. *in vivo* fix proteins interacting with DNA in place using formaldehyde
2. split the DNA-protein construct into random fragments, usually by *sonication*
3. enrich DNA fragments that are linked to the protein or show the chromatin modification by use of a specific antibody against the protein/modification by *immunoprecipitation* (IP)
4. reverse the DNA-protein binding and wash away the proteins

5. label the IP-enriched DNA fragments with one dye and label control DNA fragments, such as non-IPed sonicated genomic DNA (*input*) with another dye, or possibly the same dye if a set of one-channel microarrays are used
6. hybridise the IP-enriched DNA fragments and the control fragments on a microarray

Steps 1 – 4 make up the chromatin immunoprecipitation (ChIP) steps, and the second "chip" refers to the hybridisation to a microarray.

ChIP-chip experiments have been successfully applied for the detection of binding events of transcription factors, such as Gal4 [39] or the oestrogen receptor [31], for pinpointing the positions of nucleosomes [40], and for identifying genomic regions, in which the histones bear certain post-translational modifications [41, for example]. In Chapter 3, I describe a ChIP-chip study, in which we identified the genomic positions of four histone modifications and the DNA binding events of four transcription factors in heart and muscle cell and related these findings to gene expression and the functional role of the modifications and transcription factors.

1.5 Microarray-based investigation of transcription throughout the cell cycle

1.5.1 Budding yeast cell cycle

Eukaryotic cells reproduce themselves by duplicating their DNA and then dividing into two daughter cells. The daughter cells repeat the same process, and so forth. The cell division (*mitosis*) phase is followed by a gap phase, in which the cells grow and fulfil their physiological function, after which they enter the S phase of DNA duplication. In a second gap phase following the S phase, synthesis mistakes in the replicated DNA are corrected before the cell enters mitosis and divides into two cells [2]. Budding yeast (*S. cerevisiae*) is unusual in that its cells give rise to unequal cells in

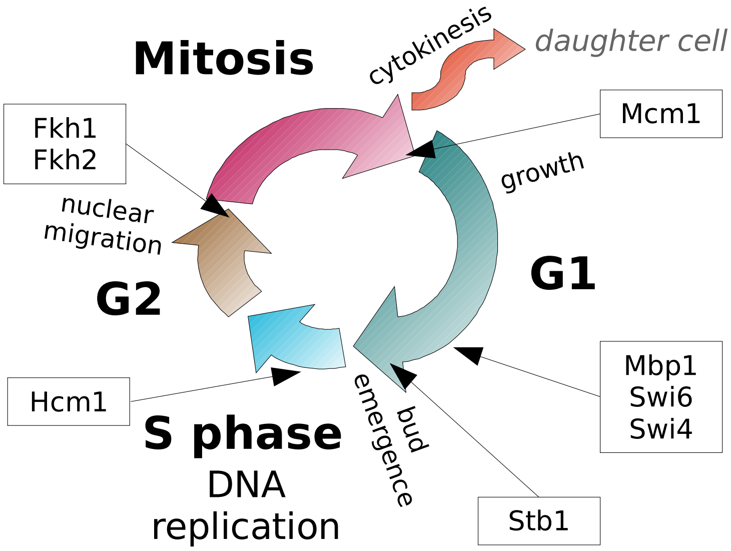


Figure 1.3: *Diagram showing the cell cycle of budding yeast. In addition to the phases, certain key events at stages of the cycle are also indicated, as are a few transcription factors (in the boxes) whose target genes are needed for these key events and/or drive the progression through the cycle. The length of the coloured arrows indicate the approximate proportional length of each of the four phases. One total iteration of the cycle when growing in rich media takes about 60–100 minutes (after release from cell cycle arrest; see Chapter 4).*

division, that is to a larger "mother cell" and a smaller "daughter cell". The smaller daughter cell evolves from a bud extension of the mother cell that emerges at the end of the G1 phase [42].

Figure 1.3 shows a schema of the cell cycle of *S. cerevisiae*. In addition to the phases of the cycle, certain key events are listed. Boxes hold a few selected transcription factors, and the arrows indicate at which stages of the cycle these TFs have been reported to regulate the expression of their target genes. These TFs include Mbp1, Swi6 and Swi4, which are all active in the late G1 phase [43], Stb1 at the G1/S phase transition [44], and the S phase regulator Hcm1 [45]. The listed TFs further include the two transcription factors Fkh1 and Fkh2, which are involved in the regulation of gene expression in the late G2 and early M phase [46], together with Mcm1, which is also a regulator of expression in late M- and early G1 phase [47].

Transcripts that are regulated by these TFs show periodic expression profiles concordant with the cell-cycle progression. Many periodically expressed genes are involved in events that only occur once in the cycle, such as budding, DNA replication and cytokinesis. Some periodically expressed genes, such as cyclins, are also involved in regulating the progression of the cycle itself [48].

1.5.2 Microarrays and the cell cycle

Cho *et al.* [49] used oligonucleotide expression microarrays to investigate the cell-cycle expression of the annotated *S. cerevisiae* genes over two independent time courses of samples taken after release from prior arrest of the cell cycle. Cells were synchronised to the same stage of the cell cycle in distinct ways for the two time courses, in one case using a temperature-sensitive mutant of the cycle-dependent kinase Cdc28, arresting the cells in the late G1 phase, and in the other case a mutant of Cdc15, arresting the cells in late G2 phase. The authors focused on those genes showing at least two-fold changes in expression over the time course and visually inspected them for periodicity in their expression patterns. Cho and co-workers identified 416 genes that were periodically expressed in concordance with the cell cycle progression and assigned them to different stages of the cycle.

Spellman *et al.* [48] used cDNA expression microarrays to analyse the expression of all annotated *S. cerevisiae* genes in three cell cycle time course data sets. The time courses differed in the methods used for synchronising the cells. The first method supplied the α-factor pheromone to the cells, the second one used a temperature-sensitive mutant of Cdc15, and the third method enriched small G1 phase cells through centrifugation. On the microarrays, cDNA from the synchronised samples was hybridised against cDNA from a control sample of cells asynchronously growing in rich media. Spellman and co-workers also included the previously published data of Cho *et al.* (see above) into their analysis. The periodicity of gene expression profiles was assessed by a Fourier transform algorithm and by correlating the expression profiles of all genes to the expression profiles of known cell-cycle genes. The authors identified 800 genes showing periodic expression. They assessed the co-expression patterns of these genes in two separate ways, firstly by the phase of the genes' peak expression (derived from the Fourier transform) and secondly by clustering them using correlation distance between expression profiles [50]. About half of the periodic genes were assigned to well-understood functional groups by the clustering. By the clustering, peak time assignment and by searching for overrepresented sequence motifs in the promoter regions of the clusters, Spellman *et al.* could provide a functional description of about 500 of the 800 periodically expressed genes.

A common challenge with cell cycle time courses is how to identify the periodically expressed transcripts. De Lichtenberg *et al.* compared different methods for this task and concluded that simple approaches for assessing periodicity, such as the ones used in [48] and [49],perform remarkably well in comparison to later suggested more complex modelling approaches [51]. The authors also concluded that the aspects needing to be considered are the expression-profile periodicity and the variation in the expression profile across the time course.

In Chapter 4, I describe a tiling microarray time-course study of the *S. cerevisiae* transcriptome along the cell cycle. For these time courses, we used two distinct established methods for synchronising the cells [48, 49] and identified periodically expressed transcripts. Besides known cell-cycle regulated ORF transcripts, these periodically expressed transcripts included antisense transcripts and intergenic transcripts from unannotated genome regions. By clustering the periodic transcripts and providing a functional description of the clusters, we could segregate them into modules of transcripts with regulatory roles at different stages of the cycle. The antisense and unannotated intergenic transcripts were divided into these clusters as well, indicating that such transcripts may play a role in the cell cycle progression.

Microarrays can only recover changes in transcript abundance along the cell cycle. There are sufficient indications that many genes that are important for cell cycle regulation are certainly regulated at the stage of transcription (Section 4.3, [48, 49]). However, cases of cell-cycle related genes that are not regulated by transcription but only post-translationally are known [49], as are genes that show periodic expression patterns but have constitutive roles regardless of the cell cycle stage [48].

1.6 Gene annotation

There is substantial variation in how much is known about individual genes. For only about 20% of *H. sapiens* and 10% of *M. musculus* proteins, the biological function has been experimentally characterised (as of October 2007) [52].

In this dissertation, I mostly work with two kinds of gene information, the gene coordinates in the genome and the gene's annotation in the Gene Ontology (GO) [53], if present. The GO provides structured annotation of gene functions. It consists of three directed graphs that structure single pieces of gene function, the individual *terms*, or *nodes*, and links them in meaningful ways, such that more specialised terms are child nodes of related general concepts. For example, the term *biological regulation* (GO:0065007), amongst others, is an ancestor of *regulation of programmed cell death* (GO:0043067), which in turn is an ancestor of *negative regulation of apoptosis* (GO:0043066)[2].

In detail, the GO consists of three independent, directed graphs, the ontologies, namely *biological process*, *cellular component* and *molecular function*, which allow for complementary descriptions of gene functions.

Genes, or the proteins encoded by protein-coding genes, are annotated to GO terms based on different sources of evidence, and this source is stored as the *evidence code* with each gene-term relation. The evidence codes are listed in Table 1.1.

When considering the GO annotation for genes, I omitted annotations

[2] GO term identifiers and ordering as of May 2008

Abbreviation	**evidence code**
IC	inferred by curator
IDA	inferred from direct assay
IEA	inferred from electronic annotation
IEP	inferred from expression pattern
IGI	inferred from genetic interaction
IMP	inferred from mutant phenotype
IPI	inferred from physical interaction
ISS	inferred from sequence similarity
NAS	non-traceable author statement
ND	no biological data available
TAS	traceable author statement

Table 1.1: *Gene Ontology evidence codes*

with evidence codes IEA, NAS and ND, since these are less reliable than the other types of annotation. The majority of annotated gene-term relations for *S. cerevisiae* are *inferred from electronic annotation* (IEA) (31,385 out of 50,452, as of January 2008).

The GO graphs provide a structured vocabulary of gene annotations. This structure has to be taken into account when performing statistical tests on individual GO terms, as the graph structure introduces statistical dependencies between terms within the same ontology.

Chapter 2

Coexpression of adjacent genes

2.1 Introduction

Comparative studies of genomes and transcriptomes have shown that genes that are concordantly expressed over time in the same tissue or co-expressed in many tissues are often located next to each other in certain chromosomal regions.

2.1.1 Gene clusters

Clusters of coexpressed genes were first identified in *Saccharomyces cerevisiae* [49, 54] and *Caenorhabditis elegans* [55, 56]. In prokaryotes and in *C. elegans*, such clusters are co-transcribed in the form of *operons*, an uncommon concept in eukaryotes other than *C. elegans*, although recent tiling microarray data sets have indicated that other eukaryotic transcriptomes may also contain small numbers of operons [9] (see Chapter 5, Figure 5.2). In *Drosophila melanogaster*, clusters of 10-30 coexpressed genes that span, on average, 125 kb of genomic DNA can be observed, if one allows a limited number of intervening genes with deviant expression patterns [57]. In *Homo sapiens*, genes with high expression levels tend to conglomerate in chromosome domains [58, 59]. Furthermore, certain gene clusters, such as the beta-globin or HoxD genes, are regulated by a single control locus [60], or by a global control region [61]. Transcription of functionally related genes from bidirectional promoters has also been described; for example, the human genes *PPAT* and *AIRC*, both of which are involved in purine biosynthesis, are divergently transcribed (see Figure 1.1) from a 700 bp promoter region on chromosome 4 [62].

It has been implied that coexpressed, clustered genes are mainly housekeeping genes, i.e. genes that are involved in fundamental cell functions and thus expressed in all tissues [63, 64, 65]. Some reports indicate that clusters of coexpressed genes tend to be conserved through evolution. For example, when looking at traces of inter-chromosomal rearrangements that occurred during the diverging evolution of human and mouse, clusters of coexpressed genes are more conserved than other sets of adjacent, orthologous genes [66]. Such clusters may have proven evolutionarily beneficial, and the cluster arrangement may be preserved by natural selection.

Fundamental questions about coexpression clusters remain unanswered. How frequent are such clusters in eukaryotes? By what mechanism is the transcriptional coupling of clustered genes brought about? And what is the evolutionary origin of cluster formation?

2.1.2 Measures of gene coexpression among clusters

The most commonly used measure for coexpression of a pair of genes is the Pearson correlation coefficient (CC) of the expression profiles of the two genes. For example, Cohen *et al.* [54] considered a pair of genes to be coexpressed if the two gene expression profiles had a Pearson CC greater than 0.7. Clusters of n genes were considered coexpressed if the CC of each of the $n-1$ pairs of adjacent genes was greater than 0.7. One known shortcoming of the Pearson CC, however, is its susceptibility to outliers, an issue that cannot be neglected with data that are as noisy as gene expression data [67].

Singer *et al.* [66] suggested three complementary measures of expression similarity for a pair of genes in their analysis of the Novartis gene expression atlas data set from 2002 [68]. First, they quantified the "housekeepingness" of the pair as the product of the proportions of tissues in which each of the two genes was considered to be expressed. Second, the pair's overall expression was calculated as the average of the pair's expression levels across all tissues of the data set. Third, the coexpression of the two genes across tissues was measured by the Pearson correlation coefficient of the two genes' expression profiles. For each of the three similarity measures, clusters of coexpressed genes were then identified using a sliding-window algorithm with a fixed window width of 10 genes and a step size of one gene, but only windows in which the 10 genes spanned less than 500 kb were analysed. In each window, the 9 similarity scores of each pair of consecutive genes were summed up to yield a coexpression score for the window. The window scores were compared to scores of 100,000 random sets of 10 genes, and scores exceeding the 95% or 99% quantile of the scores of the random sets were taken to indicate that the window contains a cluster of coexpressed genes. One drawback of the method of Singer *et al.* is that the three measures for assessing coexpression of a gene pair are considered separately from each other. Another drawback of the method is the fixed window size of 10 genes for finding gene clusters. Due to the latter limitation, this method may miss shorter clusters of less than 10 coexpressed genes.

I considered two measures for assessing coexpression of pairs of genes. I first used the Hamming distance as a way to assess coexpression, since gene expression profiles are binary vectors in our data. Later on, I devised a more informative two-dimensional measure of gene coexpression.

2.2 Material and Methods

Data sets For this analysis, two publicly available data sets were considered, namely the *Fantom3* transcription data (*Mus musculus*) [69] and the 2004 Novartis *Symatlas* expression data (*H. sapiens*) [65]. The Fantom3 data set contains expression profiles of 39,593 transcriptional units (TUs) mapped to the *mm5* assembly of the mouse genome (obtained from `http://fantom31p.gsc.riken.jp/cage/mm5/`). A TU is a set of (m)RNAs that share at least one transcribed genomic nucleotide and are transcribed from the same segment of the genome in the same orientation [70]. In the following, I use the term *gene* synonymously for transcriptional unit.

Expression information was obtained by four methods: full-length sequencing of isolated cDNAs, cap analysis of gene expression (CAGE), gene identification signature (GIS) and gene signature cloning (GSC). CAGE, GIS and GSC aim at identifying transcripts by recognition of short characteristic sequence tags. For our analysis, we focused on the following tissue transcriptomes (tissues are named as in the Fantom3 publications; numbers in brackets denote the number of genes expressed in that tissue): adipose (19,166), brain (13,766), cerebellum (18,753), diencephalon (6,567), heart (8,423), liver (30,721), lung (30,560), macrophage (26,746), muscle (8,829), prostate gland (10,795), somatosensory cortex (17,193), testis (13,347), visual cortex (17,216).

The Symatlas data [65] were generated on a combination of two oligonucleotide microarray designs (HG-U133A and GNF1H) from Affymetrix, and contained measurements for approximately 34,000 probe sets. The MAS5 algorithm [71] was used to preprocess the data and to assign a call to each probe set's expression in each tissue, one of *present*, *marginal*, or *absent*. Probe sets were associated to genes (and thereby chromosomal locations) according to the probe set annotation supplied by the microarray manufacturer. A gene was considered to be *expressed* if any of its associated probe sets had a *present* or *marginal* call. Probe sets with incomplete location information, which did not unambiguously relate to a gene, were excluded. The resulting data set consisted of binary expression calls for 19,358 genes with distinct chromosomal locations.

Adjacency of transcriptional units For the purposes of this study, two TUs are considered to be genomic neighbours, or adjacent to each other, only if, according to their genome location annotation, they are located on the same chromosome and there is no other TU annotated to the genomic region in between them on either strand. The actual number of base pairs between the TUs is irrelevant for this designation, as is whether the TUs are on the same DNA strand.

Transcription factor binding sites (TFBSs) We considered only TFBSs that were considered to be conserved in human/mouse/rat alignments, and annotated as such in the UCSC Genome Browser (`http://genome.ucsc.edu`). All TFBSs annotated in the 10 kb upstream region of each TU were obtained, based on the Ensembl gene identifier of the TU. The TFBS annotation of the UCSC Genome Browser is conservative and likely to contain few false positives but may miss many true positive TFBSs. Two genes were considered to have common *cis*-acting regulatory units if both had binding sites of the same transcription factors annotated in their 10 kb upstream region.

Gene Ontology I investigated the similarities between the Gene Ontology (GO) [53] annotations for pairs of genes. For each gene in the Fantom3 data, the GO annotations were obtained by using the Bioconductor package *biomaRt* [72] to query the Ensembl database [34](release 33, May 2005). Only the most specific GO terms annotated for each gene were kept and their ancestor terms[1] were disregarded.

The similarity between GO annotations for a pair of genes was determined as follows. Let $T(g_i)$ be the set of most specific GO terms annotated for gene g_i and $|T(g_i)|$ denote the cardinality of this set. Two genes g_i and g_j were considered to have a similar GO annotation if

$$|T(g_i) \cap T(g_j)| \geq 0.5 \cdot \min\left(|T(g_i)|, |T(g_j)|\right) \quad , \tag{2.1}$$

i.e. if 50% of the most specific GO terms of the gene with fewest such annotated terms were also annotated for the other gene.

Protein domain information I also investigated the similarities between the protein domain annotations for pairs of genes. For each gene in the

[1] See Section 1.6 in Chapter 1 for a description of the GO structure.

Fantom3 data, the domain information for the encoded protein was obtained, where available. I used the Bioconductor package *biomaRt* [72] to query the Ensembl database (release 33, May 2005). Two genes were considered to have a similar domain annotation if at least 50% of the domains of the gene with fewest domains annotated were also present in the protein of the other gene.

2.3 Results

2.3.1 Permutation scheme for estimating the null distribution of coexpression scores

I permuted the order of the TUs in the genome, while keeping the actual positions at which TUs are located and the expression profile of each TU across tissues fixed. More formally, given N ordered identifiers of TUs $\Psi = (\psi_1, \ldots, \psi_N)$ located at N ordered genomic locations $\Lambda = (\lambda_1, \ldots, \lambda_N)$, the original mapping consists of the tuples (ψ_i, λ_i), $i \in [1, N]$. In each permutation, a new set of locations $\overline{\Lambda}$ is constructed from Λ by randomly reordering the elements of Λ, and the permuted mapping consists of tuples $(\psi_i, \overline{\lambda}_i)$, $i \in [1,\ N]$. The mapping between the N identifiers and the N expression profiles, however, is the same in the original data and in each permutation. The underlying null hypothesis of this permutation scheme is that coexpression of genes within tissues and across tissues is independent of the genes' genomic location. By only permuting their location, but keeping the genes' identifiers, the number of genes expressed in that tissue as well as coexpression patterns between genes are preserved.

2.3.2 Chromosomal clustering of transcriptomes

We investigated the genomic organisation of 13 *M. musculus* tissue transcriptomes that had been analysed in the Fantom3 project [69]. First, the scale of clustering of tissue-expressed genes along the genome was assessed. A set of two or more adjacent genes that were expressed in a particular tissue was called a *tissue coexpression cluster*. 30-75% of the genes expressed in each tissue were arranged in such clusters along the genome, without any obvious prevalence for particular chromosomes. The clusters consisted mainly of two or three genes.

To evaluate the significance of this observation, we compared the observed number of genes expressed in clusters with the numbers seen in 10,000 permuted versions of the data (see above for the permutation scheme). This permutation approach shows that while a large number of genes can already be expected to be part of tissue coexpression clusters under the null hypothesis, the observed number of clustered genes is significantly larger. For 10 out of the 13 tissues, none of the 10^4 permutations showed a greater or equal number of genes in such clusters. For instance, in the tissue *brain*, there are $8,214$ genes in such clusters in the actual Fantom3 data, while the 10^4 permuted versions of the data contain between $7,714$ and $8,107$ genes in such clusters. This observation corresponds to an empirical p-value $p < 10^{-4}$ for each of these 10 tissues; the p-values for the other three tissues were $4 \cdot 10^{-4}$ (diencephalon), 0.203 (lung) and 0.2954 (liver).

2.3.3 Evaluating coexpression across tissues

One-dimensional measure of coexpression

After preprocessing both data sets, Fantom3 and Symatlas, matrices X are obtained with the rows holding the TUs and the columns holding the tissues. These matrices are binary, with $X_{ij} = 1$ if TU i is deemed to be expressed in tissue j, and $X_{ij} = 0$ otherwise. The expression profiles of two adjacent TUs i and $i+1$ are both binary vectors of equal length. An obvious way to compare such vectors is the *Hamming distance* (HD) [73], the number of positions in which the two binary vectors differ, i.e. the number of tissues in which the two expression profiles differ.

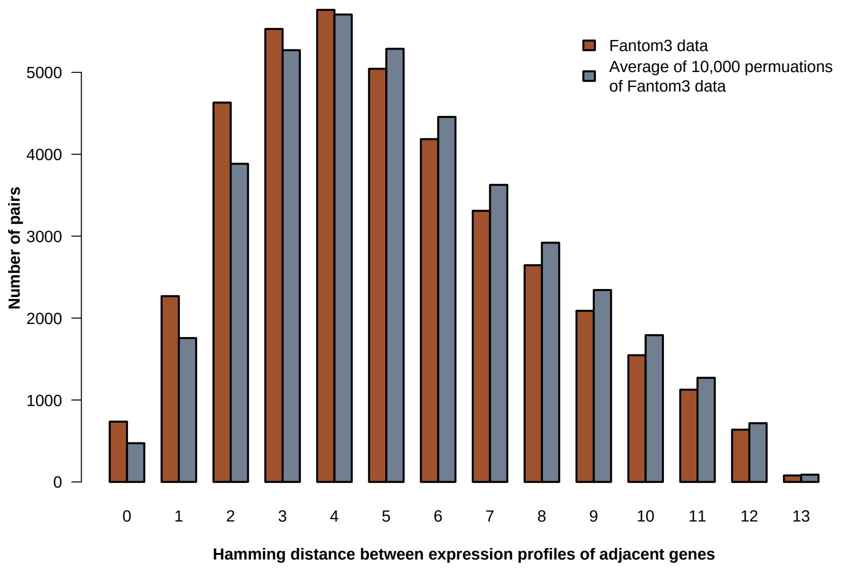


Figure 2.1: *Fantom3: bar plot comparing observed Hamming distances between the expression profiles of adjacent genes across tissues with the average Hamming distances over 10,000 permutations of the gene order.*

A probabilistic description of the HD The coexpression of a given pair of genes across N tissues can also be described in terms of a 2×2 contingency table.

		Gene 1 expressed	
		0	1
Gene 2 expressed	0	v	x
	1	y	z

with $v + x + y + z = N$.

Once this contingency table is filled in for a gene pair, the Hamming distance d_H for this pair is given by

$$d_H = N - v - z = x + y \, . \tag{2.2}$$

I computed the Hamming distance for all pairs of adjacent genes in the actual data and compared the observed distances to the ones obtained in $10,000$ permutations of the gene order. See Figure 2.1 for the bar plot. The observed Hamming distances tend to be smaller with the actual Fantom3 data than the average distances over $10,000$ permutations of the gene order. A χ^2 test also suggests rejection of the null hypothesis that the bin frequencies of observed Hamming distances are consistent between the actual data and the data with permuted gene order ($p < 2.2 \cdot 10^{-16}$, the test statistic approximately follows a χ^2_{13} distribution under the null hypothesis).

For pairs of highly coexpressed genes, the Hamming distance between their binary expression profiles is small. One shortcoming of the Hamming distance for scoring coexpression, however, is that the reverse is not necessarily true, since the HD completely disregards the actual expression of each TU across tissues. Consider these two example pairs of adjacent TUs:

Pair I

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
TU Ia	0	1	1	1	1	1	1	1	1	1	1	1	1
TU Ib	0	1	1	1	1	1	1	1	1	1	1	1	0

Pair II

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
TU IIa	0	0	0	0	0	0	1	0	0	0	0	0	0
TU IIb	0	0	0	0	0	0	0	0	0	0	0	0	0

where T1...T13 indicate the 13 tissues of the Fantom3 data set. In both pairs, the TU a is expressed in one more tissue than its neighbour TU b. Thus, both pairs have a Hamming distance of 1. With pair I, however, one of the TUs is expressed in 12 tissues and its partner is also expressed in 11 out of these 12 tissues. These two adjacent TUs can clearly be considered to be *highly coexpressed*. In contrast, with pair II, the first TU is expressed in a single tissue, while its neighbouring TU is not expressed in any tissue. Even though, both pairs have an equally small Hamming distance of 1, the TUs in the pairs show very different degrees of coexpression.

Scaled Hamming distance To get a more informative measure of the coexpression of two genes, the observed Hamming distance is scaled by a

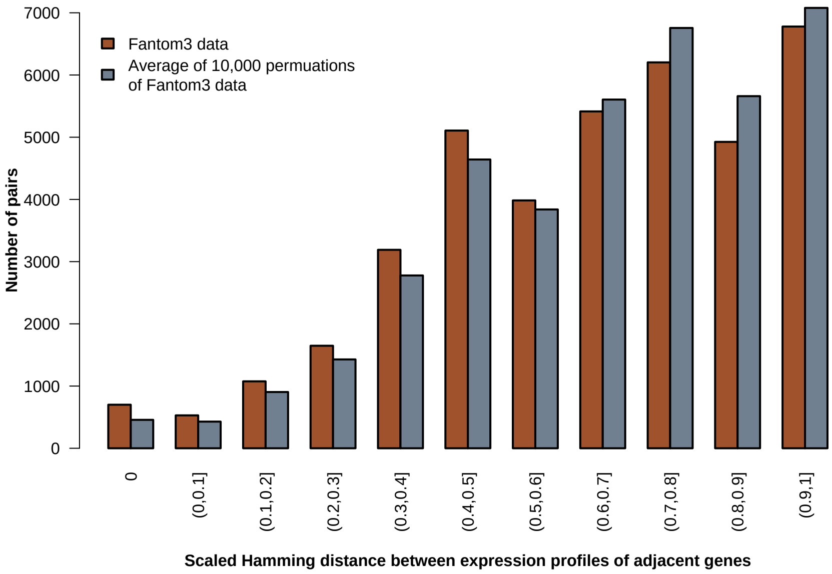


Figure 2.2: *Bar plot comparing observed scaled Hamming distances between the expression profiles of neighbouring genes and average scaled Hamming distances over 10,000 permutations of the gene order in the Fantom3 data. The scaled distances are binned into intervals, which are specified in the open interval notation. For example, (0.1, 0.2] summarises all scaled HDs d_{SH}, for which* $0.1 < d_{SH} \leq 0.2$.

factor that indicates the overall expression of the gene pair. The observed HD is multiplied by the scaling factor $1/\tau$, where τ is the number of tissues in which either one or both genes are expressed.

Considering the previously described 2×2 contingency table for coexpression of a pair of genes across tissues (page 27), the scaled Hamming distance d_{SH} for the a pair is computed as

$$d_{SH} = \frac{N - v - z}{N - v} = \frac{x + y}{x + y + z}. \tag{2.3}$$

The scaled Hamming distance d_{SH} is a number between 0 and 1 with lower numbers indicating higher degrees of coexpression. For pair I in the previous example, the distance is $d_{SH} = 1/12$, and for pair II it is $d_{SH} = 1/1 = 1$. The scaled HD thus clarifies that pair I is highly coexpressed and that pair II is not coexpressed at all.

I computed the scaled Hamming distances for all pairs of adjacent genes in actual data and compared them to the distances seen in the 10,000 permuted versions of the data (see Figure 2.2). The scaled HDs observed on the Fantom3 data are smaller than the distances in the permuted versions of the data. A χ^2 test also suggests rejection of the null hypothesis that the bin frequencies of observed scaled Hamming distances are consistent between the actual data and the data with permuted gene order ($p < 2.2 \cdot 10^{-16}$, the test statistic approximately follows a χ^2_{10} distribution under the null hypothesis).

Nevertheless, the scaled HD does also not provide a satisfying resolution in quantifying the coexpression of gene pairs across tissues. Consider the following two example pairs of TUs:

`Pair III`

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
TU IIIa	0	1	1	1	1	1	1	1	1	1	0	0	0
TU IIIb	0	0	0	0	1	1	1	1	1	1	1	1	1

`Pair IV`

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
TU IVa	0	0	0	0	0	1	1	0	0	0	0	0	0
TU IVb	0	0	0	0	0	0	1	0	0	0	0	0	0

where `T1`...`T13` indicate the 13 tissues of the Fantom3 data set. Both pairs `III` and `IV` have a scaled HD of 0.5, but while the TUs in pair `III` are expressed in many tissues, and coexpressed in 6 tissues, pair `IV` consists of two rarely expressed TUs, which are coexpressed in one single tissue. Moreover, if one of the positive measurements of pair `IV` turns out to be a False Positive the qualitative statement of the pairs' coexpression changes completely, while for pair `III` the coexpression statement is more solid.

I conclude that the Hamming distance, or a scaled version thereof, is not an appropriate measure of gene coexpression. In fact, no one-dimensional measure was found able to appropriately evaluate coexpression of adjacent TUs and also to distinguish between coexpressed tissue-specific TUs, housekeeping gene pairs and silent gene pairs.

A two-dimensional measure of coexpression

To quantify coexpression of a pair of genes in a set of n tissues, two coefficients are defined. Ω is the proportion of tissues in which either one or both genes are expressed, and A is the proportion of tissues in which both genes are expressed. Both coefficients are numbers between 0 and 1, and $A \leq \Omega$. If $A = \Omega$ the two genes have an identical expression pattern across tissues, while a small ratio ${}^{A}/_{\Omega}$ indicates that the two genes are rarely coexpressed.

Considering the previously described 2×2 contingency table for coexpression of a pair of genes across tissues (page 27), these coefficients correspond to

$$\Omega = \frac{x+y+z}{N} \tag{2.4}$$

and

$$A = \frac{z}{N}. \tag{2.5}$$

Two thresholds Θ_{coex} and Θ_{unc} to the ratio ${}^{A}/_{\Omega}$ were introduced for assigning each pair of neighbouring TUs to one of the following coexpression categories:

1. *house-keeping*, if $A = 1$
2. *highly coexpressed*, if ${}^{A}/_{\Omega} \geq \Theta_{coex}$ and $A < 1$
3. *uncorrelated*, if ${}^{A}/_{\Omega} \leq \Theta_{unc}$
4. *silenced*, if $\Omega = 0$.

I computed these coefficients for each pair of adjacent genes in the Fantom3 data and compared them to the values expected under the null model.

Figure 2.3 shows the empirical p-values for each tuple (A_i, Ω_j). The frequency of each tuple in the Fantom3 data was compared with the tuple's frequencies in $10,000$ permuted versions of the data. For each tuple (A_i, Ω_j), the empirical p-value is given by the proportion of permutations in which equally many or more gene pairs display this coexpression pat-

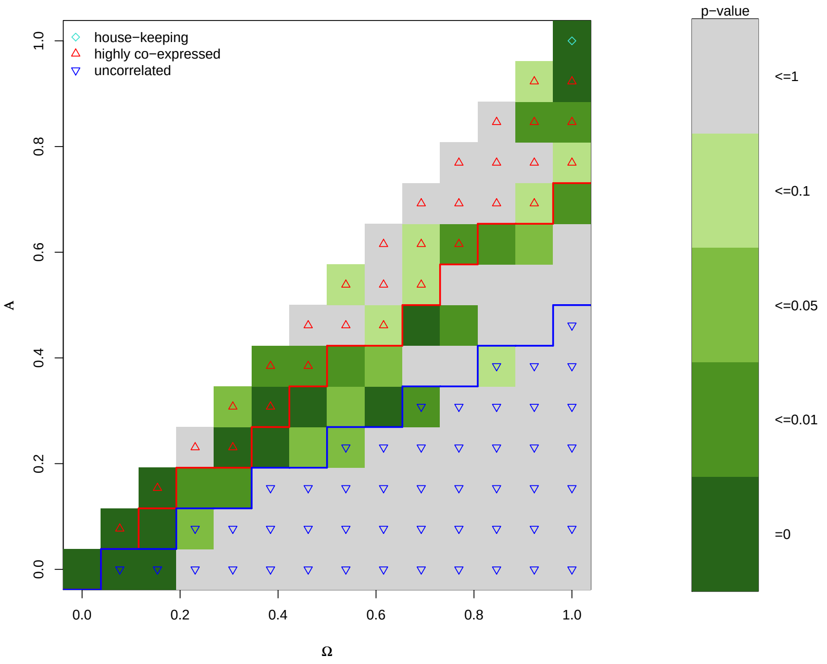


Figure 2.3: *Empirical p-values for numbers of gene pairs in Fantom3 transcription data binned by general expression (A) and coexpression (Ω) over 13 tissues. For each tuple/bin, the p-value indicates the proportion among 10,000 permuted versions of the Fantom3 data, in which the frequency of this tuple was equal or higher than in the actual Fantom3 data. Red and blue lines indicate the thresholds on the ratio $^{A}/_{\Omega}$ for highly coexpressed and uncorrelated gene pairs, respectively.*

tern (A_i, Ω_j) as in the actual data. The distribution of the bivariate coexpression measure (A, Ω) is non-random; certain tuples (A_i, Ω_j) occur more frequently in the genome than expected if coexpression were independent of genomic location.

For the Fantom3 data, we set the thresholds $\Theta_{coex} = 0.75$ and $\Theta_{unc} = 0.5$. These thresholds resulted in 3,230 highly coexpressed pairs (HCPs), 154 housekeeping pairs, 36 silenced pairs and 27,287 uncorrelated pairs (UCPs). Figure 2.3 shows that the number of HCPs is larger than expected under the null model. Similarly, there are more housekeeping pairs, and more silenced pairs, than expected.

2.3.4 Highly coexpressed clusters

As a generalised version of the highly coexpressed pair, a highly coexpressed cluster (HCC) is defined as a set of two or more neighbouring genes, in which each consecutive pair of genes is considered to be a highly coexpressed pair. It follows that a HCC of length 2 is the same as a HCP. The majority of HCCs consist of two or three clustered genes (see Table 2.1).

Genes in HCC	2	3	4	5	6	7
# such HCCs	2,155	359	74	27	3	2

Table 2.1: *Numbers and lengths of observed highly coexpressed clusters in the Fantom3 data. #: 'number of'.*

2.3.5 HCPs and housekeeping functionality

It has been reported that housekeeping genes are often arranged in clusters along the genome [64], and the same can be seen in the Fantom3 data. For 298 out of 1, 915 TUs that are expressed in all 13 tissues, the adjacent TU is coexpressed in all tissues as well, which is a significantly higher number than expected by chance (see Figure 2.3). The reverse, however, is not true: most of our highly coexpressed pairs are expressed in nine or fewer of the thirteen analysed tissues, indicating that highly coexpressed genes are not necessarily housekeeping genes (see Figure 2.4).

2.3.6 Location, orientation, and dimension of HCCs

Highly coexpressed pairs were homogeneously distributed over all chromosomes and chromosomal regions. Regarding the transcript pair orientation (see Figure 1.1), the observed frequencies of divergent, convergent and tandem transcript orientation were similar in HCPs and genomic neighbour pairs in general. The intergenic distances and distances between the transcription start sites (TSSs) were computed for HCP and compared with the respective distances for all pairs of genomic neighbours. HCPs show smaller intergenic distances (median of 7,662 bp versus

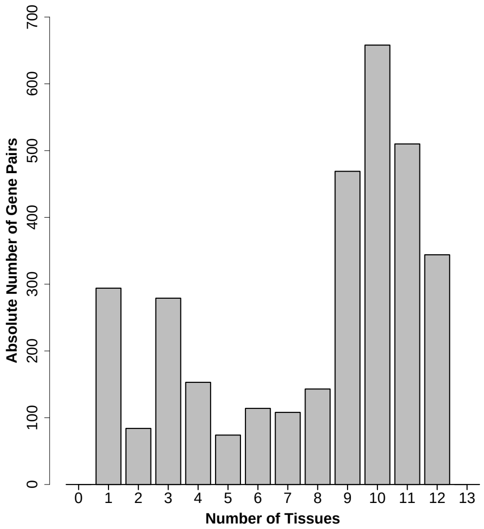


Figure 2.4: *Histogram relating highly coexpressed gene pairs to the number of investigated tissues (0 to 13) in which the genes of the pair are expressed.*

18,665 bp for all pairs, $p = 3 \cdot 10^{-5}$, Wilcoxon Rank Sum Test) and smaller distances between their TSSs (median of 28,781 bp versus 34,491 bp, $p = 8 \cdot 10^{-8}$, Wilcoxon Rank Sum Test).

Highly coexpressed clusters showed an upper bound in size. The extension of a cluster in base pairs is proportional to the number of coexpressed genes in the cluster. For HCCs, we observed a maximal number of 7 genes in a cluster, and the 95% quantile of the cluster extension was 320 kb, as compared to 810 kb for clusters of uncorrelated genes (see Figure 2.5).

2.3.7 Functionality, paralogy and transcriptional regulation of highly coexpressed gene clusters

The similarity in GO annotation, protein domain annotation and sharing of TFBSs among pairs of adjacent genes were investigated. Pairs of highly coexpressed genes were compared with all pairs of adjacent genes in the Fantom3 data. Each analysis was limited to the genes annotated with Gene Ontology terms (36% Fantom3 genes), protein domain information (42%)

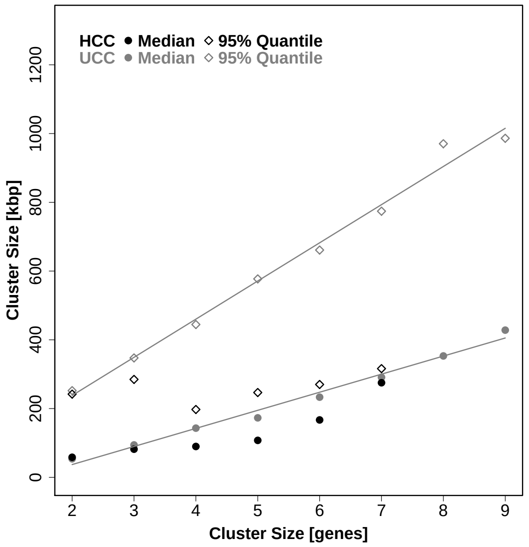


Figure 2.5: *Scatter plot showing median and 95% quantile of the empirical distribution of the genomic extension (in base pairs) of a cluster versus the number of genes in the cluster. These values are shown for highly coexpressed clusters (*HCC*) and for clusters of uncorrelated, consecutive genes (*UCC*). The two lines indicate the least-squares linear-regression fits for median and 95% quantile of the extension of UCC.*

or TFBSs (33%), respectively. Table 2.2 shows that the sharing of protein domains and GO terms is slightly less frequent in HCPs than for genomic neighbours in general, whereas sharing of common TFBSs occurs at a similar rate.

To further investigate the relationship between coexpression and paralogy, gene pairs that had highly similar protein domains but showed only weak coexpression (in total 1,307 gene pairs) were examined. Among such pairs were members of well-known gene families that have previously been described to be clustered at certain genomic locations but to display tissue-specific expression nonetheless. These included the family of S100-calcium binding proteins [74].

	# annotated pairs	% of neighbour pairs with similar annotation	# annotated HCPs	% of HCPs with similar annotations
GO terms	5,586	17.1	1,272	8.8
Protein domains	7,335	18.1	1,567	10.8
TFBSs	4,800	27.4	770	29.7

Table 2.2: *Functional and transcriptional properties of genomic neighbours in* M. musculus*. Genomic neighbours irrespective of their coexpression share Gene Ontology (*GO*) terms and protein domains to a slightly higher extent than do highly coexpressed gene pairs (*HCPs*), whereas a similar number of both groups of neighbours are potentially regulated by the same transcription factors through their respective binding sites (*TFBSs*). The phrase 'pairs with similar annotation' means 'pairs in which the annotation of the partners is considered to be similar'. #: 'number of'.*

2.3.8 *H. sapiens* microarray data

To verify that the observations from the Fantom3 data are not limited to one single data set and organism, we repeated the analysis for the 79 tissues in the *Homo sapiens* part of the GNF Symatlas data set [65].

These data include a greater variety of tissue transcriptomes than the Fantom3 data. In particular, the Symatlas data includes a number of embryonic transcriptomes, which can be expected to show a small overlap with adult tissue transcriptomes. To account for this increased variation of transcriptomes, the thresholds for the definition of HCPs and UCPs were set lower, namely $\Theta_{coex} = 0.50$ and $\Theta_{unc} = 0.33$. These lower thresholds also ameliorated the lower coverage and higher false negative rate that is to be expected in the Symatlas microarray data as compared to the combination of sequencing and tag-based techniques of the Fantom3 data. One indication of the lower coverage of the Symatlas data is that it contained measurements for 19,358 TUs, while the Fantom3 data contains expression information for 39,593 TUs. The transcriptome of *H. sapiens*, however, is not expected to be half the size of the *M. musculus* transcriptome.

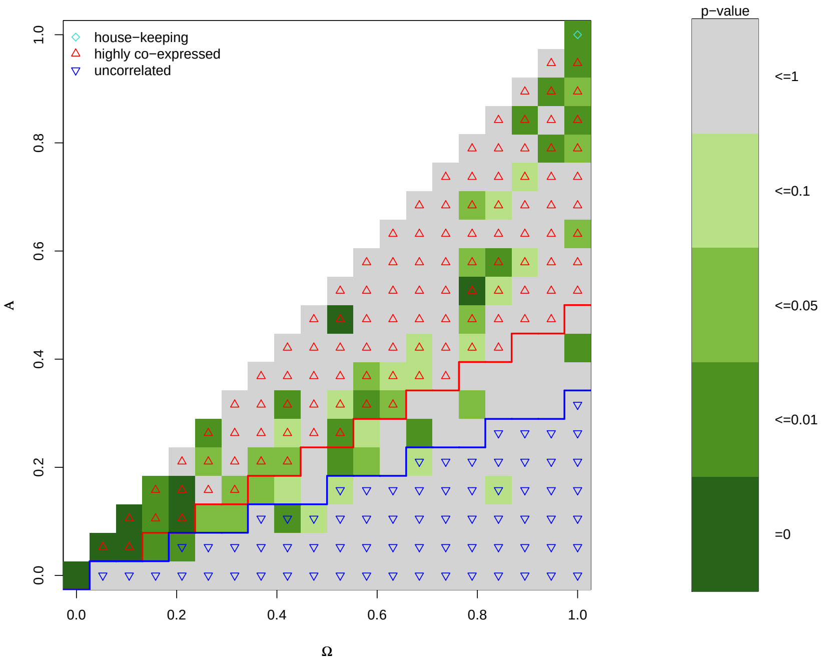


Figure 2.6: *Empirical p-values for numbers of gene pairs in GNF H. sapiens expression data binned by general expression (A) and coexpression (Ω) over 79 tissues divided in 20 equal-sized bins. Red and blue lines indicate the thresholds on the ratio* $^{A}/_{\Omega}$ *for highly coexpressed and uncorrelated gene pairs, respectively.*

We repeated the complete analysis (Sections 2.3.2 – 2.3.7) using the Symatlas data set and obtained very similar results. Figure 2.6 shows the empirical *p*-values per expression/coexpression tuple. Very similar results were also obtained with respect to the chromosomal clustering of genes expressed in human tissues, the tissue distribution of observed HCCs as well as for the relationship of functional similarity, paralogy and transcriptional regulation. In agreement with the results on the Fantom3 data (see Table 2.2), Table 2.3 shows for the *H. sapiens* Symatlas data that sharing of protein domains and GO terms is slightly less frequent in HCP than in genomic neighbours in general, whereas sharing of common TFBS occurs at a similar rate.

We conclude that our observations on chromosomal clusters of coex-

	# annotated pairs	% of neighbour pairs with similar annotation	# annotated HCPs	% of HCPs with similar annotations
GO terms	5,336	10.3	1,047	6.9
Protein domains	5,511	13.3	1,066	9.9
TFBSs	3,174	20.4	600	20.2

Table 2.3: *Functional and transcriptional properties of genomic neighbours in H. sapiens. Genomic neighbours irrespective of their coexpression share GO terms and protein domains to a slightly higher extent than highly coexpressed gene pairs (*HCPs*, coexpression as measured in the GNF H. sapiens Symatlas data), whereas a similar number of both groups of neighbours are potentially regulated by common transcription factors through their respective binding sites (*TFBSs*).*

pressed genes are not specific properties of the Fantom3 data set, or the *M. musculus* transcriptome, and hypothesise that these observations may more generally hold true for mammals.

2.4 Discussion

There is plenty of evidence that eukaryotic genes are ordered along the genome in an organised manner. We focused on large-scale, qualitative features of transcriptional regulation and considered a set of adjacent, coexpressed genes to be a coexpressed gene cluster, irrespective of the quantitative expression levels of the genes. We analysed the *M. musculus* transcriptome using the Fantom3 data set [69] and reproduced the results using *H. sapiens* microarray data (Novartis Symatlas [65]).

In tissue transcriptomes, large numbers of genes were found to be expressed in chromosomal clusters. Observations based on random permutations of the gene order, however, suggested that only a subset of these clusters might be actively transcriptionally coupled. The observed clustering also is an effect of crowding a given number of genes into a genomic region of fixed length.

I assessed the coexpression of gene clusters across tissues and observed a significant proportion of highly coexpressed gene clusters (HCCs) and a small number of housekeeping gene clusters. HCCs are characterised by an upper limit on physical cluster size and on the number of genes making up these clusters (see Figure 2.5), possibly due to the underlying mechanism of coexpression. This finding may point to uncharacterised *cis*-acting units regulating the coexpression of certain sets of genes. The coupling of highly coexpressed clusters could be controlled by histone modifications interacting with specific proteins that loosen or tighten the chromatin structure. Histone-modifying enzymes can affect large chromosomal regions, which are delimited by boundary elements [75, 64] and possibly contain the coexpressed gene clusters. In contrast, uncorrelated clusters may have emerged as a consequence of intervening genes being transcriptionally silenced, for example during cell differentiation. In haematopoietic cells, it has been shown that the stem cells are characterised by a mostly loose chromatin structure and each differentiation step of the cells is accompanied by silencing of genes and tightening of the chromatin structure in specific chromosomal regions [76]. Modifications that affect chromatin structure can be stably inherited across cell division by DNA methylation [77], slowly reversed by histone lysine methylation or rapidly modulated by histone acetylation.

For *S. cerevisiae*, it has been reported that genes regulated by the same sequence-specific transcription factor tend to be periodically spaced across the genome [78]. Other reports suggest that transcriptional regulation determines the organisation of genes in transcriptional units on the chromosome [79]. For example, target genes apparently regulated by the transcription factor *aire* were shown to occur as clusters along the genome. Nevertheless, *aire* sometimes had opposing effects on adjacent genes, upregulating one and down-regulating the other [80]. Our finding that 20–30% of all genomic neighbours show the same transcription factor binding sites (TFBSs) in their upstream regions is in agreement with previous reports. However, sharing of TFBSs does not necessarily result in the coexpression of adjacent genes across tissues, as such sharing is equally frequent in highly coexpressed neighbours and genomic neighbours in general. Our focus on TFBSs that are conserved between *H. sapiens*, *M. musculus* and *Rattus norvegicus* seemed reasonable at that time. Recently, it has been reported that, while the binding motif is often conserved between species, the exact position of the motif with respect to the TSS is not [81]. In the light of this finding, we would probably not limit ourselves to strongly conserved binding sites again.

Considering a broader definition of a gene clusters, irrespective of the genes' transcription rate, genes coding for proteins involved in the same metabolic pathways have been reported to be arranged in clusters in all eukaryotic genomes, albeit to a different degree [82]. Genes involved in stable protein-protein complexes tend to be located on the same chromosome and to be closer to each other than expected by chance [83]. We assessed paralogy and functional similarity as potential explanations for the arrangement of genes in clusters. Previous reports have indicated co-functionality of coexpressed gene clusters [54, 57, 84], without comparison to co-functionality of genomic neighbours in general. Adjacent genes that were highly coexpressed in the Fantom3 or Symatlas data were not found to show a higher degree of co-functionality than genomic neighbours in general. This unexpected finding might be due to the incompleteness of the used annotation information (e.g., the Gene Ontology). Another possible explanation of this observation is provided by models of gene duplication where duplication leads to neofunctionalisation and subfunctionalisation. Neofunctionalisation, in which duplicate genes have diverged in function from the ancestral gene, can result in expression of the duplicate genes in tissues lacking expression of the ancestral gene [85]. Subfunctionalisation, on the other hand, can result in division of the ancestral expression pattern onto the duplicates [86, 85].

The assignment of pairs of adjacent gene pairs into the categories "highly coexpressed" and "uncorrelated" was based on arbitrary thresholds for the two-dimensional measure of coexpression. As the biological mechanisms for the co-transcription of adjacent genes are currently unknown and the data might contain false-positive and false-negative expression calls, a rigorous evaluation of these thresholds is difficult. Some pairs of adjacent transcripts might have been miscategorised. However, the permutation approach and the observations about genomic extent and size of HCCs indicate that the chosen thresholds can be used for a meaningful categorisation of the gene pairs.

We hypothesise that HCCs trace back to large-scale, persistent reorganisations of the transcriptome, whilst TF regulation is likely to fine-tune co-transcription on shorter time scales. At present, we can only speculate on how transcriptional coupling of adjacent genes is brought about. Studies addressing the chromatin remodelling process and the factors involved therein may provide insights into the underlying mechanism.

Contributions to this project

This study was conducted between May 2005 and March 2006. I designed and performed the described bioinformatic analysis, except for the analysis on TFBS sharing between genomically adjacent genes. I wrote the Methods part and most of the Results part of the publication [87] and contributed to the interpretation of the results. Antje Purmann created the binary expression matrices and annotation tables for both data sets. Markus Schüler performed the analysis of TFBSs shared by adjacent genes. Silke Sperling conceived the study and wrote most of the biological introduction and discussion of results in the publication, excerpts of which I have adopted in this chapter to put my methods and analyses into perspective. Wolfgang Huber supervised my analyses. All authors contributed to the publication [87].

Chapter 3

Transcriptional regulation in developing cardiomyocytes

The embryonic development of the mammalian heart has been well documented at the anatomical and morphological level, but the underlying genetic regulation is poorly understood. A few key transcription factors and chromatin modification processes are known to be crucial in heart cell development, but the complete regulatory network of transcription factors and epigenetic factors involved remains to be determined. I contributed to a study in which ChIP-chip and expression microarray data were used to investigate the influence of post-translational modifications of the histone proteins and transcription factors (TFs) in the development of *Mus musculus* heart and muscle cells. First, we investigated the presence of four histone modifications and their relation to gene expression (see Section 3.1 and [88]). Second, we used ChIP-chip to determine the DNA binding sites of four known key transcription factors (see Section 3.2).

3.1 Histone modifications

3.1.1 Introduction

Post-translational histone modifications

Post-translational modifications of the histone proteins have been associated with changes in the transcription rate of genes from DNA adjacent to the modified nucleosomes [14, 15]. (See Section 1.2.1 for an introduction to histone modifications in transcriptional regulation.)

Distinct histone modifications have been suggested to act together, inasmuch as their combination determines the downstream effect on transcription [17]. This "histone code" hypothesis is still under discussion. In particular, there has been no agreement on whether histone modifications are working in combination to bring about transcription by recruiting specific transcription factors [89] or if the combinations of modifications rather are a consequence of active transcription [90, 91].

Liu *et al.* mapped 12 different types of histone H3 and H4 acetylation and methylation in 0.5 Mb of the *S. cerevisiae* genome, using ChIP-chip with single-nucleosome resolution. Some of the modifications were found to be transcription-dependent, whilst the others seemed to be independent of transcription [91].

A combination of one activating and one repressive histone modification has been associated with reduced gene transcription [92]. This finding could be explained by the histone code hypothesis, or simply by the fact that the repressive modification is dominant.

We investigated four different histone modifications, which have all previously been associated with gene activation, for their combinatorial effect on gene expression in three cell types. These modifications are two types of acetylation of lysine residues, and di- and tri-methylation of one specific lysine residue. Acetylation of lysine residues in the histone tails has been observed to have an activating effect, which can be explained by a change in charge. The addition of the acetyl group neutralises the positive charge of the lysine residue. The affinity of the histone protein to the negatively charged backbone of the DNA is reduced, and the chromatin structure is opened up [16]. The interplay between histone methylation and transcriptional regulation seems to be more diverse, as some methylation types have been associated with gene activation and others with gene repression, depending on which residues were being methylated and how many methyl groups were added. However, the modifications that we analysed (di- and tri-methylation of lysine 4 of histone H3) have been associated with gene activation only [15, 93], although the di-methylated form has also been observed near inactive genes in *S. cerevisiae* [15].

Previous approaches to the analysis of ChIP-chip for histone modifications

There are important differences between ChIP-chip against histone modifications and ChIP-chip against transcription factors. With transcription factors, it is safe to assume that the majority of genomic regions will not show a real binding site for that transcription factor. Hence, most reporters on the microarray will not indicate true enrichment, at least not when the tiling microarrays represent the whole genome or an unbiased subset of the genome. This situation is beneficial for the data preprocessing and for identifying ChIP-enriched regions, since most of the data can safely be assumed to show non-enrichment. With histone modifications, on the other hand, the degree and extent to which the genome shows a certain histone modification can only be guessed at the present time. This situation make estimation of the background distribution of reporter levels under non-enrichment difficult.

Moreover, transcription factor binding sites are highly localised point effects, meaning that the transcription factor binds at one specific position directly or indirectly to the DNA and the signal will show a peak shape around this position. The highest point of the signal peak will be as close to the actual binding site as the reporter-tiling on the microarray allows (see [94] for an extended discussion and for a derived model of TF ChIP-

chip data). With histone modifications, the enzyme which modifies the histone tail is unlikely to act on only one single histone protein, but will modify a number of nearby histones. A single-nucleosome resolution study of histone modifications in *S. cerevisiae* has shown that modifications occur in the form of broad modified domains and that adjacent nucleosomes mostly share the same modifications [91].

Bernstein *et al.* [95] have mapped di- and tri-methylation of histone H3 lysine 4 (H3K4me2/H3K4me3) and acetylation of histone H3 lysine 9 or lysine 14 (H3ac[1]) across the non-repetitive regions of chromosome 21 and 22 in *H. sapiens*, using ChIP-chip on oligonucleotide tiling microarrays with one 25mer reporter starting every 35 bp. To identify ChIP-enriched regions, they used a sliding window approach. A 400 bp sliding window was moved along the chromosome, and the authors tested whether the preprocessed reporter levels within the window were higher in the ChIP samples than in control DNA samples using a one-sided Wilcoxon rank sum test. Positions with a p-value less than 10^{-4} (alternative hypothesis: levels of reporters mapped inside the window are higher in the ChIP samples than in the control DNA samples) were considered to be enriched, and enriched positions less than 200 bp apart were merged into ChIP-enriched regions.

Liu *et al.* [91] conducted a ChIP-chip study for 12 different types of histone H3 and H4 acetylation in *S. cerevisiae*. They used an oligonucleotide microarray that tiled selected regions of the *S. cerevisiae* genome with one reporter every 20 bp. This dense reporter tiling allowed the authors to reach a single-nucleosome resolution. The positions of the nucleosomes had been determined by the authors in a previous study. ChIP-chip levels for each histone modification were summarised over the ~ 7 reporters per nucleosome. Nucleosome modification levels were then categorised according to the positions of the nucleosomes relative to annotated ORFs in the *S. cerevisiae* genome. For each type of histone modification, the nucleosome modification levels in each positional category were compared

[1] H3ac is used as a summary term for H3K9ac and H3K14ac.

with the levels in all other categories using two-tailed *t*-tests, and the resulting *p*-values were corrected for multiple testing using a false discovery rate (FDR) estimate [96].

Koch *et al.* [41] analysed histone modifications from ChIP-chip data in 1% of the human genome as part of the ENCODE project [97], using a Hidden Markov model (HMM) [98] approach. The microarrays that Koch and colleagues used, however, were cDNA arrays containing spotted PCR products of about 1 kb length. The advantage of using an HMM for finding ChIP-enriched regions, especially for finding possibly large regions enriched for histone modifications, is that the autocorrelation in nearby reporter match positions is modelled in the analysis. A large enriched region would correspond to a sequence of observations showing an "enriched" hidden state in the Viterbi path. However, a fairly good idea about the distributions of reporter levels in enriched and non-enriched regions is required for the construction of an appropriate HMM for ChIP-chip data, as the numbers of "non-enriched" and "enriched" hidden states and the respective emission distributions have to be set.

Li *et al.* [99] also suggested an HMM approach for finding enriched regions in ChIP-chip from high-resolution tiling microarrays and demonstrated its use on a data set of *p53* binding on *H. sapiens* chromosomes 21 and 22. The authors used a two-state HMM and different normal distributions for the emission probabilities of the two states.

3.1.2 Material and methods

We investigated the presence of four post-translational histone modifications using ChIP-chip on tiling microarrays. We also analysed gene expression microarray data for the relationship between histone modifications and transcript expression.

Chromatin immunoprecipitation data

ChIP-chip experiments were performed using three different *M. musculus* cell types: undifferentiated and differentiated skeletal muscle cells (cell line C2C12; I use the terms C2C12U for undifferentiated C2C12 cells and C2C12D for differentiated C2C12 cells), and cardiomyocytes (cell line HL1). Immunoprecipitation was performed with separate antibodies recognising four different post-translational histone modifications: two acetylation types, H3ac, which is histone H3 acetylated at lysine residues 9 and/or 14, and H4ac, which is histone H4 acetylated at lysine residues 5, 8, 12, and/or 16; and di- or tri-methylation of the lysine residue 4 of histone H3, denoted H3K4me2 and H3K4me3. For each antibody and cell-line, the whole ChIP-chip process was performed twice, resulting in two biological replicates per antibody-cell-type combination and 24 samples in total.

Initially, we also had 6 ChIP samples, two per cell type, using an antibody against Polymerase II. The resulting microarrays, however, showed strong artifacts during quality assessment and we decided to exclude them from further analysis.

Microarray design

ChIP microarray For this study, we used custom-designed oligonucleotide microarrays from NimbleGen Systems. The reporters on the ChIP microarrays were 50 nucleotides long. Human or mouse transcripts of 8,585 genes expressed in heart, skeletal or smooth muscle were selected from several sources[2]. All supplied gene identifiers were mapped to the Ensembl database (version 26). Human-mouse orthologs were identified and redundant entries were removed. Transcript coordinates corresponded to those annotated in the *mm5* assembly (NCBI build 33, May 2004) of the *M. musculus* genome.

Reporters were designed to match regions surrounding the transcription

[2] The sources included review publications of genes involved in muscle and heart development, e.g., [100], and lists of genes considered to be expressed in *H. sapiens* or *M. musculus* heart or muscle tissues in the Symatlas expression data [65]

start sites (TSSs) of the transcripts. The regions were selected as follows[3]. The human-mouse conserved non-coding blocks in the 5 kb region upstream of the TSSs and in the first intron up to 10 kb downstream of each TSS were considered. If no such block with $\geq$ 10% sequence conservation could be found for a transcript in the list, a fixed region extending from 2.2 kb upstream to 0.8 kb of the first intron was selected. Repeats in the selected regions were masked and 50mer reporters covering the selected regions were designed by NimbleGen Systems, with one reporter-matched genomic position starting every 85 bp where possible. Note that this is not a whole-genome tiling array, but rather it is similar to commercially available promoter tiling arrays in which selected promoter regions are densely tiled with reporters but the rest of the genome is not represented on the array.

Expression microarray Reporters on the expression microarray were 24 nucleotides long. Gene identifiers of the 8,585 selected genes (see above) were mapped to the Ensembl database (version 26), and all transcripts annotated for these genes were retrieved. Each transcript was represented on the expression microarray by 15 reporters.

Quality assessment

The first step of data analysis was a thorough quality assessment of the raw microarray data. The chromatin immunoprecipitation process and the hybridisation to the microarray are delicate processes, in which a number of artifacts can arise. First, I considered the spatial distribution of raw reporter intensities. This visualisation can show artifacts that were introduced if the array was mishandled, such as fingerprints, scratches, and edge effects, or incorrectly scanned. The raw intensities are visualised according to the position of the reporters on the microarray surface. The colour of the reporter-wise dots represents the raw intensity. One expects

[3] Steffen Grossman and Silke Sperling from the Max Planck Institute for Molecular Genetics in Berlin, Germany, decided which regions should be represented on the microarray and designed the microarray in collaboration with NimbleGen Systems.

a more or less homogeneous distribution of low and high intensities across the whole microarray surface. Regions with predominantly low or high intensities can arise due to hybridisation artifacts, such as for example fingerprints on the array surface.

Another quality assessment step was to assess the similarity of raw reporter intensities between microarrays. With our ChIP-chip data, within each cell type, all microarrays had a comparable *input* sample in the Cy3 channel, the untreated genomic DNA from cells of the same cell type. Thus, between all the arrays of one cell type, the raw Cy3 reporter intensities would show a strong correlation. In the microarrays that we finally used in the analysis, I observed Pearson correlation coefficients of 0.73 to 0.91 between the *input* samples.

Another factor that should be checked at this stage is whether the antibody has led to any enrichment. Functional antibodies should lead to some points lying above the diagonal in the scatter plot of ChIP reporter intensities versus *input* intensities [101]. Moreover, when positive control regions are visualised (see page 136), a possibly noise enrichment signal should already be obvious in the raw data. Positive control regions are genomic regions in which ChIP enrichment is to be expected a priori. If negative control regions are also known, a preliminary assessment of the antibody's specificity is possible as well. We did not have any control regions, but exploratory visualisations of the raw data indicated that specific enrichment was present.

ChIP array reporter remapping

The microarrays had been designed based on the *mm5* assembly of the *M. musculus* genome, which was an incomplete draft assembly. I used the Blast-like alignment tool *BLAT* [102] to re-map the reporters to the *mm8* assembly (NCBI build 36, February 2006), then the current draft of the genome (in March 2006). In the re-mapping step, I allowed for one mismatch per 50mer reporter. The implicit assumption is that if 49 out of 50 nucleotides are complementary that would be sufficient for hybridisation.

I did not consider a more complex hybridisation model, in which the position of the mismatch in the reporter sequence and its impact on secondary structure formation are taken into account. Reporters were mapped to 389,918 genomic positions of the *mm8* assembly. These positions are dispersed over 16,882 genome stretches that are covered by five or more reporters which uniquely match inside that stretch and have an offset between their matches' start positions of $\leq$ 100 bp, mostly exactly 85 bp, as intended in the array design for *mm5*.

ChIP microarray preprocessing

Intensities of each channel were normalised and glog-transformed[4] using the *vsn* method [25]. One common assumption of most normalisation methods, including *vsn*, is that the variation of most reporter levels does not reflect biological variation between samples/conditions (*input*, ChIP) but is non-biological variation, e.g., due to differences in sample processing and hybridisation. This assumption probably does not hold in this case, especially since the fraction of histones bearing post-translational modifications cannot safely be assumed to be small. The *vsn* normalisation was still useful to reduce unspecific errors in reporter measurements. However, it cannot be excluded that while the *vsn* certainly increased the signal-to-noise ratio, the *vsn* may also have masked weak ChIP enrichment. Log-ratio enrichment levels for each reporter were calculated by subtracting the preprocessed Cy3 (*input*) levels from Cy5 (ChIP) levels, which were both reported on a log scale by *vsn*.

Expression array reporter remapping

The reporters were mapped to the mouse genome assembly mm8 using BLAT [102], allowing up to one mismatch per 24mer reporter. Only reporters that uniquely matched one segment of 23 or 24 nucleotides in the *M. musculus* genome and that were mapped within the boundaries of

[4] $\mathrm{glog}_\Delta(x) = \log\left(x + \sqrt{x^2 + \Delta}\right)$

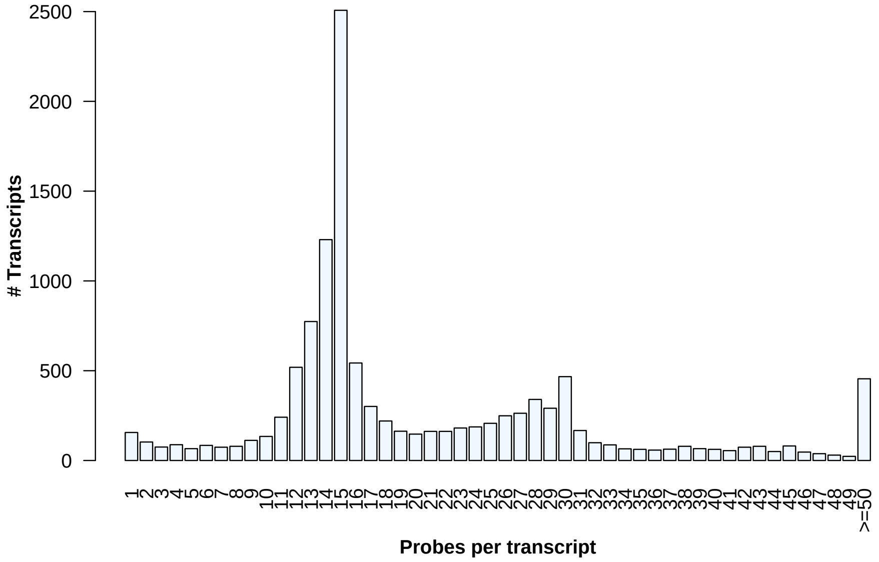


Figure 3.1: *Bar plot showing how many reporters were mapped within the boundaries of each of the 11,865 M. musculus transcripts annotated in the Ensembl database (version 39) that had at least one reporter mapped inside. #: "number of".*

any transcript annotated in the Ensembl database (version 39, June 2006) were further analysed. 11,865 transcripts had at least one reporter mapped within their boundaries, the majority of which (8,130) had 15 or more reporters mapped to them (see Figure 3.1). Alternative transcripts of the same gene were allowed to share reporters.

Expression array data preprocessing

Transcript expression levels were computed from raw reporters intensities using the robust multiarray analysis (RMA) method [24]: raw intensities were background-corrected by subtracting a global background estimate, quantile normalised, log-transformed (base 2) and summarised into transcript expression levels using the median polish procedure.

Identification of ChIP-enriched regions

To identify genomic regions in which the histones bear one or more of the four investigated modifications (H4ac, H3ac, H3K4me2, H3K4me3), we employed a three-step procedure:

1. smoothing reporter levels using a sliding-window approach
2. determining a threshold above which smoothed reporter levels should indicate *enrichment* at that genomic position
3. combining adjacent enriched positions into ChIP-enriched regions.

Smoothing of reporter levels Individual reporters measure the same amount of DNA with varying efficiency due to reporter sequence characteristics, such as GC content, secondary structure, and cross-hybridisation [103]. Normalised reporter levels were averaged across the two biological replicates and smoothed along chromosomal coordinates using a sliding-window method. To ameliorate the reporter effects as well as the stochastic noise, I performed a smoothing over individual reporter intensities before looking for ChIP-enriched regions. A window of 800 bp width was slid along each chromosome and the intensity at each reporter-matched genomic position x_0 was replaced by the median over the intensities of those reporters inside the window centred at x_0. Factors taken into account in the choice of the sliding-window width were the size distribution of DNA fragments after sonication (600–1200 bp, mean: 900 bp) and the spacing between reporter matches on the genome (median: 86 bp, first quartile: 84 bp, third quartile: 216 bp). We chose a window-width of 800 bp, which was slightly less than the average fragment size and meant that most windows included $>=$ 4 reporters. With this window width, we could be sure that the signal is not smoothed over many fragments and was calculated as the median over at least four reporters. At any position x_0 at which the window comprised less than three reporter-matched positions, the smoothed level was flagged as missing, as the data were insufficient to provide information about ChIP enrichment at such a position.

Enrichment threshold As threshold for enrichment, we took a high quantile of a simulated distribution of smoothed reporter levels under non-enrichment. We permuted the reporter-matched genomic positions and repeated the smoothing procedure on the preprocessed reporter levels with the permuted positions. The 99% quantile of this estimated, empirical "null" distribution of smoothed reporter levels made up the threshold for each antibody. Note that this 99% quantile was an arbitrary choice and was not used for assessing significance of determined ChIP-enriched regions. To allow for different efficiencies of antibodies, a threshold was defined for each type of histone modification separately.

Combining enriched positions into ChIP-enriched regions We called a reporter-matched position *enriched* if it had a smoothed reporter level greater than the threshold. Enriched positions were merged into regions using an agglomerative approach. Initially, each position was considered to be an individual region. Two regions r_i and r_j were combined into a single region if any enriched position in r_i was less than 600 bp apart from any enriched position in r_j. This procedure corresponds to the *single-linkage* agglomeration method in hierarchical clustering. ChIP-enriched regions containing fewer than three reporter-mapped positions were discarded. To require such a minimum number of enriched positions might at first seem redundant with the smoothing median computation (since a smoothed reporter intensity is already the median of all the reporter intensities in the window), but it plays an important role in reporter-sparse regions, where a window might only contain one, or a few reporters, and we wanted to avoid making calls in such regions.

Co-occurrence of ChIP-enriched regions with different modifications

Two ChIP-enriched regions r_{i,h_1} and r_{j,h_2}, which were enriched for different histone modifications, were considered to be co-occurring if

$$\text{length}\left(r_{i,h_1} \cap r_{j,h_2}\right) \geq 0.75 \cdot \min\left(\text{length}(r_{i,h_1}), \text{length}(r_{j,h_2})\right) \tag{3.1}$$

Code	**Histone modifications for which such a region shows enrichment**
H3ac	histone H3 acetylated at lysine residues 9 and/or 13
H3K4me2	histone H3 di-methylated at lysine residue 4
H3K4me3	histone H3 tri-methylated at lysine residue 4
H4ac	histone H4 acetylated at lysine residues 5, 8, 12, and/or 16
H3acK4me2	H3ac & H3K4me2
H3acK4me3	H3ac & H3K4me3
H3acK4me2/3	H3ac & H3K4me2 & H3K4me3
H4ac-H3ac	H4ac & H3ac
H4ac-H3acK4me2	H4ac & H3ac & H3K4me2
H4ac-H3acK4me3	H4ac & H3ac & H3K4me3
H4ac-H3acK4me2/3	H4ac & H3ac & H3K4me2 & H3K4me3

Table 3.1: *Codes used to indicate from which kinds of ChIP-enriched regions combined enriched regions were constructed.*

where "∩" denotes region intersection. Thus, the two regions were considered to be co-occurring if 75% or more of the shorter of the two regions (in terms of base pairs) was also included in the longer of the two regions. The two co-occurring regions were merged into a combined region with start and end being the extremal positions of the two co-occurring regions. Other regions could in turn co-occur with the combined region and be included into a larger combined region, and so on.

Thus, we obtained all (combined) genomic regions enriched by ChIP with any combination of the four histone modifications. We used a coding scheme to indicate the combination of the four antibodies (see Table 3.1).

Relating ChIP-enriched regions to transcript expression

A ChIP-enriched region was considered to be associated to a transcript if its centre position was located less than 10 kb upstream of the TSS, or between the TSS and the 3′ end of the transcript. A region could be associated to more than one transcript, for example in the case of nearby alternative transcripts of the same gene or other tandem, convergent or divergent transcripts (see Figure 1.1).

Transcripts for which we had expression levels measured on the microarray were stratified into categories depending on to what kind(s) of (combined) ChIP-enriched regions they were associated to. Transcripts without any associated region were assigned to a separate category. I compared the expression levels between categories in a formal way using a linear model that related the presence or absence of histone modifications to transcript expression levels.

Software

The computational analysis of the data was performed in R, using my software package *Ringo* (see Section 5.2) and other Bioconductor [104] packages.

3.1.3 Results

First I performed an extensive quality assessment of the data and performed a preprocessing step to increase the signal-to-noise ratio in the data. During the quality assessment step, I observed a number of problematic microarrays that showed hybridisation artifacts[5]. See Figure 3.2 for an example picture of one microarray that showed hybridisation artifacts. The bright rim on the picture suggests that all reporters near the rim of the array display very high raw intensities, which is probably due to mishandling of the microarray rather than due to biological effects. The second artifact is the wave pattern on the surface. This effect is known as a Moiré pattern in image processing and emerged during the scanning process of the microarray.

[5] Hybridisation and scanning of the microarrays were conducted by employees of NimbleGen Systems. We had no influence on these steps.

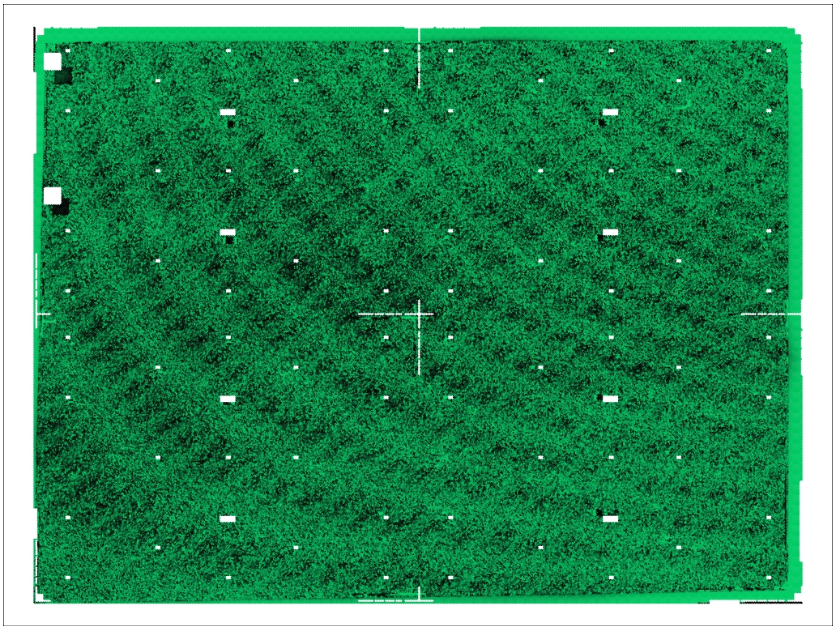


Figure 3.2: *This picture shows the spatial distribution of raw reporter intensities of one example microarray that shows two artifacts. Coordinates in the picture correspond to coordinates on the surface of the microarray. The colour of the dots represents the value of the raw reporter intensity, with brighter shades of green corresponding to higher intensities. For well-hybridised microarrays, a homogeneous picture can be expected. Here, the bright rim of the array and the wave-like patterning are artifacts introduced during hybridisation and scanning of the microarray.*

After communication with the microarray manufacturer, the hybridisations of most below-standard microarrays were redone. Almost all of the new hybridisations resulted in microarrays of acceptable quality. However, for the ChIP-chip experiments with the antibody against Polymerase II, we did not receive acceptable replacement microarrays and therefore decided to drop these experiments from the study.

I applied a custom heuristic algorithm (Section 3.1.2, page 52) for the detection of modified histone sites in the ChIP-chip data, and hence obtained the positions and co-occurrence of ChIP enriched regions for the four histone modifications in each of the three cell types (HL1, C2C12U, C2C12D). The number of ChIP-enriched regions for each type of histone modifica-

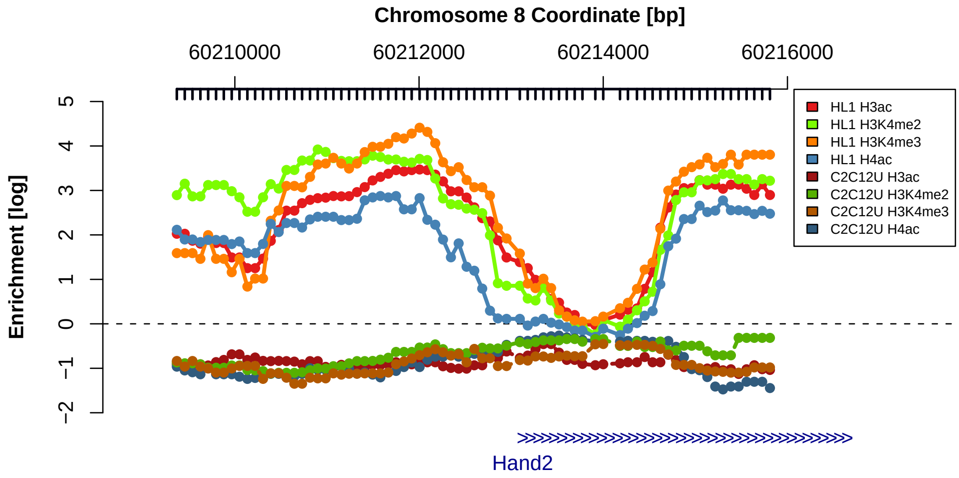


Figure 3.3: *Normalised reporter levels at the TSS of the gene* Hand2 *in HL1 and C2C12U cells. The brighter lines correspond to the four histone modifications in HL1 cells, the darker ones to the histone modifications in undifferentiated C2C12 cells. The ticks below the genomic coordinate axis on top indicate genomic positions matched by reporters on the microarray. The blue arrows on the bottom mark the* Hand2 *gene with the arrow direction indicating its transcription direction, i.e. the gene is located on the Watson strand.*

tion was similar in all three cell lines (see Table 3.2), although the genomic positions of ChIP-enriched regions varied between cell types. To illustrate, Figure 3.3 shows one example genome region which is enriched for all four histone modifications in HL1 cells, but shows no enrichment in undifferentiated (and differentiated) skeletal muscle cells. This genome region contains the TSS of the gene Hand2, which is a transcription factor

Modification	**HL1**	**C2C12U**	**C2C12D**
H4ac	2,657	2,682	2,802
H3ac	2,950	2,765	2,982
H3K4me2	3,087	2,999	2,940
H3K4me3	3,051	3,195	3,260

Table 3.2: *Frequency of ChIP-enriched regions for the four histone modifications in the three investigated cell types.*

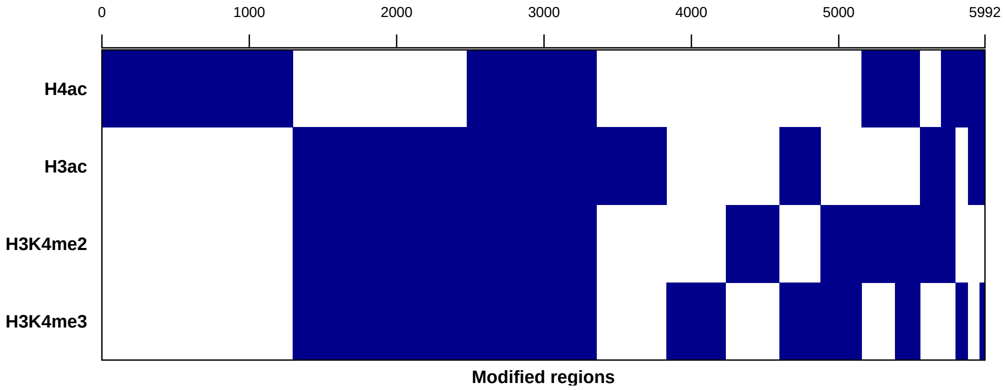


Figure 3.4: *This figure displays the 5,992 modified regions discovered in HL1 cells (x-axis) and which combinations of histone modifications are enriched in these regions (y-axis). Dark-blue indicates that a histone modification is enriched in a region, white indicates it is not. The regions are ordered by the frequency of the combination among all modified regions.*

that is required for the development of the right heart ventricle [105, 100]. We saw that the ChIP-enriched regions of the four histone modifications frequently co-occurred, i.e. showed an overlap of more than 75% of their genomic width (Section 3.1.2, page 53). Such combined enriched regions of overlapping ChIP-enriched regions from single histone modifications were classified according to the combination of histone modifications (see Table 3.1 for the class labels). For the sake of consistency, ChIP-enriched regions for each modification that showed less than 75% overlap (or no overlap at all) with any ChIP-enriched regions of other modifications, were grouped into four single-modification classes. Thus, the 11,745 ChIP-enriched regions of the four individual histone modifications were merged into 5,992 modified regions in HL1 cells; 6,202 in C2C12U; 6,125 in C2C12D. Figure 3.4 shows all the combined regions for the HL1 cells. The regions are grouped by classes and the classes are ordered according to the size of the class among all combined regions. The most common type are regions that only show enrichment for H4ac (1,295 regions), followed by combined regions of class H3acK4me2/3, i.e. regions that show co-occurrence of the three analysed modifications of histone H3 (1,181 regions).

I assessed the relationship between histone modifications upstream and within transcripts, and the expression of these transcripts, in detail using the expression microarray data from the three cell lines.

Histone modifications should be considered in combination rather than individually in order to get a precise description of the association between histone modifications and transcription. Tri-methylation of lysine 4 of histone H3 (H3K4me3) has previously been reported to coincide with elevated transcript levels [106], but in our data we saw that this in not the case if some histones in that genomic stretch were only di-methylated at that residue. Acetylation of histone H3 residues is a far more reliable marker for increased transcription (see Figure 3.5). These relations were highly similar in all three cell types.

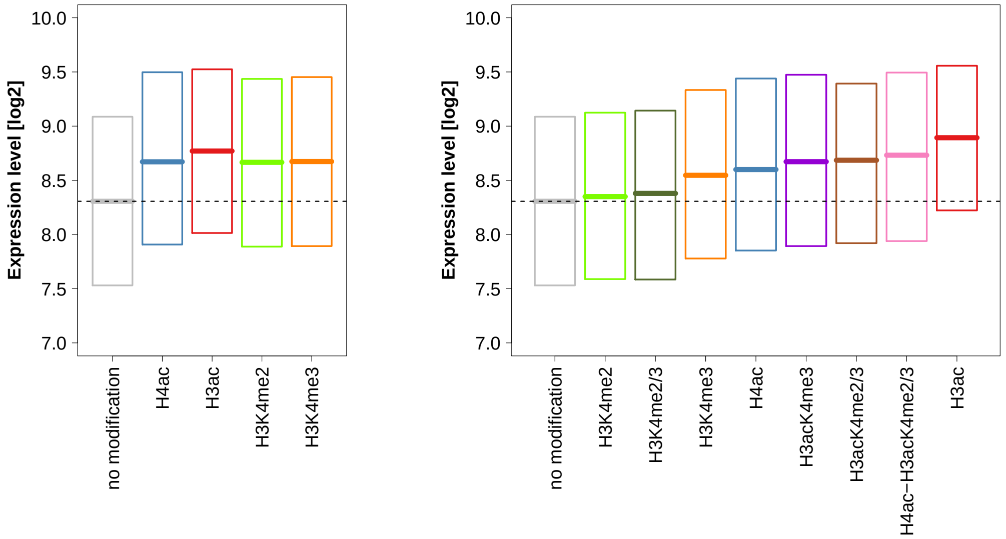
A more formal approach was required to assess the relation between histone modifications and transcript expression.

Linear model: transcript expression related to histone modifications

I employed a linear model relating the absolute expression level of every transcript to cell type, presence of associated ChIP-enriched regions for each histone modification, median reporter GC content and interactions between the histone modification terms. The model (in `S-plus/R` formula notation):

```
y ~ H3ac + H4ac + H3K4me2 + H3K4me3
    + GC + cell.type + H3ac:H4ac + H4ac:H3K4me2
    + H4ac:H3K4me3 + H4ac:H3K4me2:H3K4me3
    + H3ac:H4ac:H3K4me2 + H3ac:H4ac:H3K4me3
    + H3ac:H4ac:H3K4me2:H3K4me3 + H3ac:H3K4me2:H3K4me3
    + H3ac:H3K4me2 + H3ac:H3K4me3 + H3K4me2:H3K4me3
```

where
`y`: log2 expression level of transcript in cell line
`H3ac`: indicator variable for transcript's associated modification H3ac; it is



(a) (b)

Figure 3.5: *Relationship between histone modifications and expression. The boxes show the quartiles of the empirical distributions of expression levels for the transcript categories. The lower and upper border indicate the first and third quartile, respectively, and the bold line in the middle of each box marks the median. Transcripts were categorised according to which (combination) of histone modification(s) had been associated to them.* (a) *When the four modifications are considered independently of each other, they are similarly associated with elevated expression levels compared to the no-modification group.* (b) *When considered in combination, we see different associations. H3ac combined with the other three modifications shows comparatively lower transcript levels than H3ac alone. Transcripts that are associated to H3K4me3 and H3K4me2 in combination and H3K4me2 alone show no elevated expression levels.*

1 if at least one ChIP-enriched region for H3ac is associated to the transcript, 0 otherwise.

`H4ac, H3K4me2, H3K4me3` : analogous to H3ac

`GC` : median GC content (in percent) of the reporters on the expression microarray that had been mapped to transcript

`cell.type` : factor variable, one of "C2C12U", "C2C12D" or "HL1"

and the expression "`A:B`" denotes the interaction term between predictors `A` and `B`.

The function `lm` of R (version 2.4) was used to fit the model. Table 3.3 shows the resulting coefficients of the model and an assessment of whether these are significantly different from zero. The table also specifies for each predictor variable the coefficient estimate, its standard error and the p-value for the null hypothesis that the coefficient is equal to 0 (two-tailed one-sample t-test, alternative hypothesis: the coefficient is different from 0). The p-values were corrected for multiple testing using the Bonferroni correction method [107].

Term	**Estimate**	**Std. error**	t**-value**	p**-value**	**Sig.**
Intercept	4.26	0.06	76.24	$< 2 \cdot 10^{-16}$	$\star$
H3ac	0.58	0.05	11.07	$< 2 \cdot 10^{-16}$	$\star$
H4ac	0.38	0.03	13.28	$< 2 \cdot 10^{-16}$	$\star$
H3K4me2	0.06	0.05	1.19	1	
H3K4me3	0.39	0.04	9.01	$< 2 \cdot 10^{-16}$	$\star$
GC	8.22	0.02	75.22	$< 2 \cdot 10^{-16}$	$\star$
cell.type.C2C12U	-0.02	0.02	-0.99	1	
cell.type.C2C12D	0	0.02	-0.25	1	
H3ac:H4ac	-0.23	0.10	-2.34	0.37	
H4ac:H3K4me2	0.08	0.09	0.93	0.35	
H4ac:H3K4me3	-0.33	0.09	-4.44	0.0031	$\star$
H3ac:H3K4me2	-0.23	0.11	-2.09	0.7	
H3ac:H3K4me3	-0.37	0.11	-3.11	$1.7 \cdot 10^{-4}$	$\star$
H3K4me2:H3K4me3	-0.33	0.08	-4.10	$7.9 \cdot 10^{-4}$	$\star$
H4ac:H3K4me2:H3K4me3	0.05	0.15	0.31	1	
H3ac:H4ac:H3K4me2	-0.11	0.18	-0.63	1	
H3ac:H4ac:H3K4me3	0.24	0.16	2.33	0.38	
H3ac:H3K4me2:H3K4me3	0.46	0.14	3.40	0.013	$\star$
H3ac:H4ac:H3K4me2:H3K4me3	-0.08	0.23	-0.33	1	

Table 3.3: *Coefficients of the linear model that relates transcript expression to presence/absence of histone modifications according to our ChIP-chip data. For an explanation of the predictors, see Section 3.1.3. The column "***Sig.***" indicates predictors that are significantly different from* 0 *according to the model by* $\star$*. p-values have been corrected for multiple testing using the Bonferroni procedure.* Std.:*"Standard".*

The intercept, and the predictors for H4ac, H3ac and H3K4me3 are significantly different from zero, as are a number of interaction terms between modifications. The median GC content of each transcript's reporters on the expression microarray has a significant positive effect on the measured transcript expression level. Cell type is not deemed to be a significant predictor in the model.

3.1.4 Discussion

We analysed ChIP-chip data for four histone modifications: acetylation of lysine residues in histones H3 and H4 and di-/tri-methylation of lysine residue 4 of histone H3. ChIP enrichment for these modifications was measured in three cell types, using a custom developed two-channel oligonucleotide microarray. The reporters on the microarray represented selected regions surrounding the TSSs of known muscle- and heart-specific genes. We investigated the relation of regions that showed enrichments for these modifications to annotated transcripts and the expression levels of these transcripts. Transcript expression levels were measured using expression microarrays of another custom design.

One caveat about using expression microarrays for this purpose is that the arrays did not directly measure the quantity that we were interested in, i.e. the transcription rate of genes related to ChIP-enriched regions. Expression microarrays measure the mRNA steady-state levels, which strongly depend on the mRNA transcription rates, but are also influenced by the mRNA degradation rates [108]. Initially, we had also intended to measure the occupancy levels of Polymerase II (Pol II) in the transcribed regions with ChIP-chip, as these are suggested to provide a better estimate of the actual transcription rate than expression microarrays can provide [91]. However, the respective ChIP-chip microarrays with the Pol II samples were found to show serious artifacts from the hybridisation, and were therefore excluded from further analyses.

The other microarrays partially showed artifacts as well, but the microarray manufacturer provided us with replacement hybridisations of acceptable quality in those cases. Microarrays may be an established technology by now, but this example shows that thorough quality assessment is still mandatory. The sample purification, hybridisation and scanning of the microarrays are such delicate processes that even microarrays generated by dedicated specialists, as in our case, can show serious artifacts. Such artifacts lead to biologically irrelevant findings from microarray studies, if not properly accounted for. The results from ChIP-chip studies are meant to generate hypotheses that need to be validated in more accurate, small-scale experiments. To avoid wasting time on testing inaccurate hypotheses that are due to artifacts in the data, concise quality assessment of the data is required, before searching for ChIP-enriched regions.

We developed a heuristic algorithm for finding genomic regions enriched for histone modifications. This algorithm provides an estimated threshold, above which smoothed reporter levels are considered to be enriched. The threshold is estimated in a three-step procedure: the reporter match positions for each chromosome are permuted at random; the levels of the permuted reporters are smoothed using the same sliding-window method as for the original levels; finally the threshold for enrichment is set equal to a chosen quantile of the distribution of smoothed reporter levels after permutation. This algorithm worked well for the ChIP-chip data in this study. However, the algorithm would have to be tuned to allow detection of histone modified sites in other ChIP-chip data with high sensitivity and specificity. In Section 5.2, I describe a more generally applicable algorithm.

We did not perform a straightforward Wilcoxon rank sum test to compare ChIP and *input* reporter levels in a sliding window, as done by Bernstein *et al.* (see Section 3.1.1) to find ChIP-enriched regions for two reasons. First, with only two ChIP samples per modification and an average reporter spacing of ≥ 85 bp, the test would have low power unless a window of large size was used. Second, one assumption for getting meaningful *p*-values with the Wilcoxon rank sum test is that all orderings of the ranks are equally likely. However, levels of reporters with adjacent genomic match positions cannot be considered to be independently identically distributed, since sample DNA fragments after sonication are longer than the distance between the positions. Hence, not all rankings of reporter levels are equally likely with ChIP-chip data and thus the *p*-values of the Wilcoxon test are not meaningful.

The algorithm used does not make strong assumptions about the data, besides smoothed levels in ChIP-enriched regions being stochastically larger[6] than the smoothed levels in non-enriched regions and enriched regions encompassing a certain number of nearby reporter match positions. Our heuristic algorithm is simpler than an HMM method. An HMM approach (see Section 3.1.1) would have required further assumptions about the distributions of reporter levels in enriched and non-enriched regions and about how many ChIP enrichment states there are for histone modification. If these assumptions are appropriately stated, however, the HMM may be more powerful in detecting enrichments.

Unlike the data of Liu *et al.* [91], our ChIP-chip data did not have a single-nucleosome resolution, so our findings might partly be due to aggregate signals over two or more nucleosomes. However, Liu *et al.* [91] have stated that the modification patterns occur as broad regions and rarely differ between adjacent nucleosomes. The resolution of the microarrays that we used is approximately one reporter every 85 bp (in most regions). Thus, only a small part of our findings could be due to inappropriate averaging over many nucleosomes.

ChIP-enriched regions of the four histone modifications were frequently found to co-occur (Figure 3.4) in the same genomic positions. Whether two regions were considered to be co-occurring was based on the arbitrary requirement that the two regions needed to overlap to at least 75% of their genomic width. However, which fractions of the whole chromatin are marked by each histone modification is currently unknown, and thus an appropriate null model for the overlap of histone modifications cannot be formulated. If such a model existed, more meaningful requirements for region overlap could be derived from it.

Our results indicate that histone modifications form a combinatorial code, in the sense that different combinations lead to distinct outcomes with respect to observed expression levels. Individual modifications were found

[6] A random variable X is said to be *stochastically larger* than the random variable Y, if their cumulative distribution functions $F_X(x), F_Y(y)$ satisfy the inequality $F_X(a) \geq F_Y(a)$ for all a and there exists at least one a_0, for which $F_X(a_0) > F_Y(a_0)$.

to be approximately equally associated with higher transcript levels, as previously reported. However, when transcripts are stratified by co-occurrence of the measured histone modifications in their upstream or downstream regions, certain combinations were found to be associated with higher expression levels, while other combinations were not related to an increase in expression at all.

In a more formal approach, I analysed the effect of each histone modification on transcript expression levels, using a linear model (Section 3.1.3). The cell type, in which the expression level has been measured, and the median GC content of each transcript's reporters were additional predictors in the model. The coefficients of the model suggest an interpretation that is in line with the previous results. Of the four modifications, presence of H4ac, H3ac and H3K4me3 seem to have a significant and positive effect on gene expression, while by itself H3K4me2 is not associated with changes in expression rate. Moreover, the fact that a number of interaction terms between the modifications are significantly different from zero indicates that the *combination* of present modifications need to be considered for a clear description of the association between histone modifications and gene expression. In combination with other modifications, H3K4me2 can also have a significant association with transcript expression.

The significant effect of the GC content indicates that this reporter-specific effect has not been cancelled out during the summarisation of reporter levels into transcript levels. However, due to the required remapping of reporters to the newer genome assembly, transcripts are represented by a variable number of reporters per transcript. Thus all transcripts' reporters cannot be expected to show an equal range of GC proportions. Therefore, the median GC content should be taken into account as a predictor in the transcript expression level, and the model indicates this significant effect. A model that excludes the reporter GC would be less able to explain the expression data.

The highly significant intercept is to be expected, since there certainly are other unobserved factors apart from the four analysed histone modifications that influence the expression levels of the transcripts, which the intercept summarises. These unobserved factors likely include other histone modifications, transcription factors, further epigenetic factors, the decay rate of the transcripts' mRNAs, and the availability of nucleotides and energy for the transcription process. Since the expression levels are specified on a logarithmic scale, the intercept may also contain any factors of proportionality between the predictors and the expression level[7].The fact that the cell line variable was not deemed to be a significant predictor in the model indicated that the associations of histone modifications to transcript expression level are comparable in the three analysed cell types.

Our findings about the four observed histone modifications agree with the histone code hypothesis [17], as long as we agree to a rather lenient definition of the word "code". For a strict definition of a code being a set of rules for converting one form of information into another form or representation, neither all of the rules nor all components of the input information are known yet. Thus it remains to be seen whether such a histone code does exist.

Histone modifications may primarily function as signalling markers for specific effectors, and thus increase the combinatorial possibilities amongst many such markers in the regulation of transcription. The four analysed modifications are only a subset of all currently known histone tail modifications (see [13] for a recent review), and thus these findings may only represent a piece of the puzzle. A more complete understanding of the role of histone modifications in transcriptional regulation will only emerge once the interactions of multiple modifications haven been considered in detail. In addition, enzymes that modify histone tails are known to have other, non-histone substrates as well. For example, the enzyme SETD7 (SET9), which can methylate histone H3 at lysine 4, was also shown to methylate P53 [109]. Hence, observed associations between histone modifications and gene activation or repression could at least partly be byproducts of the effect on these other targets rather than the histone

[7] since $\log_2(c \cdot x) = \log_2(c) + \log_2(x)$

modifications being directly involved in gene activation.

Without further experiments, we cannot say what the relationship is between the observed histone modifications in genomic regions and the chromatin structure of that region. Recent reports suggest that the same modification state can have opposite functional results, depending on which proteins bind to the modified histones. H3K4me3 and H34me2 can be recognised by both BPTF and ING2. BPTF is part of the NURF chromatin-remodelling complex that activates transcription [110]. In contrast, ING2, which is part of a deacetylase transcription repressor complex, can also bind to H3K4me3 and H3K4me2 [111].

Post-translational histone modifications are certainly not the single determinant of chromatin structure formation. For example, nucleosomes that contain variant histone proteins in place of standard histone proteins, such as H3.3 in place of H3, are seen in characteristic locations with respect to actively transcribed regions, indicating that these may be involved in modifying the chromatin structure [112]. There are indications that the histone variants are mainly responsible for inheritance of gene activation status and histone modifications through cell division (reviewed in [113]). Accordingly histone variants would seem to be more important in the hierarchy of transcriptional regulation than histone modifications.

Further results of this study can be found in [88].

3.2 Transcription factor binding events

3.2.1 Introduction

Subsequent to the ChIP-chip study of histone modifications, we analysed the binding events of four transcription factors in the *M. musculus* cardiomyocyte cell line HL1. The four transcription factors, GATA4, MEF2A, NKX2.5, and SRF, are all known to be involved in different stages of heart development [100, 114].

MEF2A is a member of the *myocyte enhancer factor* MEF family of transcription factors that has been associated with the differentiation of all muscle cell types [100], but has also been found to be involved in post-synaptic differentiation in the brain [115]. NKX2.5 has been described as a cardiac transcription factor and its expression was found to be crucial for the development of the heart's conduction system [116, 100]. GATA4 is a member of a family of zinc finger transcription factors that bind a core GATA motif and are expressed during development in the heart and tissues derived from the endoderm [117]. Mutations of GATA4 lead to a number of congenital heart defects [100]. SRF regulates the development of all muscle cell types, and *Srf*$^{-/-}$ knock-out mice embryos fail to form a mesoderm and die at an early stage of embryonic development. A heart-specific silencing of SRF also results in embryonic lethality due to heart chamber malformation [114].

The genes *Gata4*, *Mef2a*, *Nkx2.5*, and *Srf* are highly conserved between *M. musculus* and *H. sapiens* [100, 114].

3.2.2 Material and methods

Microarray design For this study, we designed an oligonucleotide microarray for ChIP-chip analysis. This new platform superseded the microarray platform we had used in the histone ChIP-chip study. This new platform consisted of a set of two microarrays with 390k reporters each. We compiled a list of genes to be represented. This list includes genes that are known to play a role in heart and muscle development, all transcription factors annotated in the TRANSFAC database [118] and known positive control genes, for which binding sites of the four TFs analysed had been reported previously. For these genes, we used the following method to select genomic regions to be represented by reporters on the microarray. We obtained the coordinates of all transcripts of these genes, as annotated in the Ensembl database [34](release 39, June 2006). For each transcript, one region extending from 2 kb upstream to 100 bp downstream of the annotated TSS was taken. Additionally, the conserved non-coding blocks in the

10 kb upstream genome region and in the 3 kb downstream region of the annotated TSSs were taken as additional regions. Bases were considered to be conserved if annotated with a PhastCons [119] score ≥ 0.2. The selected genome regions were merged if less than 300 bp apart and extended to a minimum size of at least 1 kb. For the selected genome regions, which encompassed approximately 89 Mb, isothermal reporters[8] of 50–60 bp length were designed by NimbleGen Systems. The design resulted in a set of two microarrays with 390,000 reporters each, thus 780,000 reporters in total representing the selected genome regions. I remapped the final list of reporters to the mouse genome build *mm8* and reporters with multiple hits in the genome were excluded from further analyses. After remapping, 17,814 transcripts of 12,942 genes are represented on the microarray, each by 1 to 613 reporters with a median number of 78 reporters (mean: 81.8) per transcript.

ChIP-chip data The enrichment of the four TFs (GATA4, MEF2A, NKX2.5, and SRF) was measured in the cardiomyocyte cell line HL1, which we had already used for histone ChIP-chip (Section 3.1). For each of the four transcription factors, the ChIP experiment was performed twice, and the ChIP samples were hybridised against *input* DNA samples. Hence, there were two biological replicates per transcription factor.

Data preprocessing The raw reporter intensities were normalised and glog-transformed using the *vsn* method [25], separately for Cy3 and Cy5 channel. Preprocessed reporter levels were then computed by subtracting the normalised Cy3 from the Cy5 levels, yielding reporter-wise enrichment as glog fold changes between ChIP and *input* DNA.

Since each sample had been hybridised to two microarrays that contained two non-overlapping sets of reporters, I normalised the data in two batches, first the microarrays with the first reporter set and then the arrays with the second set. The control reporters on the microarrays, whose main

[8] Isothermal array design aims to achieve uniform reporter performance, by adjusting each reporter's genome position and length to obtain an approximately equal melting temperature across the entire reporter set of the microarray [120].

purpose is to act as guiding points during the scanning of the microarrays, were dropped from the data prior to normalisation. After normalisation, the results from the two microarrays per sample were combined, yielding an expression matrix with levels for 764,087 reporters in 8 samples, two samples per transcription factor.

Finding ChIP-enriched regions The two replicate samples per transcription factor were analysed together. We employed a two-step approach for finding ChIP-enriched regions, similar to the one for identifying regions enriched for histone modifications (Section 3.1.2), but differing in the choice of enrichment threshold. First, normalised reporter levels were smoothed using a sliding window. A window of 600 bp width was slid along the genome and the score at each genomic position, which is matched by a reporter, was calculated as the median over all reporter levels within the window. Positions for which the window encompassed less than four reporters were excluded from further analyses. I computed a threshold, above which smoothed reporter levels should considered to be enriched. See Section 5.2.3 (page 138ff.) for details on the heuristic algorithm that I used to compute this threshold, and implemented in *Ringo*. Enriched genome positions that were less than 210 bp apart from each other were merged into a single ChIP-enriched region. ChIP-enriched regions were identified individually for each TF.

Relating ChIP-enriched regions to transcripts An enriched region was counted as being related to a transcript if the region's centre position was located less than 10 kb upstream of the transcript's TSS or between its start and end coordinates.

Software For this analysis, we mainly used my R/Bioconductor package *Ringo* (see Section 5.2).

3.2.3 Results

We investigated the *in vivo* binding sites of the transcription factors GATA4, MEF2A, NKX2.5 and SRF, using a custom-design microarray plat-

TF	**# enriched regions**	**mean # related genes / region**	**# unique related genes**	**mean # related regions / gene**
GATA4	447	1.02	345	1.32
MEF2A	999	1.04	701	1.49
NKX2.5	383	1.00	276	1.39
SRF	1,335	1.09	1,150	1.26

Table 3.4: *Numbers of ChIP-enriched regions for the four transcription factors and of genes related to them. The table also shows the average number of genes that every enriched region is related to and the average number of ChIP-enriched regions per gene for each TF. #: "number of".*

form with reporters representing potentially regulatory regions at the TSSs of 17,814 transcripts. Based on our previous experience with ChIP-chip microarrays from the same manufacturer, I performed an extensive quality assessment of the raw data. All microarrays were deemed to be of acceptable quality.

For each TF, we found multiple ChIP-enriched regions. Table 3.4 shows the numbers of enriched regions per TF and how many genes these were considered to be related to. Most enriched regions are related to the transcripts of one gene only, but every gene that has enriched regions of an TF related to them on average has about 1.35 of them. In total, 1,671 genes had one or more regions enriched for any of the four TFs and of these:

- 1,173 showed enrichment for one TF only
- 286 showed enrichment for two of the TFs
- 121 showed enrichment for three TFs and
- 91 genes were related to enriched regions of all four investigated TFs.

Figure 3.6 displays the transcription factor ChIP-chip reporter levels at the TSSs of the genes *Tbx20* and *Tbx3*. *Tbx20* has associated ChIP-enriched regions for all four transcription factors upstream and downstream of the TSS. *Tbx3*, on the other hand, only shows enrichment for SRF and MEF2A, downstream of the two TSSs; and these enriched regions have a lower maximal level in comparison to the enrichments seen for *Tbx20*.

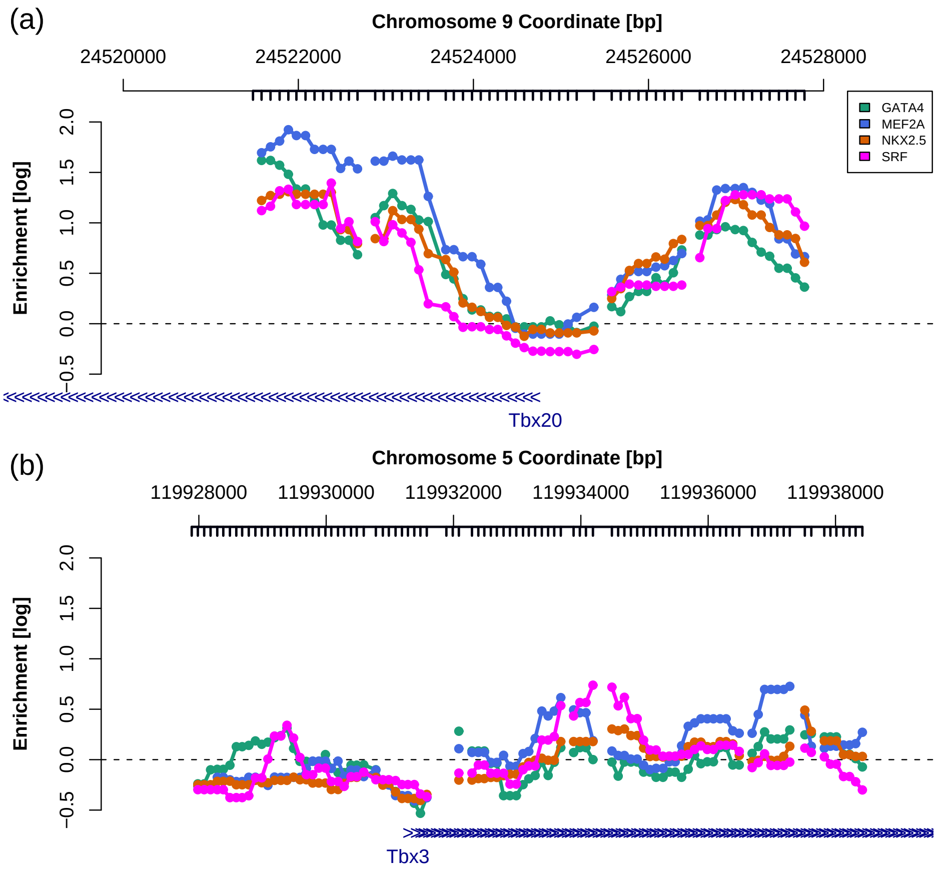


Figure 3.6: *Heart transcription factor enrichments at the TSSs of the genes Tbx20* ((a), *top panel) and Tbx3* ((b), *bottom panel). The ticks below the genomic coordinate axis on top indicate genomic positions matched by reporters on the microarray. The blue arrows on the bottom mark the transcripts of the genes Tbx20 and Tbx3, with the arrows indicating their respective transcription direction. Tbx3 has two transcripts with separate TSSs in this region.*

Figure 3.7 displays a subset of the genes that showed ChIP-enriched regions for the four transcription factors. The displayed genes are the four analysed TFs, genes that show enrichment for all four TFs and a few selected genes that have previously been reported to be regulating cardiomyocyte development.

When comparing the TF data with the histone data (Section 3.1), we observed that some of the enriched regions for TF enrichment were almost completely overlapping enriched regions for histone modifications.

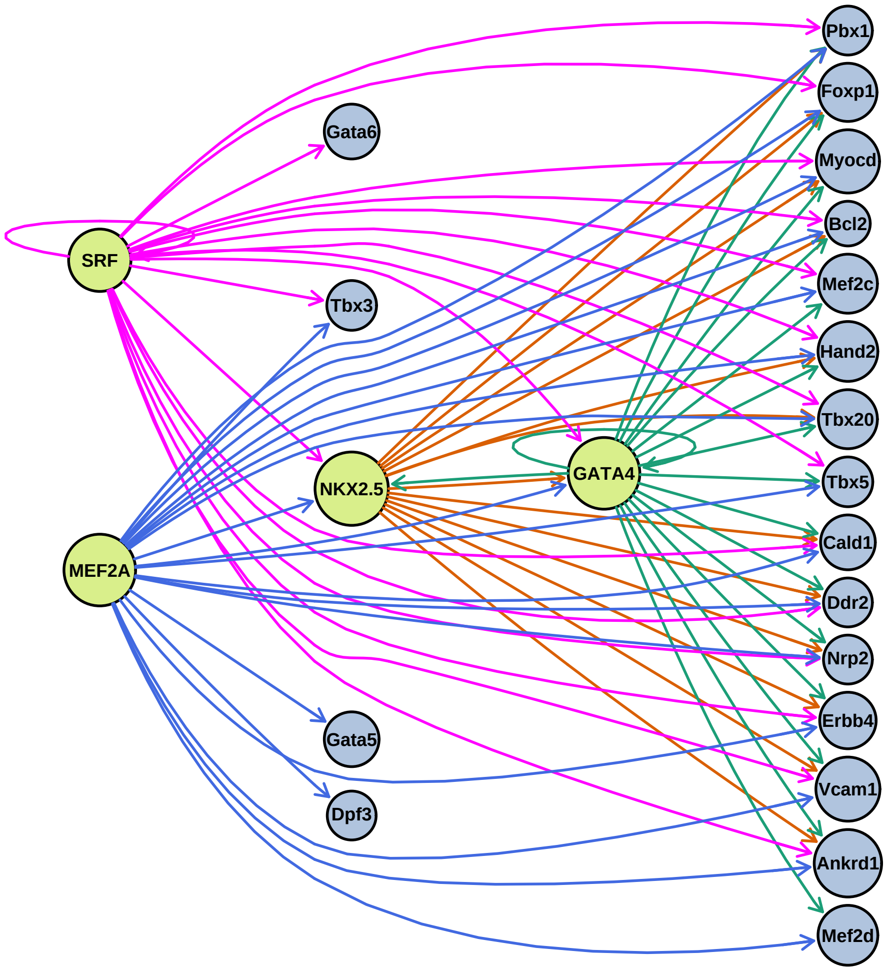


Figure 3.7: *Heart transcription factor network as seen in the ChIP-chip data. The green coloured nodes are the four transcription factors for which we sought binding sites with ChIP-chip. Blue nodes are a subset of the genes with identified TF binding events. An edge from a TF to a gene indicates that the TF had a ChIP-enriched region upstream or inside one of the respective gene's transcripts. Edge colouring is used to distinguish the edges from the four TFs.*

Figure 3.8 displays the reporter levels for the four transcription factors at the TSS of the *Hand2* gene on chromosome 8 (bottom panel). For comparison, the ChIP-chip data for the four histone modifications (Section 3.1) in this region is also shown (top panel).

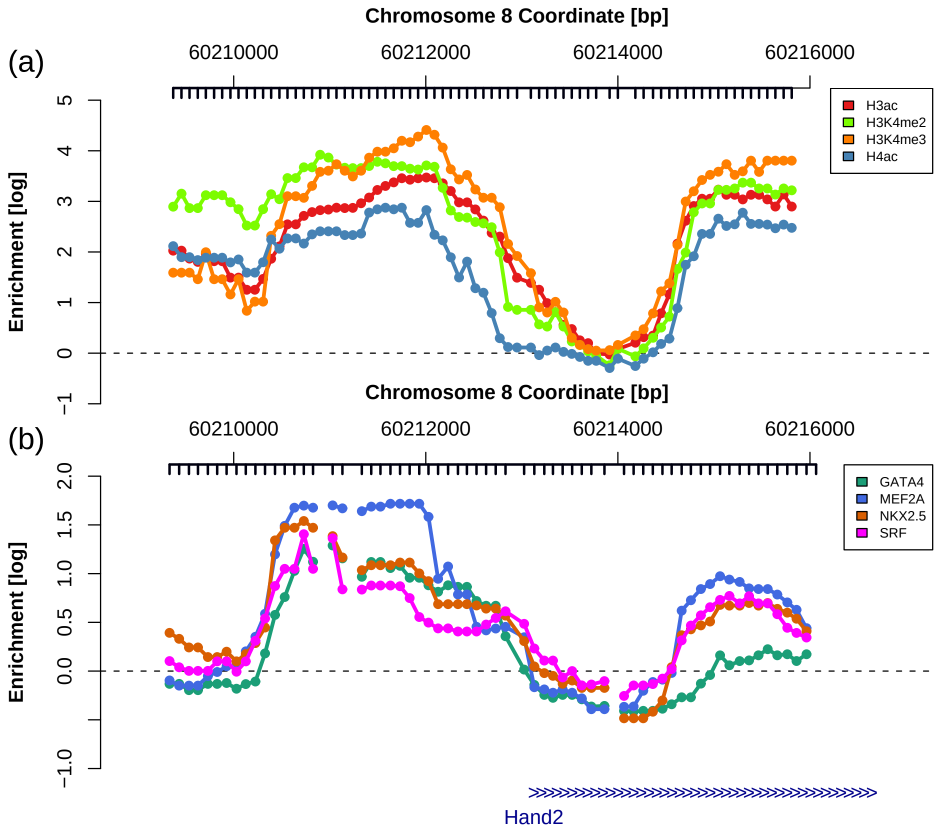


Figure 3.8: *Normalised histone modifications ChIP-chip and TF ChIP-chip reporter levels at the TSS of the gene* Hand2 *in HL1 cells.* (a) *(top panel): the four histone modifications* (b) *(bottom panel): the four transcription factors. The ticks below the genomic coordinate axis on top indicate genomic positions matched by reporters on the microarray. The blue arrows on the bottom mark the* Hand2 *gene with the arrow direction indicating its transcription direction, i.e. the gene is located on the Watson strand.*

3.2.4 Discussion

We analysed DNA-binding events of four transcription factors, GATA4, MEF2A, NKX2.5 and SRF, in HL1 cardiomyocyte cells, using ChIP-chip on custom oligonucleotide microarrays with reporters tiling potential regulatory regions surrounding the TSSs of selected genes. Figure 3.7 shows the core regulatory network of the four TFs, as it is suggested by our ChIP-chip data. The graph indicates a tight regulatory interplay between the transcription factors. NKX2.5 and GATA4 show ChIP-enriched regions in the upstream or transcribed genome region of each other, indicating that NKX2.5 potentially regulates the expression of *Gata4*, and vice versa. GATA4 and SRF show ChIP-enriched regions in their own transcript region, indicating potential auto-regulation of their own transcription rate. The genes of *Nkx2.5* and *Gata4* both show potential regulation by the transcription factors SRF and MEF2A, as do many of the indicated NKX2.5- and GATA4-target genes. This could indicate that SRF and MEF2A are at a higher hierarchical level in the regulatory network. The other displayed genes that show enrichment for all four TFs are all known to play roles in the development of heart cells. See Section 3.1.3 for the earlier discussion of *Hand2*. In addition to HAND2, the T-box transcription factors TBX3, TBX5 and TBX20 have also been implicated as components of the core regulatory network of heart development [100]. The apparent co-regulation of *Tbx3*, *Tbx5* and *Tbx20* by multiple transcription factors underlines their reported function (see Figure 3.6). *Tbx20*, which shows ChIP-enriched regions for all four analyses TFs, has been described as an activator in the formation of the chambers and the conduction system. *Tbx3*, which only shows borderline ChIP-enrichment for SRF and MEF2A (Figure 3.6), was implicated as a repressor of chamber myocardium development [100].

Not every identified ChIP-enriched region necessarily corresponds to an actual binding site of that transcription factor *in vivo*. Unspecific binding of the antibody and cross-hybridisation of the reporters in that region result in false positive enriched regions. Preliminary validation of the TF enrichment regions using quantitative PCR indicated a low false positive rate in the data, however. Moreover, multiple previously described regulatory interactions were also implied in our data.

A second caveat is that the presence of a genuine binding site of a transcription factor does not necessarily imply that the bound transcription factor affects the expression of any nearby genes [121].

Appropriate experiments are required to follow up on the observed ChIP-enriched regions for transcription factors and histone modifications. Experiments can validate hypothesised binding sites and help to distinguish between functional binding sites and inconsequential binding events or false positive identified ChIP-enriched regions. Moreover, investigations of the interaction between these transcription factors and the modified histones can provide further insights into the transcriptional regulatory network in heart cell development. Follow-up experiments may also clarify at which particular stage(s) of the transcription process these transcription factors and the histone modifications are involved.

Contributions

These studies were undertaken as part of the European Union Framework Project *HeartRepair* (FP6, LSHM-CT-2005-018630) and were a collaboration between our research group and the Cardiovascular Genetics group at the Max Planck Institute for Molecular Genetics in Berlin, Germany.

I planned and performed the data analyses described here. Jenny J. Fischer and Silke Sperling performed the microarray experiments and were in continuous communication with me about the progress of the study. Tammo Krueger and Matt Ritchie contributed R source code for data processing. Silke Sperling planned the experiments. Wolfgang Huber supervised my data analysis. I designed the oligonucleotide microarray for the transcription factor ChIP-chip study, together with Tammo Krueger, Silke Sperling and the support team of NimbleGen Systems.

Chapter 4

A high-resolution view of transcription in budding yeast

4.1 Introduction

We studied the transcriptome of *Saccharomyces cerevisiae* (budding yeast) in exponential growth phase and during cell cycle progression, using a whole-genome tiling microarray. The microarray contains over 6.5 million 25mer reporters, which cover the *S. cerevisiae* genome (strain S288c) in unprecedented detail and resolution. Figure 4.1 displays the layout of the strand-specific reporter tiling of the *S. cerevisiae* genome, using a microarray design which was devised by Lars Steinmetz in collaboration with Affymetrix Inc. (Santa Clara, California, USA). The reporters were designed to tile each strand, with a reporter starting every 8 base pairs. An offset of 4 bp between the two strands' tiling paths allows for strand-specific readouts. For each of the approximately 3 million *perfect match* (PM) reporters that match the genome without mismatches, there is a *mismatch* (MM) reporter. A MM reporter has the same sequence as the corresponding PM reporter, apart from its central base at the thirteenth position, which is replaced by its complementary base.

Prior to our analyses, we re-aligned the sequences of all reporters, regardless of whether the reporters had been designed as PM or MM reporters

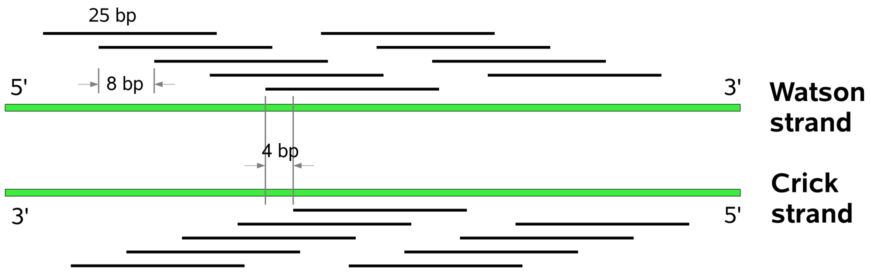


Figure 4.1: *Layout of reporters tiling the S. cerevisiae genome on the used Affymetrix arrays. On each strand, the offset between reporter start positions is 8 bp; between the two strands, there is a 4 bp offset between the two tiling paths, which allows for strand-specific readouts.*

by the manufacturer, to the genome sequences of *S. cerevisiae* strain S288C (obtained from SGD, `ftp://genome-ftp.stanford.edu/pub/yeast/data_download`, version of 7 August 2005). Only reporters with a perfect match to a 25-nucleotide segment in the genome sequence were used in the following analyses. For each genomic position matched by a reporter, it was recorded whether it was a unique reporter match position (RMP) or whether that reporter also matched other genomic positions[1].

4.2 Redefining the transcriptome

The first study was an analysis of the *S. cerevisiae* transcriptome in exponential growth phase. We used the microarray readouts to identify which parts of the yeast genome are being transcribed in cells in exponential growth phase and how the observed transcripts are related to annotated genome features.

[1] More precisely, a reporter matches a 25-nucleotide "segment". That segment, however, is uniquely defined by the genomic position of its central nucleotide, which is why I refer to the "position" that is matched by a reporter.

4.2.1 Material and methods

Samples were taken from the *S. cerevisiae* 288c background strain S96, which had been grown in rich media to exponential growth. Poly-A RNA[2] and genomic DNA were extracted from the samples and hybridised to tiling microarrays of the design described above (see Section 4.1).

During quality assessment, a number of the hybridisations showed obvious artifacts on the microarray surface that rendered them useless for later analyses. However, three poly-A RNA hybridisations and three genomic DNA hybridisations were deemed to be of acceptable quality. First, we normalised the data to increase the signal-to-noise ratio in the data.

Normalisation We used the three genomic DNA hybridisations for normalisation of the RNA hybridisations. See Section 5.1.3 for a detailed description of this normalisation approach, which we implemented in our custom R/Bioconductor package *tilingArray*. Briefly, from the reporter levels from the RNA hybridisation, we first subtracted a background level estimated from the intensities of reporters that mapped to the genome outside any annotated genome feature[3]. The background-corrected reporter levels from the RNA samples were then divided by a reporter-wise affinity level, which was estimated by the geometric mean of that reporter's intensities in the three genomic-DNA hybridisations. Finally, the reporters with the 5% lowest geometric mean intensities in the DNA hybridisations were dropped from the data. These were deemed to be dead reporters, the readouts of which would be dominated by noise. When taking the ratio during normalisation (Equation (5.5), page 122), such readouts would result in artificially high reporter levels, as a consequence of dividing two small numbers by each other.

The preprocessed levels are fold changes of expression in the RNA sam-

[2] We also prepared total RNA extract samples (Total-RNA) for the study [9]. The results described here, however, were obtained on the Poly-A RNA samples only, since their data were deemed to be cleaner than the Total-RNA data.

[3] Genome feature is used as a summary term for annotated ORFs (confirmed or dubious ones), upstream open reading frames (uORFs), pseudogenes, ncRNAs, repeat regions, and transposable elements.

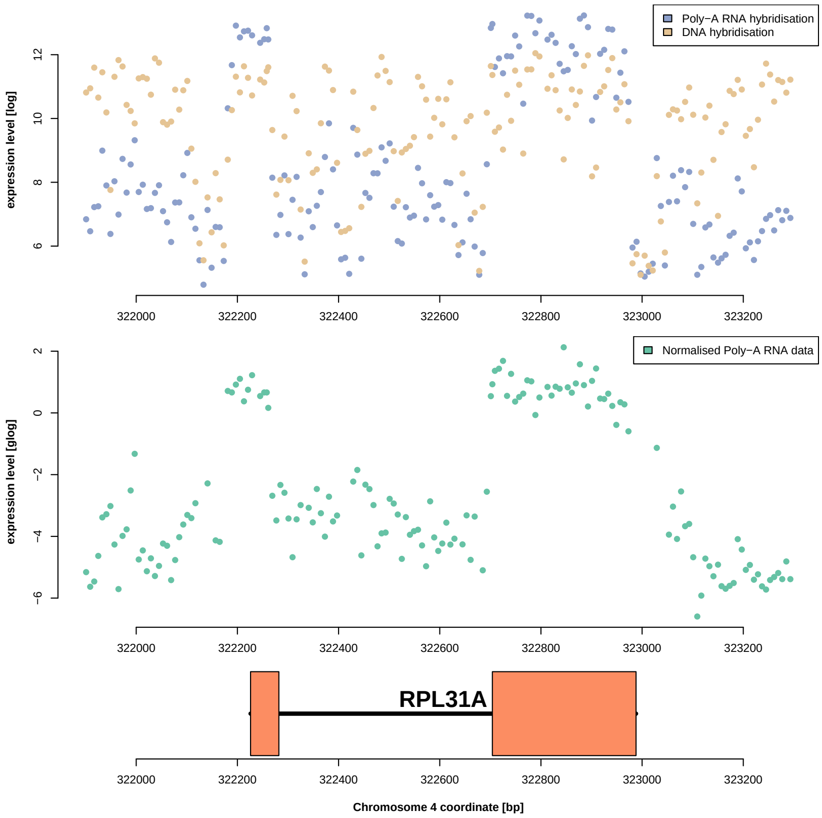


Figure 4.2: *Example picture displaying the effect of the normalisation method, tilingArray, which uses a DNA hybridisation to increase the signal-to-noise ratio for RNA hybridisations. Shown is a genomic region on the Watson strand of chromosome 4 of S. cerevisiae. The upper panel shows the raw Poly-A RNA and DNA readouts from reporters having a match position in this region. The middle panel shows the normalised Poly-A RNA signal. The lower panel shows the gene annotated in this region, with orange boxes indicating the exons of this gene.*

ples in comparison to the DNA samples. Thus, high reporter levels indicate that the genome segment matched by this reporter may be part of a transcript. I will refer to the preprocessed reporter levels as *expression levels*.

Figure 4.2 shows the effect of normalisation on the data in the region around the ORF of gene *RPL31A*. Reporter levels matching the two exons of the transcript are elevated in contrast to the reporter levels outside of the ORF or in the intron. The expression levels confirm that the two exons are part of the ORF's mRNA, while the intron is not. One can also see that a few reporter levels to the left and right of the ORF boundaries are also elevated. These correspond to the 5′ UTR and 3′ UTR of the ORF's transcript, respectively.

After normalisation, the reporter expression levels were interpreted as signal along each strand. Each strand was divided into segments of similar expression level. See Section 5.1.4 for details about the segmentation algorithm. Briefly, the signal along each strand is divided into S segments. Expression levels within each segment are assumed to be constant, and the segment boundaries are chosen using a dynamic programming algorithm that minimises the residual sum of squares (that is, the sum of squares of all expression levels minus their respective segment mean expression levels). The single parameter of the algorithm is S, the number of segments. We set it based on the assumption that the average length of a transcript or non-transcribed segment would be 1.5 kb[4]. Dividing the length of each chromosome by 1.5 kb gave the setting for the parameter S, the number of segments on each strand of that chromosome.

Segment expression level A segment k is characterised by two segment boundaries t_k and t_{k+1}, with $t_k < t_{k+1}$. The segment boundaries are nucleotide positions on a chromosome, as are the RMPs. Each reporter with a unique RMP z_i that satisfied $t_k \leq z_i < t_{k+1}$ was considered to be within the segment. The expression level of segment k was defined to be the median of the expression levels of all reporters within the segment k.

Threshold for transcribed segments We estimated an expression threshold, above which a segment is considered to be transcribed. To estimate the expression threshold, most unannotated regions were assumed to be untranscribed. The distribution of segment expression levels from unannotated regions was used to estimate the expression level distribution of

[4] The average exon length in *S. cerevisiae* is 1,305 bp; and 6,322 out of 6,609 genes contain only one single exon, as annotated in the Ensembl database [34, release 50].

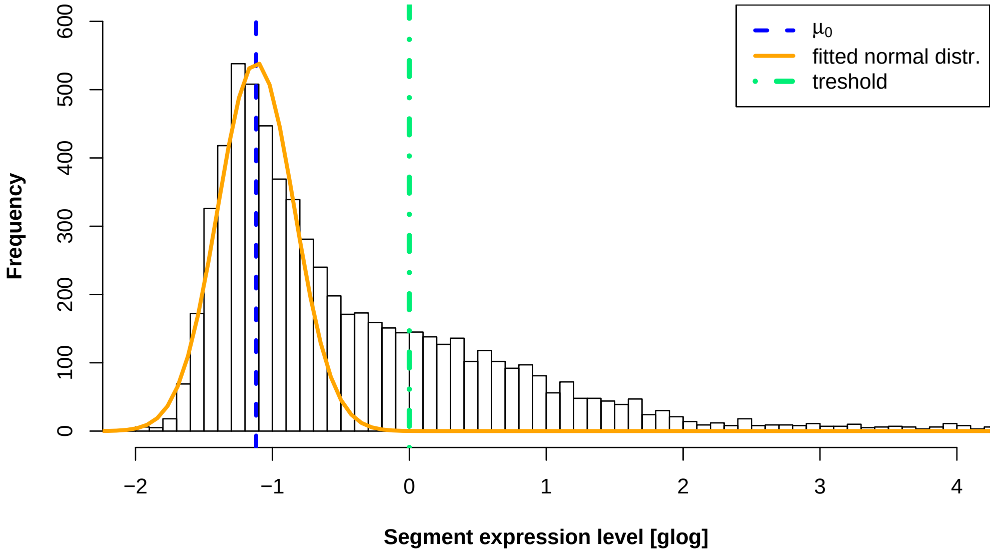


Figure 4.3: *Histogram displaying the median segment expression levels (after normalisation) for the segments that do not overlap any annotated genome features. The distribution of these segment levels was used to estimate the threshold y_0 (see the text for details). Segments with median segment levels exceeding y_0 were considered to be transcribed. Note that after the threshold had been determined, the segment levels were scaled such that the threshold is equal to* 0.

untranscribed regions. We considered the distribution of expression levels of those segments that did not overlap with any annotated genome feature, but for which the majority of RMPs were unique to the segment. Some of these segments may have corresponded to previously unannotated transcripts, but the majority of them were assumed to be untranscribed. The observed distribution of average segment levels y is shown in Figure 4.3. The distribution has a sharp peak on the left, corresponding to untranscribed segments, and a flat shoulder on the right, which presumably includes the unannotated intergenic transcripts. This distribution can be interpreted as a mixture between a null distribution (the peak) that corresponds to a normal distribution $\mathcal{L}_0 \simeq \mathcal{N}(\mu_0, \sigma_0)$ and some other distribution $\mathcal{L}_{\text{alt}}$. The alternative distribution $\mathcal{L}_{\text{alt}}$ was assumed to contain negligible mass at $y < \mu_0$.

The normal distribution $\mathcal{L}_0$ was estimated from the observed distribution of segment expression levels of unannotated regions, as follows. The mode of the distribution was estimated by the midpoint of the shorth[5] of the distribution This mode was taken as an estimate for $\widehat{\mu}_0$, the mean of the normal distribution $\mathcal{L}_0$. For an estimator of the standard deviation σ_0, first the values $y < \mu_0$ were reflected onto $y > \mu_0$. The median absolute deviation over these values, which are symmetrically distributed about μ_0, was taken as an estimate for $\widehat{\sigma}_0$. We assumed that the null distribution of expression levels in untranscribed segments is the (μ_0, σ_0) normal distribution and assigned a p-value to every segment; that is, how probable it is to observe a similar or greater average segment level under such a null distribution. The p-values were adjusted for multiple testing using the Benjamini-Yekutieli false discovery rate (FDR) procedure [122]. We selected the segment level threshold y_0 that corresponds to a FDR of 0.1% to get a conservative set of transcribed unannotated regions. Any segment, whether corresponding to an annotated genome feature or not, was considered to be transcribed if its median expression level y satisfied $y \geq y_0$.

A method that is similar to our estimation of the null distribution of expression levels in untranscribed segments has been described previously for estimating the distribution of microarray reporter intensities that are due to non-specific binding and background noise [24].

Categorisation The segments were categorised according to existing genome feature annotation[6]. Segments in which more than 50% of the reporters had other match positions (i.e. $>$ 50% of the RMPs were non-unique) were assigned to the category "excluded". The remaining segments were divided into categories "transcribed" and "untranscribed", depending on whether the segment expression level exceeded the threshold (described above).

[5] The *shorth* of a univariate distribution is defined as the shortest interval that contains at least half of the data. Its midpoint is a robust estimator of the mode of an unimodal distribution.

[6] Genome feature annotation was obtained from the *Saccharomyces* Genome Database (SGD) FTP site at `ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature` in August 2005.

For transcribed segments, first the overlap on the same strand with any verified, uncharacterised or dubious ORF, any transposable element, or any rRNA, tRNA, snRNA or snoRNA was assessed. If there was any overlap with any such feature, the segment was recorded as overlapping the feature a) completely, b) more than 50%, or c) less than 50% of the segment length. In the case of a segment overlapping more than one type of feature, the category was determined by the "most important" feature type in that segment, where features were ranked by importance as follows: verified ORF > uncharacterised ORF > dubious ORF > rRNA > tRNA > snRNA > snoRNA > transposable element. For transcribed segments that did not overlap any such feature on the same strand, we checked the overlap to features on the opposite strand. Any transcribed segment that did not overlap any such feature on the same strand, but did overlap a feature on the other strand, was categorised as "antisense". Any transcribed segment that did not overlap any feature on either strand was categorised as "unannotated intergenic". As additional annotation, annotated transcription factor binding sites, which were derived in ChIP-chip experiments by Harbison and co-workers [123], were obtained from `http://jura.wi.mit.edu/fraenkel/download/release_v24/GFF`. These transcription factor binding sites (TFBSs) were not considered for categorisation of transcribed segments, but were added to the along-chromosome visualisations in the database (page 88) as additional annotation of genome regions.

Pruning of short segments Short, apparently transcribed, segments could have arisen during the reverse transcription step (from RNA to cDNA), by spurious second-strand transcription from the first-strand cDNA. We applied computational filters to distinguish "real" transcribed segments from such artefactual transcript. Transcribed segments that did not overlap annotated genome feature were only considered for further analyses if they were longer than 48 bp, had higher expression levels than the adjacent segments at both sides, and if they had a higher expression level than the region defined by the same boundaries on the opposite strand.

UTR lengths For each transcribed segment that completely contained a single annotated ORF, the 5′ and 3′ UTR lengths were determined by comparing the positions of the segment boundaries with the boundaries of the ORF. The 5′ UTR length was calculated as the distance between the first base of the first reporter and the first base of the ORF start codon. The 3′ UTR length was calculated as the distance between the last base of the stop codon and the last base of the last reporter in the segment. 5′ UTR and 3′ UTR lengths were grouped by Gene Ontology annotation (Section 1.6) of the ORFs. For each GO group, I compared the UTR lengths of ORF transcripts in the group with the UTR lengths of all ORF transcripts not included in group, using a two-sample, two-tailed Wilcoxon rank sum test. ORF transcripts in each GO group were considered to have significantly long or significantly short UTR lengths if the group's p-value satisfied $p \leq 0.002$.

Software The R software which was written for the analysis, and which I co-authored, is available in package *tilingArray* (Section 5.1) from the Bioconductor project (`www.bioconductor.org`), as is the microarray data in package *davidTiling*.

4.2.2 Results

Our microarray data describe the transcriptome of *S. cerevisiae* in unprecedented detail. 11,412,977 bp ($\approx$ 94%) of the genome of strain S288c are represented uniquely by reporters on the microarray that we used. Of these, 85% showed an expression level above background (where the background level was determined as described above on page 81).

Most transcribed segments corresponded to annotated ORFs. Table 4.1 shows the categories and lengths of observed transcripts.

2,223 transcribed segments contained one single ORF and had a significantly higher expression level than the adjacent segments at both sides. For these ORF transcripts, we could accurately measure the lengths of their UTRs by comparing the transcript segment boundaries with the annotated ORF boundaries. 3′ UTRs had a median length of 91 nu-

Transcribed segment category	#	Median length [bp] (Q_1 – Q_3)
annotated ORF	5,942	1,321 (769 – 2,039)
dubious ORF	183	1,073 (571 – 1,789)
ncRNA (all)	80	413 (249 – 1,301)
unannotated intergenic	214	305 (91–637)
antisense	228	1,169 (739–1,797)

Table 4.1: *Categories and lengths of 6,647 transcribed segments in the yeast tiling microarray data. Samples were taken from strain S288c in exponential growth phase. The lower and upper quartiles of the empirical lengths distributions are specified in brackets beside the median lengths. "ncRNA (all)" is used as a grouping term for rRNA, tRNA, snRNA and snoRNA. #: "number of".*

cleotides and were found to be longer than 5′ UTRs, which had a median length of 68 nucleotides. The UTR lengths were found to differ between ORFs of distinct functional categories. I grouped the UTR lengths by the genes' Gene Ontology annotations (see Section 1.6) and compared the UTR lengths between each group and the combined remaining groups. The results are shown in Figure 4.4.

Furthermore, we saw multiple transcripts of unexpected structure, such as pairs of genes being transcribed as single, apparently *bicistronic*, transcripts. See Figure 5.2 (Chapter 5, page 130) for an example of such an operon-like transcript.

In accordance with other recent publications from tiling microarray experiments, the proportion of the genome that is actively transcribed was found to be much larger than currently annotated. We found about 200 unannotated intergenic transcripts from DNA sections that do not have any existing genomic feature annotation on either strand. None of these transcripts (which were validated using PCR) were indicated to be protein-coding in an evolutionary analysis (see [124, page 121ff.] for details). With a median length of 305 bp, these unannotated intergenic transcripts are shorter than ORF transcripts and previously annotated ncRNAs (median length: 413 bp). Figure 4.5 shows one such unannotated intergenic transcript on chromosome 16. In addition, there were about 200

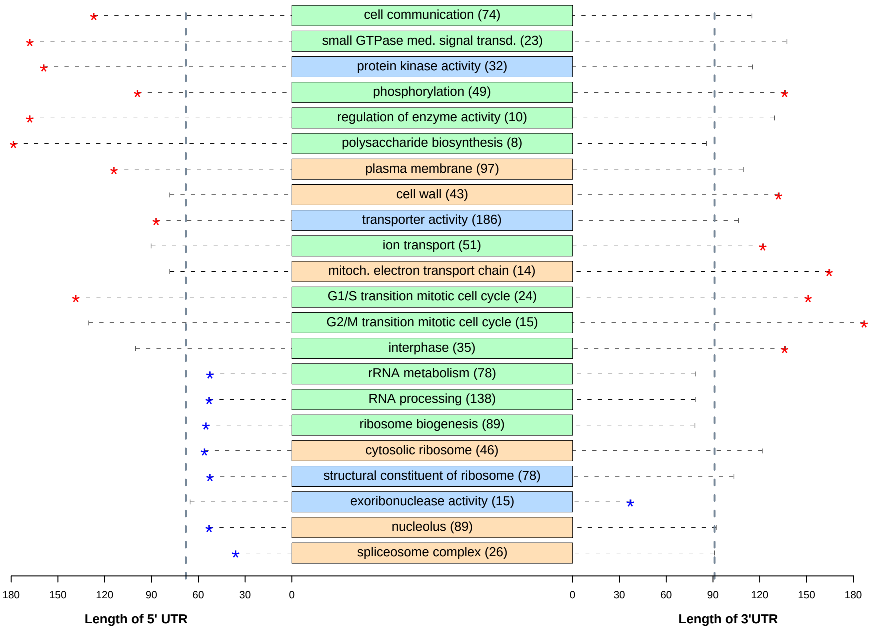


Figure 4.4: *Shown are the median 5′ and 3′ UTR lengths for ORF transcripts annotated to each GO category shown. The colour of the box marks which ontology a category belongs to (orange: cellular component, green: biological process, blue: molecular function). The grey vertical lines denote the median 5′ and 3′ UTR lengths over all ORF transcripts. Stars indicate categories for which the UTR lengths are significantly longer (red star) or shorter (blue star) than the UTR lengths of transcripts not annotated to the category (two-sided Wilcoxon test, $p \leq 0.002$). med.: mediated; mitoch.: mitochondrial.*

transcripts antisense to known genes. Figure 4.5 display an identified transcript that is antisense to the ORF *YPR027C*.

In summary, our data suggest that transcriptomes, even from organisms that have been extensively studied, are far more complex than currently annotated. All biological findings from this study are presented in detail in the related publication [9]. Further details about the normalisation and segmentation methods used in this study can be found in Chapter 5 (Section 5.1).

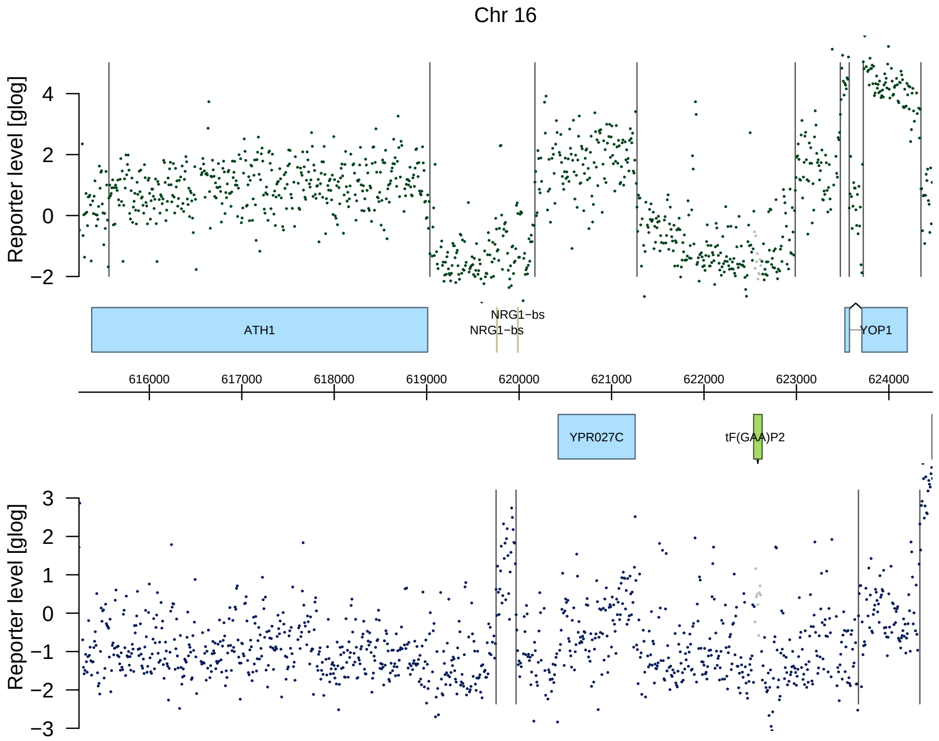


Figure 4.5: *Visualisation of the Poly-A-RNA data in a 9 kb region of chromosome 16. The green dots in the top panel correspond to the preprocessed reporter intensities on the Watson strand, the blue dots in the bottom panel are the intensities from the Crick strand. The middle panel shows the genomic coordinate and genome features that are annotated in this region. Blue boxes are annotated ORFs; the golden vertical bars are experimentally determined transcription factor binding sites. The bright-green box denotes the genomic template of a tRNA. On the Crick strand, a unannotated intergenic transcript of 233 bp length can be seen. Moreover, there is a transcript antisense to the ORF YPR027C.*

Database

Visualisations of our transcription data in all genomic regions (similar to Figure 4.5) are provided in an online database. The database can be searched by gene name, gene alias and chromosomal coordinate. I implemented the database and its interface in `HTML` and `Perl CGI`. The database can be accessed at `http://www.ebi.ac.uk/huber-srv/David2006`.

4.2.3 Discussion

We have presented a high-resolution survey of the transcriptome of *S. cerevisiae* in unprecedented detail, available in an online database. Whole-genome tiling microarrays, despite their technical shortcomings and analysis caveats, provide an unbiased view of the transcriptome, since the reported transcribed regions are not restricted to previously annotated genome features. Using this technology, we have seen that even the transcriptome of a thoroughly studied organism, *S. cerevisiae*, still contains many unexpected regions of transcriptional activity.

Many genomic regions that are considered to be transcribed corresponded to annotated open reading frames. The segmented tiling microarray data allowed us to obtain precise measurements of the lengths of the UTRs. I investigated the relationship between UTR length and gene function as annotated in the Gene Ontology (Figure 4.4). In general, 3′ UTRs were found to be longer than 5′ UTRs, which fits well with reports of the diverse roles of the 3′ UTR. For example, the 3′ UTR was reported to contain sequence motifs for mRNA localisation [125] and motifs for regulation of mRNA stability [126]. The observed relationship between gene function and UTR length also indicates that long UTRs are required for more complex functions outside of the nucleus. The mRNAs that are exported to the mitochondria (GO:0005746) and the cell wall (GO:0005618) show longer 3′ UTRs, which contain the localisation sequence motifs.

Not all ORFs annotated in the *S. cerevisiae* genome were found to be transcribed in our data. Genes that are required for sporulation and mating, for example, did not show any transcripts. This observation was to be expected, considering that the data were derived from *S. cerevisiae* samples grown in rich medium to exponential phase.

We observed more than 200 unannotated intergenic transcripts outside any annotated genome feature and more than 200 antisense transcripts. The function of these unannotated transcripts is unknown, and they do not appear to be protein-coding [124, page 121ff.]. Several recent high-throughput transcriptome studies have reported large numbers of unexpected transcripts of unknown function in different organisms [127]. Among such transcripts of unknown function are antisense transcripts, as seen in our survey of the *S. cerevisiae* transcriptome. *S. cerevisiae* lacks core components of the RNA interference (RNAi) pathway[7] [128]; therefore a possible role of the observed antisense transcripts in RNAi degradation of the complementary sense mRNAs can be excluded.

A small fraction of our observed antisense transcripts may have been due to antisense artifacts. A recent study by Perocchi *et al.* has shown that, during the reverse transcription step from the RNA sample to cDNA (which is then hybridised to the microarray), second-strand artifacts are created by accidental reverse transcription of the first cDNA copy [129]. When hybridising the cDNA to a microarray that allows for strand-specific readouts, such as the microarray that we used, the artifacts will be observed on the strand opposite to the actual measured transcript. Perocchi *et al.* suggested the use of the polypeptide actinomycin D, which specifically blocks DNA-dependent DNA replication, to prevent such antisense artifacts. Our computational pruning of short transcripts (page 84), however, has likely eliminated most (but not all) of the second-strand artifacts [129].

In summary, our tiling microarray study provided important insights into the transcriptional complexity of budding yeast. In a recent, independent high-throughput sequencing study of the *S. cerevisiae* transcriptome, Nagalakshmi *et al.* found 85% of the DNA that we reported as transcribed to be transcribed in their data [10]. The high agreement between the two studies' unexpected finding, that more than three quarters of the *S. cerevisiae* DNA are actively transcribed during exponential growth, indicates how little the characteristics of to-be-transcribed DNA features are yet understood, even in genomes of well-studied model organisms.

[7] The absence of components of the RNA interference pathway in *S. cerevisiae* is remarkable, since they are conserved from *Schizosaccharomyces pombe*, another species of yeast, to multicellular eukaryotes [128].

Contributions

The experimental study, data preprocessing and segmentation were planned and performed by Lior David, Wolfgang Huber and Lars Steinmetz. I contributed to the implementation of the accompanying R/Bioconductor package *tilingArray* for data preprocessing and analysis (Chapter 5; Section 5.1), in particular the implementation of the segmentation algorithm. I performed the analysis and visualisation of the results of differing UTR lengths per GO category. I also programmed the web interface that provides easy access to the high-resolution visualisation of our data. The protein-coding potential of the unannotated intergenic transcripts was assessed by Lee Bofkin.

4.3 Cell-cycle data

We worked with two further data sets, in which the cells had been synchronised to the same stage of the cell cycle using two different methods (see Section 1.5, page 14ff., for a description of the *S. cerevisiae* cell cycle and previous related microarray studies). In one data set, the α-factor pheromone was used to achieve the synchronisation, in the second set a temperature-sensitive mutant of the the cyclin-dependent kinase Cdc28 (denoted $cdc28_m$) was employed. Both synchronisation methods are well established [48, 49]. Following the release of cell-cycle arrest, one sample was taken every five minutes, up to 220 (α-factor) or 215 ($cdc28_m$) minutes after the release. The samples were treated with actinomycin D to prevent the emergence of second-strand antisense transcripts during the reverse transcription step [129].

The microarray data were normalised as described above (Section 4.2), using three genomic DNA hybridisations that had also been treated with actinomycin D. The normalised data were also segmented as described above (Section 4.2), except for one change. For setting the segmentation algorithm's parameter S, the number of segments (see Section 5.1.4 for details), we assumed a shorter average segment length of 1,250 bp (instead of 1,500 bp). This shorter average segment length was more consistent with the average length of transcripts, as the median length of transcripts observed in the exponential growth phase data was 1,273 bp (Section 4.2). The parameter S for segmenting each chromosome strand was again determined by diving the length of the chromosome by the average segment length. Hence, we obtained a higher number of segments per strand than in the previous study. Both data sets were segmented together, such that the same segment boundaries applied in both data sets.

The transcribed segments were also identified and categorised according to overlap with existing genome annotation as described above[8] (Section 4.2), with one small change in the procedure. In the pruning step of the previous study (page 84), one condition for defining an unannotated intergenic transcribed segment was that the median reporter level in the segment had to be higher than the segment level on the opposite strand. This filter had been introduced as a way to handle the previously mentioned antisense-transcription artifacts. Due to the actinomycin D treatment of the samples, such artefactual reverse transcription was suppressed, and the filter could be dispensed with.

4.3.1 Methods

Determine periodicity

A combination of two methods was used to determine periodically expressed transcripts. The first method, by Ahdesmäki *et al.* [130], is a robust testing procedure for periodicity of expression, based on a spectral estimator together with Fisher's g-statistic and correction for multiple testing. The procedure is available in the R-package *GeneCycle*. Transcripts with a p-value ≤ 0.01 were considered to be periodically expressed. The second method used was suggested by de Lichtenberg *et al.* [51] and consists

[8] I obtained an updated annotation of the *S. cerevisiae* genome from the SGD web site in September 2007.

of computing a combined p-value p_{total} for cell-cycle regulation of a transcript. For each transcript, p_{total}-value is combined from $p_{\text{periodicity}}$, a p-value for periodic expression, and $p_{\text{regulation}}$, a p-value for "regulation" of the transcript. $p_{\text{periodicity}}$ assesses the periodicity in the expression profile of the transcript (regardless of the amplitude), while $p_{\text{regulation}}$ accounts for the variance in the expression profile of the transcript.

Transcripts were considered to be periodically expressed if both methods, by Ahdesmäki and by de Lichtenberg, indicated such. Transcripts were also considered to be periodically expressed, if only one of the methods indicated such and the periodic expression could be confirmed upon inspection of the visualisation of the transcript expression profiles.

Regularised correlation

I define a regularised correlation coefficient as:

$$CC_{reg}(X_{i,\bullet},\, X_{j,\bullet}) \;=\; \frac{\operatorname{cov}(X_{i,\bullet},\, X_{j,\bullet})}{\operatorname{sd}(X_{i,\bullet}) \cdot \operatorname{sd}(X_{j,\bullet}) + s_0} \tag{4.1}$$

where:

$X_{i,\bullet}$: expression profile of transcript i in the considered samples (time points)

$\operatorname{cov}(X_{i,\bullet},\, X_{j,\bullet})$: covariance of the expression profiles of transcripts i and j

$\operatorname{sd}(X_{i,\bullet})$: standard deviation of the expression profile of transcript i

s_0: a regularisation constant, set equal to a certain quantile, such as the 10% quantile or the median, of the empirical distribution of all pair-wise products $\operatorname{sd}(X_{i,\bullet}) \cdot \operatorname{sd}(X_{j,\bullet})$. This constant is used to assign a lower correlation to transcript pairs whose expression shows little variation over time. The use of a similar regularisation constant has been suggested for regularising t-statistics when assessing differential gene expression [131].

Figure 4.6 shows the expression profiles of two example gene pairs in the α-factor cell-cycle data to clarify the need for regularising the correlation coefficient. The genes in the left panel (*YAR008W* and *YJL187C*) show periodic, variable expression. The genes in the right panel (*YOR033C* and

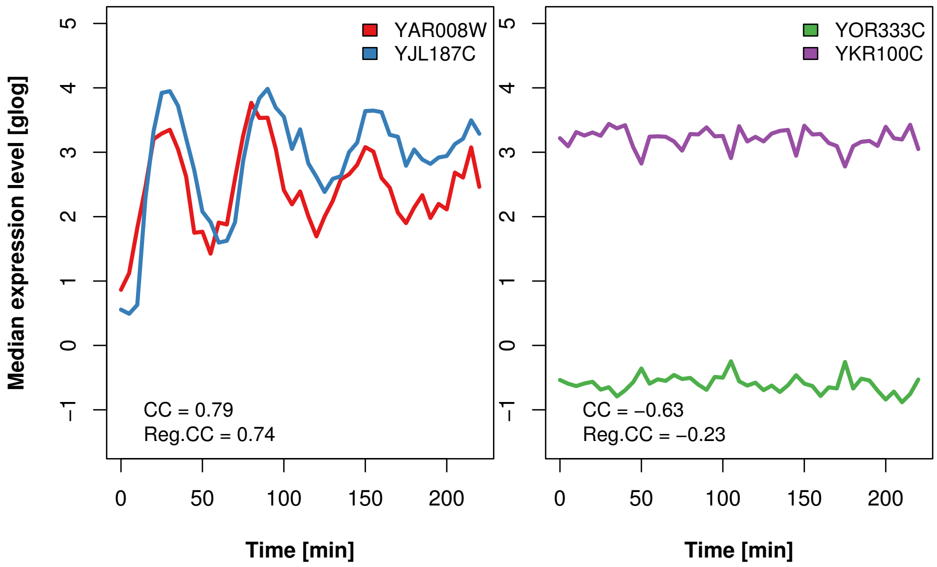


Figure 4.6: *This figure shows the expression pattern of two pairs of gene transcripts in the α-factor cell-cycle data. Also shown are the Pearson correlation coefficient (CC) of the gene expression profiles and the regularised CC_{reg} (Equation* (4.1)*). The y-axis captures the whole dynamic range of the α-factor expression data. reg: regularised; glog: generalised logarithmic scale.*

YKR100C), on the other hand, show little variation around the mean in their expression levels. This little variation happens to be in the opposite direction between the two genes, and the expression profiles thus have a Pearson correlation coefficient of -0.63. The small variation is accounted for in that the regularised correlation coefficient of this pair is considerably closer to zero. The genes in the left panel clearly show correlated expression profiles, while the negative correlation of the right pair is difficult to distinguish from noise in the expression profiles.

Regularised correlation distance The regularised correlation distance between two transcripts i and j is computed as:

$$d(i,j) \;=\; 1 - CC_{reg}(X_{i,\bullet},\; X_{j,\bullet})\;. \tag{4.2}$$

where $CC_{reg}(X_{i,\bullet},\; X_{j,\bullet})$ is defined above in equation (4.1).

Clustering

Periodically expressed transcripts were grouped into clusters using hierarchical clustering. The distance matrix between transcripts was computed using the regularised correlation distance (Equation (4.2)). This distance matrix was used as input for agglomerative, hierarchical clustering with the complete linkage agglomeration method. The height at which to cut the dendrogram was chosen by inspection, giving the number of clusters and the assignment of transcripts to clusters. The expression data along with the clustering are visualised in the form of a heatmap (Figure 4.10, Results page 104).

Transcription factor binding sites

The binding specificities of 124 *S. cerevisiae* transcription factors as position specific score matrices (PSSMs) were published by MacIsaac *et al.* [132], based on a refinement of ChIP-chip results by Harbison *et al.* [123]. I obtained these PSSMs from the supplementary web page of the publication [132] and added the later reported PSSM of the transcription factor Hcm1 [45].

The PSSM of a transcription factor is a probabilistic encoding of the TF's binding motif. The number of columns in the PSSM corresponds to the length of the motif in nucleotides, and the PSSM has four rows, one each for A, C, G, and T. Each entry $m_{x,j}$ of a PSSM M with L columns is the probability of observing nucleotide x at position j of the bound sequence, where $x \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and $j \in \{1, 2, \ldots, L\}$. Given a DNA sequence $Z = \{z_1, z_2, \ldots, z_L\}$ of the same length L, one can compute a score of how well the sequence matches the motif encoded by the PSSM M as

$$S_M(Z) = \sum_{i=1}^{L} \log \frac{m_{z_i,i}}{q_{z_i}} \tag{4.3}$$

where q_x is the background probability of observing nucleotide x in the genome. The score $S_M(Z)$ indicates how likely the TF whose binding motif is encoded by M would bind to sequence Z. Typically, a sequence is considered to match a motif if it achieves at least 70%–90% of the maximum possible score for the PSSM [133]. I required a score exceeding 80% of the maximum possible score for a match. At this threshold, a limited number of matches are found for each motif, and I tested how this number of matches changes over different sets of sequences. Note that control of the type I error rate for identifying matches in individual sets of sequences is not the concern here, but rather I compare the number of matches of each PSSM for different sets of sequences using the same threshold.

To assess whether a set of n DNA sequences matches a motif more often than expected by chance considering a total of m sequences ($m \geq n$), i.e. whether a certain binding motif is overrepresented in the smaller sequence set compared to the background frequency in the larger set of m sequences, I use the group-specificity score [134], which is based on the hypergeometric distribution. A set of n sequences, out of a total of m sequences, of which k sequences ($k \leq n$) match a certain motif, is assigned the following score, which is in essence a p-value, to the motif:

$$p(k) = \mathrm{P}\left(X \geq k\right) = 1 - \mathrm{P}\left(X < k\right) = 1 - \sum_{i=0}^{k-1} \mathrm{P}\left(X = i\right) \tag{4.4}$$

where

$$\mathrm{P}\left(X = i\right) = f(i;\ m, b, n) = \frac{\binom{b}{i}\binom{m-b}{n-i}}{\binom{m}{n}} \tag{4.5}$$

and

m: total number of sequences
n: number of selected sequences, with $n \leq m$
b: total number of sequences that match the binding motif
k, i: number of selected sequences that match the binding motif, with $k, i \leq n$.

The score $p(k)$ is the probability of observing a number of k or more sequences with matches to the motif, if the set of n selected sequences is a randomly drawn subset of the total set of all *m* sequences. A sequence set was deemed to have a significant enrichment of a transcription factor's binding motif if the score for this tuple was $p \leq 10^{-3}$.

Gene Ontology enrichment

I retrieved the GO annotation (Section 1.6) of verified *S. cerevisiae* genes from the Ensembl database (version 47), using the Bioconductor package *biomaRt* [72]. Over-representation of GO terms among gene groups was assessed using Fisher's exact test, as implemented in the Bioconductor package *topGO* [135], with the *p*-value cutoff set to 0.001. The evaluation proceeds in a bottom-up approach, starting from the most specific GO nodes. Genes that are used for evaluating a certain node are not used for evaluating any of its ancestor nodes [135, *elim* algorithm].

Software

The analyses described here, except for the TF motif enrichment, were implemented in the statistical environment R, making use of *tilingArray* (see Section 5.1) and other Bioconductor packages [104]. For the clustering and the subsequent heatmap visualisation, I used the base R functions `hclust` and `heatmap`. The enrichment for transcription factor binding motifs was investigated using the `Python` module *TAMO*, which was developed by Fraenkel and co-workers [136].

4.3.2 Results

We analysed two time-course tiling microarray data sets, that both covered more than two iterations of the cell cycle. The data sets were normalised independently of each other, but then combined for determining segments and transcripts. Table 4.2 lists the categories of transcribed segments that we observed in the cell-cycle data. Note that the number of transcribed ORF segments is greater than the number of annotated ORFs in the yeast genome (6,259 [Source: *SGD* www.yeastgenome.org, September 2007]). This can be explained by the observation that certain ORF transcripts are divided into two or more transcribed segments with different expression levels, or are interspersed with untranscribed intronic regions.

Transcribed segment category	**#**	**Median length [bp] (Q_1 – Q_3)**
ORF	7,576	937 (457 – 1,545)
Dubious ORF	194	701 (305 – 1,221)
antisense	523	721 (377 – 1,133)
unannotated intergenic	135	505 (345 – 877)
pseudogene	7	1,217 (425 – 1,741)
rRNA	13	217 (161 – 466)
snoRNA	60	287 (167 – 411)
snRNA	4	341 (153 – 643)
transposable element	8	361 (206 – 581)
tRNA	20	675 (299 – 991)

Table 4.2: *Categories of transcribed segments in the budding yeast cell-cycle data. The lower and upper quartiles of the empirical length distributions are specified in brackets behind the median lengths. #: "Number of segments".*

A number of transcript expression profiles show periodic patterns concordant with cell cycle progression. Figure 4.7 displays the scaled expression profiles of the transcripts of nine histone proteins in the two data sets. Histone transcripts have been previously reported to show varying expression patterns concordant with the cell-cycle progression, and to have their peak expression during the S phase [137, 48]. In our data sets, the histone transcripts are clearly periodically expressed. From the periodic expression pattern, it can be concluded that both data sets cover more than two iterations of the cell cycle. The periodic expression profiles of the histone transcripts (Figure 4.7) can be used to obtain an estimate of the duration of one iteration of the cell cycle in each data set. The histone transcripts show two clear expression peaks in both data sets and their expression is tightly synchronised up to 160 minutes after release from cell-cycle arrest. The synchronisation of cells wears off after 160 minutes. In the α-factor set, the histone transcripts show peak expression at 40 and at 105 minutes, indicat-

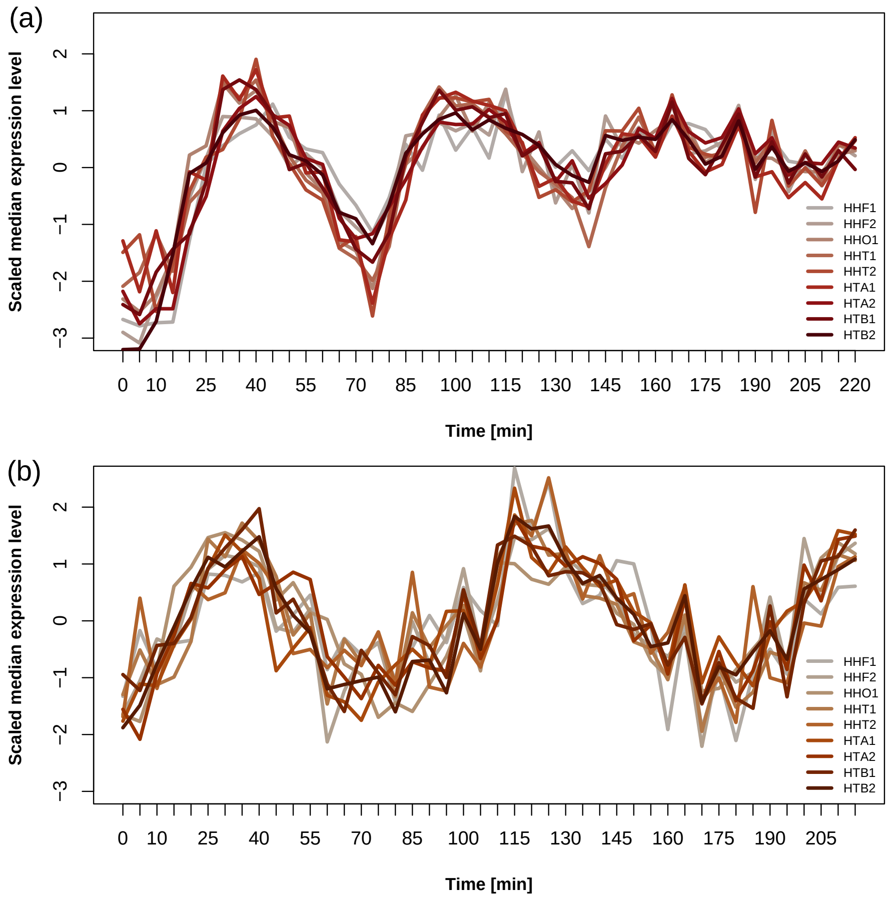


Figure 4.7: *These plots show the median expression of nine histone gene transcripts in the cell-cycle data. The top panel* (a) *shows their expression in the α-factor data, and the bottom panel* (b) *shows the expression in the cdc28$_m$ set. The expression profiles of the histone transcripts were standardised for visualisation purposes.*

ing that in these data one iteration of the cell cycle takes about 65 minutes, which corresponds well with the estimate of Spellman *et al.* (66 ± 11), who used the same synchronisation method for their data [48]. In the *cdc28$_m$* set, the transcripts show peak expression at 30 and at 120 minutes. Thus, one iteration of the cell cycle in the *cdc28$_m$* data takes about 90 minutes (the estimate of Spellman *et al.* for the *cdc28$_m$* data set of Cho *et al.* [49] was 80–100 minutes).

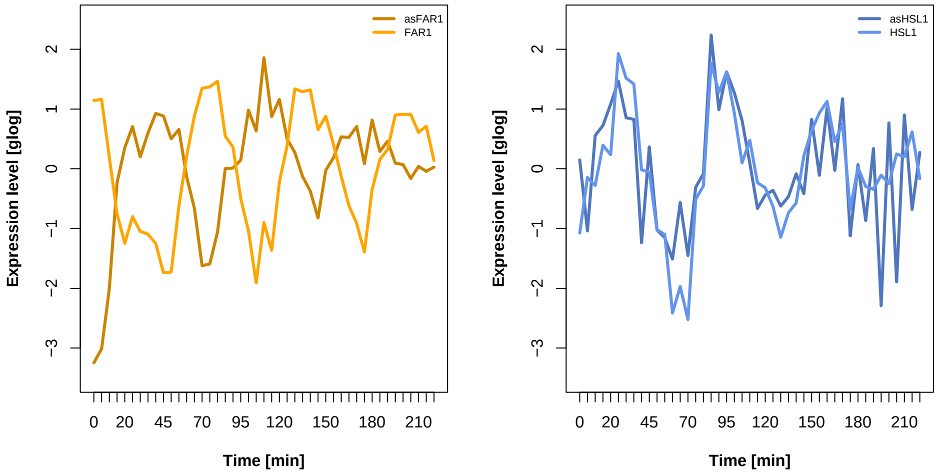


Figure 4.8: *Plots showing the expression profiles of two of the five antisense transcripts whose expression strongly correlates negatively ($CC_{reg} \leq -0.5$) or positively ($CC_{reg} \geq 0.5$) with the expression of the transcript on the opposite strand. Also shown are the expression profiles of the respective sense transcripts. The left panel shows the sense/antisense pair of FAR1 as an example for the four pairs with strong negative correlated expression profiles. The right panel shows the single sense/antisense pair (of HSL1) with strongly positively correlated expression profiles. as: "antisense".*

Periodically expressed transcripts

We determined a set of 639 transcripts that are periodically expressed in concordance with the cell cycle progression (as described in Section 4.3.1, page 92). These transcripts included 591 ORF transcripts, 37 antisense transcripts and 11 unannotated intergenic transcripts.

Periodic antisense transcripts

Among the periodically expressed transcripts, there were 37 that were antisense to annotated ORFs. Table 4.3 details these periodic antisense transcripts.

The expression profiles of five of the periodic antisense transcripts are strongly correlated with the expression profiles of the respective sense transcript ($|CC_{reg}| \geq 0.5$), 4 negatively and 1 positively. Figure 4.8 shows

Chromosome	**Strand**	**Start**	**End**	**Antisense to**	CC_{reg}	
1	-	189427	191819	*YAT1*	0.45	
2	+	372119	372543	*TIP1*	−0.49	
2	-	590100	591348	*DTR1*	−0.05	
3	+	202729	204209	*TAF2*	−0.45	
4	-	232869	233909	*VCX1*	−0.01	
4	+	722298	723026	*YCF1*	−0.04	
4	-	908055	909407	*GTB1*	0.02	
5	-	76101	77413	*YEF1*	−0.26	
7	-	454935	455087	*ALK1*	0.41	
8	+	378137	380073	*SPS100*	−0.18	
8	-	383677	384573	*CHS7*	−0.04	
8	-	456343	458487	*YHR177W*	0.01	
9	+	31185	32561	*YIL166C*	−0.07	
9	+	243633	243929	*YRB2*	0.03	
10	+	123737	125449	*FAR1*	−0.55	⋆
10	+	291913	293537	*PRY3*	−0.10	
10	-	703349	705149	*HMS2*	0.04	
11	-	6205	7749	*MCH2*	−0.15	
11	-	203832	204600	*PGM1*	−0.12	
11	-	251040	251384	*HSL1*	0.59	⋆
12	-	203261	204717	*YLR030W, YLR031W*	−0.41	
12	+	245577	246177	*YLR050C*	0.31	
12	-	807211	808699	*SPO77*	−0.04	
13	+	94457	95585	*YML087C*	−0.46	
13	-	618389	619429	*YMR178W*	−0.24	
13	+	623585	625281	*YMR181C*	−0.69	⋆
13	+	775985	776769	*YMR253C*	−0.35	
14	-	65757	66037	*YNL300W*	−0.13	
14	+	372473	374273	*YNL134C*	−0.38	
15	-	380333	380469	*BUB3*	0.31	
15	+	506945	507817	*YOR097C*	0.16	
16	+	76705	78609	*GYP5*	−0.24	
16	-	114485	116357	*USV1*	0.14	
16	+	243193	244217	*YPL162C*	−0.55	⋆
16	-	798015	799023	*MSS18*	−0.40	
16	-	799783	800391	*CTF4*	−0.61	⋆
16	+	924291	928339	*OPT2*	0.10	

Table 4.3: *This table provides details on the 37 antisense transcripts that are considered to be periodically expressed. CC_{reg}: regularised correlation coefficient of the antisense transcript's expression profile with the opposite sense transcript's expression profile. The five sense/antisense pairs whose expression profiles are strongly correlated (see text) are marked by ⋆.*

the expression levels of two of these five transcripts in the α-factor cell-cycle data. In the left panel, one of four antisense transcripts whose expression profiles have a strong negative (regularised) correlation with the expression profile of the respective sense transcript is shown. The genes of these four pairs are *FAR1*, *CTF4*, *YMR181C* and *YPL162C*. The right panel shows the only sense/antisense transcript pair (of the gene *HSL1*) whose expression profiles have a strong positive correlation.

TF binding motifs for the 37 antisense transcripts I investigated whether the binding motifs of any transcription factors were specifically enriched in the regulatory regions of the periodically expressed antisense transcripts. Enriched TF binding motifs might give indications about the function of these antisense transcripts. For each of the 37 periodic antisense transcripts, I retrieved the sequence of a respective regulatory region, extending from 600 bp upstream of the transcript's start site to 600 bp downstream of the transcript's end coordinate. A control set of sequences was made up by the total set of 639 periodically expressed transcripts, similarly extended by 600 bp both upstream and downstream. No TF binding motif was found to be enriched (with $p \leq 10^{-3}$) in the extended sequences of the 37 antisense transcripts, compared to the control set.

Clustering of periodic transcripts

Hierarchical clustering was used to stratify the 639 periodically expressed transcripts into groups of similar periodic expression. For clustering, I employed a regularised correlation distance measure (Equation (4.2), page 94) between the expression profiles of the transcripts. The expression data used for the clustering were a subset of the α-factor data set, because I considered the α-factor data to be cleaner and to show a higher degree of synchronisation than the *cdc28*$_m$ data. Only the microarray data from 15 minutes after release of the cell-cycle arrest up to 160 minutes after release were used, as the expression profiles of the histone transcripts indicated tight synchronisation during this time (see Figure 4.7).

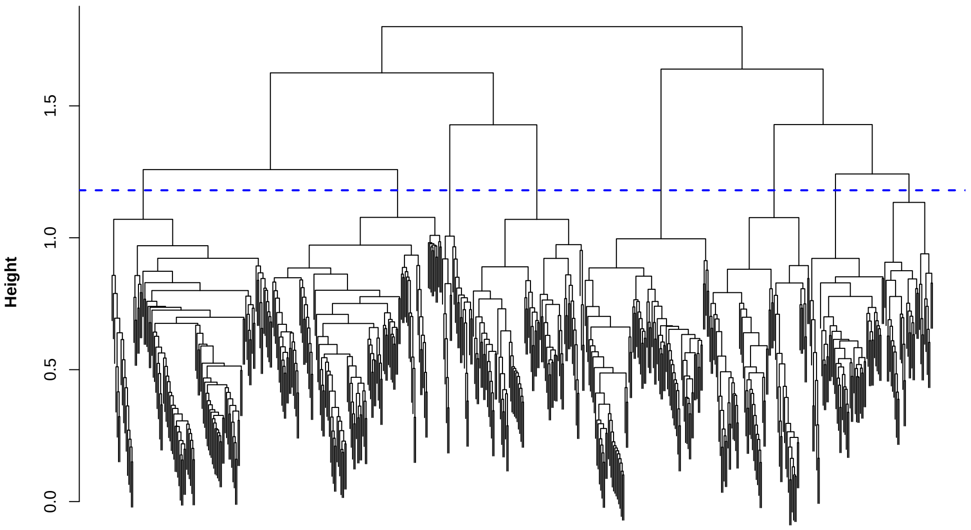


Figure 4.9: *Dendrogram visualising the result of the hierarchical clustering of the expression data of the 639 periodically expressed transcripts. The blue line indicates the height at which I decided to cut the dendrogram, resulting in assignment of the periodic transcripts into 8 clusters.*

The hierarchical clustering dendrogram is shown in Figure 4.9. The displayed horizontal line marks the height at which I decided the cut the dendrogram. This cut resulted in an assignment of the periodic transcripts into eight clusters.

Figure 4.10 shows the resulting heatmap of the expression data of transcripts in the eight clusters. Antisense (AS) and unannotated intergenic (UI) transcripts are distributed over the eight clusters (the colours indicate the mark of the cluster on the left side of the heatmap in Figure 4.10).

- **pink**: 38 transcripts (5 AS, 2 UI), including *ZPS1*
- **purple**: 58 transcripts (2 AS, 0 UI), including *RAS1*
- **red**: 78 transcripts (1 AS, 2 UI), including *CDC20*, *CDC6*, *CLN3*, *FAR1*, *MCM2*, *MCM6*, and *TAF2*
- **brown**: 97 transcripts (2 AS, 1 UI), including *BUD4,CLB1,CLB2*, and *SWI5*

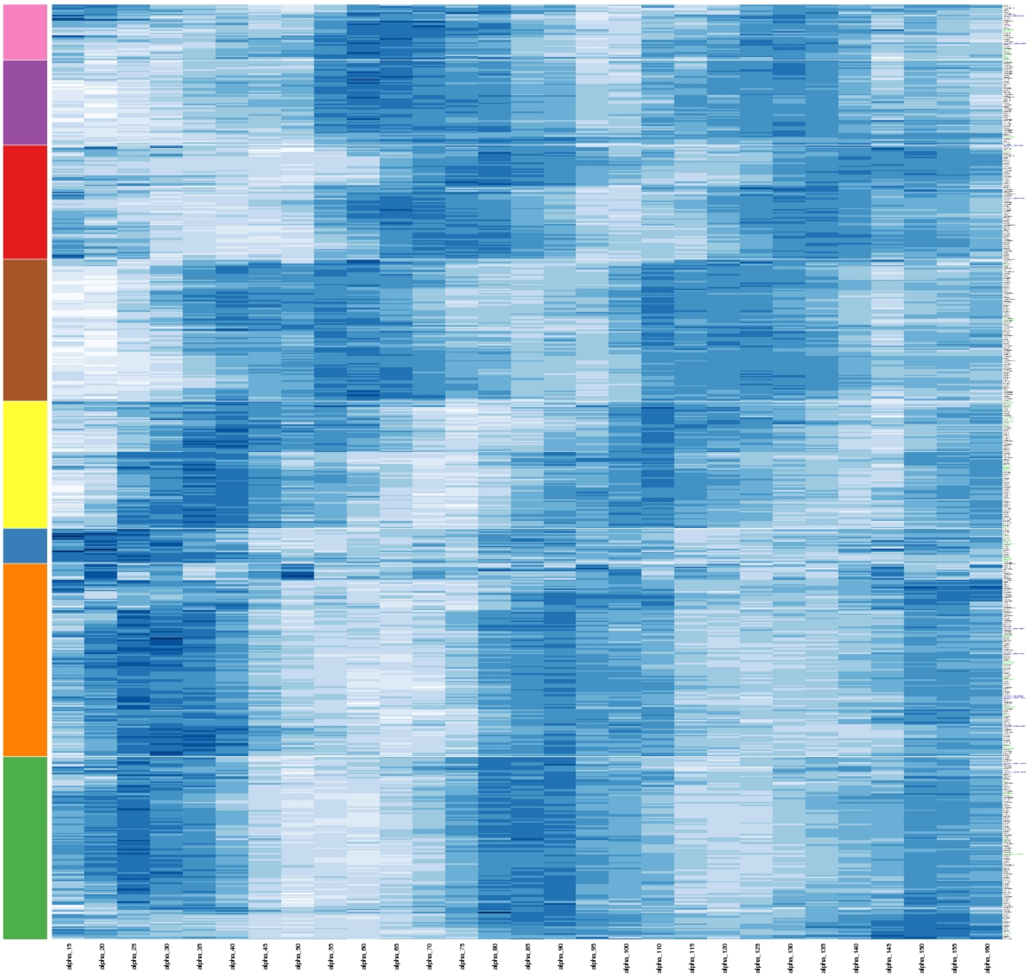


Figure 4.10: *Heatmap showing the expression profiles of the periodic transcripts in the S. cerevisiae α-factor cell cycle data. Each row corresponds to a transcript, each column represents a sample (time point). Columns are ordered by the time they were taken after release from cell cycle arrest and the shown time points correspond to more than two iterations of the cell cycle. Rows (transcripts) are ordered as in the hierarchical clustering dendrogram (see Figure 4.9). The colour bar on the left side of the plot denotes the cluster assignment of each transcript. The colours in the heatmap denote the median expression levels of the transcripts in each sample, with darker shades of blue indicating higher expression levels.*

- **yellow**: 87 transcripts (9 AS, 0 UI), including all histone genes
- **blue**: 24 transcripts (5 AS, 0 UI), including *DIG2*, *YPS1*
- **orange**: 132 transcripts (8 AS, 5 UI), including *CDC11*, *CLN1*, *MF*, and *YOX1*
- **green**: 125 transcripts (5 AS, 2 UI), including *CDC45*, *HO*, *STB1*, and *SWI4*

TFBSs in the clusters I investigated the clusters for over-represented TF binding motifs (as described in Section 4.3.1, page 95f.). For each periodic transcript, a potentially regulatory sequence that extended from 600 bp upstream of the start base to 600 bp downstream of the end base of the transcript was retrieved. For each cluster and each of the 125 PSSMs, I determined how many sequences of transcripts in the cluster matched the PSSM and contrasted that to the matches of the PSSM among all periodic transcripts. Table 4.4 shows the found significantly enriched binding motifs per cluster.

Cluster	**TF**	**# in cluster**	**# in all**	**Score** p
red: 78 transcripts	Mcm1	30	101	$1.3 \cdot 10^{-7}$
	Xbp1	30	146	$6.6 \cdot 10^{-4}$
brown: 97 transcripts	Fkh2	32	101	$3.1 \cdot 10^{-6}$
	Fkh1	28	101	$2.7 \cdot 10^{-4}$
yellow: 87 transcripts	Hcm1	27	107	$2.7 \cdot 10^{-4}$
orange: 132 transcripts	Stb1	46	122	$8.2 \cdot 10^{-7}$
	Swi4	41	108	$3.5 \cdot 10^{-6}$
green: 125 transcripts	Mbp1	43	102	$5.7 \cdot 10^{-9}$
	Swi6	62	195	$4.3 \cdot 10^{-7}$
	Swi4	34	108	$7.9 \cdot 10^{-4}$

Table 4.4: *Table listing which TF binding motifs are enriched in the regulatory sequences of the transcripts in each cluster in contrast to the sequences of all periodic transcripts. Clusters not mentioned here had no significantly enriched TF binding motifs. #: "number of transcripts whose sequences contain one or more matches to the TF's motif".*

Gene Ontology categories of clusters I investigated which GO categories were over-represented among the annotation of the ORF transcripts in each cluster. The number of transcripts with a certain GO annotation in each cluster was contrasted to the number of all periodic transcripts that had this GO annotation. Table 4.5 details which terms were enriched for each cluster, if any.

Cluster	GO Term	Annotated	In Cluster	Expected	p-value
pink	polyamine transport	3	3	0.14	$8.3 \cdot 10^{-5}$
purple	ribosome biogenesis and assembly	22	17	1.70	$1.5 \cdot 10^{-7}$
	rRNA processing	5	4	0.39	0.00014
	localisation	61	12	4.73	0.00079
yellow	microtubule nucleation	15	9	2.43	0.00011
orange	membrane lipid biosynthetic process	8	7	1.51	$4.6 \cdot 10^{-5}$
green	lagging strand elongation	9	9	1.92	$6.5 \cdot 10^{-7}$
	mismatch repair	7	7	1.49	$1.6 \cdot 10^{-5}$
	leading strand elongation	5	5	1.07	0.00040
	DNA replication	48	28	10.23	0.00047
	double-strand break repair	12	8	2.56	0.00075

Table 4.5: *Table listing which GO terms are enriched in the gene annotation for each cluster in contrast to the annotation of all periodic transcripts. "Annotated" refers to the number of all periodically expressed genes that are annotated with the respective GO term. Clusters not mentioned here had no significantly enriched GO terms with $p \leq 10^{-3}$.*

Transcription patterns

In addition to looking at periodically expressed transcripts in the two cell-cycle data sets, I surveyed the transcript orientation (see Figure 1.1) of all pairs of adjacent, nearby transcripts, regardless of whether the transcripts showed periodic or constant expression.

There have been previous studies investigating pairs of adjacent genes in *S. cerevisiae*. For example pairs of adjacent genes, the expression profiles of the two transcripts of each pair were observed to be highly correlated (Pearson CC $\geq$ 0.63) over time in *cdc28*$_m$ cell-cycle microarray data [49, 20]. The distribution of intergenic distances, i.e. the distances between annotated ORFs, has been investigated on a whole-genome scale in *S. cerevisiae* and related to the orientation that the two adjacent genes have to each other [138].

The high-resolution tiling microarray data allow us to consider the actual start and end base coordinates of transcripts in place of the annotated ORF coordinates. In addition, non-ORF transcripts are also taken into account.

Of particular interest are pairs of divergent transcripts (Figure 4.11), as such pairs with highly correlated expression profiles may be regulated by single promoter regions with bidirectional activity. Bidirectional promoter regions have been described in *H. sapiens* and were found to have the following characteristics [21, 139]:

- are short inter-transcript regions, less than 1 kb long, with the majority being less than 300 bp long.
- have a higher median GC content (66%) than unidirectional promoters (53%)
- often show specific TF binding motifs (such as the one of the GABP complex) and lack of a TATA box.

With *S. cerevisiae*, it is unclear whether bidirectional promoters are a common category of promoters and, if so, how promoters of this category might be characterised in general. Only a few examples of bidirectional promoters have thus far been described and characterised. One previously described pair of divergent transcripts is *UGA3–GLT1* and certain elements, such as the binding site of Abf1, were found to be important for the bidirectionality of this shared promoter region [140].

In general, an inter-transcript region of 1000 bp may be considered small in the human genome (size: $3 \cdot 10^9$ bp), but it would be large in com-

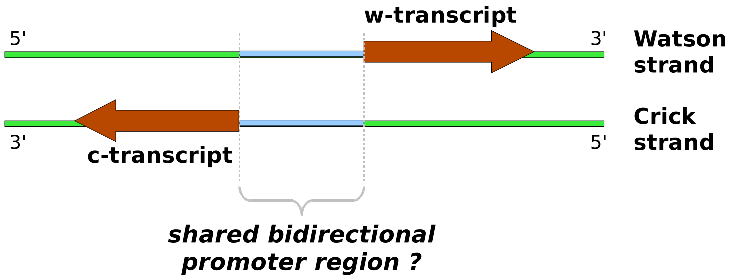


Figure 4.11: *Scheme of divergent transcripts with putative bidirectional promoter region in between.*

parison with the untranscribed part of the *S. cerevisiae* genome. With the α-factor cell-cycle data, we observed that of the 12,162,996 bp of the S288c strain's genome, 9,043,789 bp ($\approx$ 75%) are contained within transcribed segments, and the non-transcribed regions between transcripts have an average length of merely 558 bp.

Data for orientation analysis The segment tables from the cell-cycle tiling microarray data were further post-processed by merging directly adjacent transcribed segments[9] into single transcripts. Directly adjacent transcribed segments might otherwise have been interpreted as pairs of tandem transcripts, although in most cases they were subsections of the same transcripts. Non-adjacent transcribed exons of the same gene were also merged into single transcripts. Merged "super-segments" were assigned an expression level equal to the weighted average of the individual segment levels, weighted by the lengths of the individual segments. Each super-segment and each transcribed segment that had not been merged was further considered to correspond to a transcript.

For this analysis, I investigated all pairs of adjacent transcripts with untranscribed regions of length $\leq$ 400 bp in between them. Depending on the orientation of the two transcripts to each other (see Figure 1.1), a pair is considered to be *divergent*, *convergent* or *tandem*. A transcript t_i can be part of more than one pair, if more than one other transcript on either strand

[9] Two transcribed segments were considered to be directly adjacent if there was no untranscribed or excluded segment between them on the same strand.

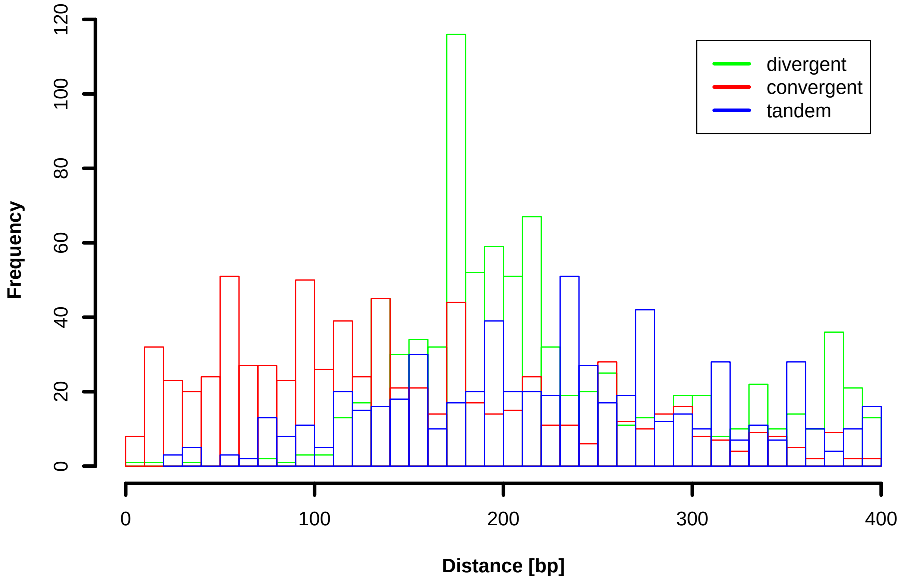


Figure 4.12: *Distance between adjacent transcripts in the α-factor cell-cycle data. The transcript pairs were stratified by orientation of the two transcripts to each other (Figure 1.1).*

has a start or end coordinate within 400 bp of the start or end coordinate of transcript t_i.

For the cell-cycle data, I observed 839 pairs of divergent transcripts, 564 pairs of convergent transcripts, and 370 pairs of tandem transcripts.

Distance between adjacent transcripts The overlayed histograms in Figure 4.12 show the base-pair distances between each pair of adjacent transcripts. Note that neither divergent nor convergent transcript pairs were allowed to contain pairs of overlapping transcripts. Thus, all pairs had an inter-transcript region that is ≥ 0 bp and ≤ 400 bp long. Convergent transcript pairs tend to show inter-transcript regions smaller than 200 bp. With tandem transcripts, the distances were more or less uniformly distributed between 20 and 400 bp. For divergent transcripts, there was a clear peak at 170–180 bp, and distances between 180 and 220 bp were also frequent.

Correlation of adjacent transcripts' expression profiles I investigated whether the orientation of two nearby transcripts affects the correlation

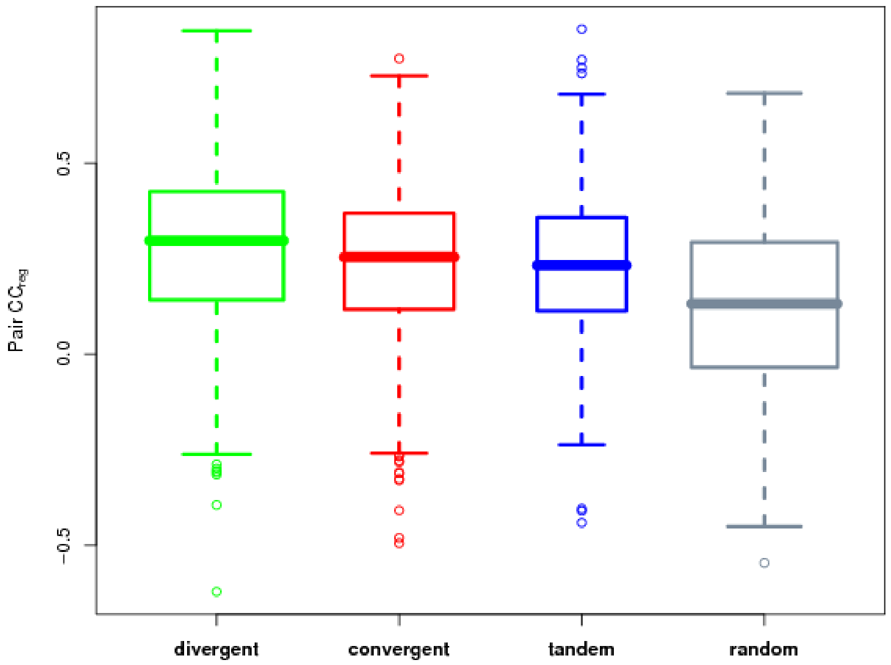


Figure 4.13: *These box plots show the distribution of the regularised correlation coefficients between the expression profiles of the two transcripts of a pair. Pairs are stratified by the orientation of the two transcripts to each other (Figure 1.1). For comparison, the correlation coefficients for 1000 random transcript pairs are also shown. The widths of the boxes are proportional to the numbers of pairs in each category.*

of their expression profiles. For each pair of divergent, convergent and tandem transcripts, the regularised correlation coefficient (Equation (4.1)) of the expression profiles of the two transcripts was computed (s_0 was set to the 10% quantile of the observed standard deviation products of all adjacent transcript pairs). For comparison, 1,000 pairs of random transcripts were generated. The two partners of each such pair were drawn at random from the set of transcripts on one chromosome without any restriction on the orientation and distance of the two transcripts to each other. The observed correlation coefficients by orientation are shown as box plots in Figure 4.13. The expression profiles of divergent transcripts were found to show significantly higher correlation than those of convergent transcripts ($p = 1.14 \cdot 10^{-5}$) and tandem transcripts ($p = 2.57 \cdot 10^{-5}$).

Divergent, convergent and tandem transcript pairs showed higher correlation than random transcript pairs ($p \leq 6.27 \cdot 10^{-14}$); p-values were derived from two-sample Wilcoxon rank-sum tests with the null hypothesis being that the two distributions of regularised CCs in pair categories a and b have the same location, and the alternative hypothesis being that the distribution of CCs in category a was shifted to higher values). Among the highly correlated ($|CC_{reg}| \geq 0.5$) divergent transcript pairs is the pair *YKR085C–YKR086W*, which has previously been reported as a pair of divergent, co-regulated transcripts [141]. Another study reported that this specific pair did not show correlated expression profiles [20], but in our data the expression profiles are clearly correlated ($CC_{reg} = 0.52$).

Promoter regions of divergent transcripts The regions between the TSSs of divergent transcripts were investigated for characteristic attributes. To this aim, I obtained the nucleotide sequences of these regions. A control set was made up of the nucleotide sequences of the inter-transcript regions from all considered divergent, convergent, and tandem transcript pairs and the 400 bp sequences upstream of the TSSs of the Watson-strand transcripts from the 1,000 random pairs. In total, the control set consisted of 2,279 sequences of potential regulatory regions of expressed transcripts. I investigated whether any of the TF binding motifs (as specified by MacIsaac and co-workers [132]) were enriched in the sequences of the putative bidirectional promoter regions between divergent transcripts (Section 4.3.1, page 95). Two TF binding motifs were found to be significantly enriched in the promoter regions between divergent transcripts. One is the motif of Abf1, which matches 57 of the 839 promoter regions between divergent transcripts, but only 101 of all 2,279 analysed regulatory regions of the control set of transcripts ($p = 3.21 \cdot 10^{-5}$). The other motif enriched for divergent transcripts is the one of Rpn4 (matches 54 promoters of divergent transcripts and 101 of all analysed sequences, $p = 3.58 \cdot 10^{-4}$).

Lin and co-workers recently published a set of motifs that they found to be overrepresented in *H. sapiens* bidirectional promoter regions, including the motifs of known transcription factors, such as the one of GABP, and new motifs that the authors found using *ab initio* motif discovery tools [139]. GABP and the other TFs specified by Lin *et al.* have no homologous TFs in *S. cerevisiae*. Thus I constructed PSSMs from the motifs as they were specified for *H. sapiens* and matched them to the analysed *S. cerevisiae* sequences. Only one of these motifs was highly overrepresented among the promoters of divergent transcripts (69/839 as compared to 120/2,279, $p = 1.79 \cdot 10^{-6}$), a motif that Lin *et al.* had discovered *ab initio*. This motif has a consensus sequence of "RAAATTTTTCA" where "R" stands for either adenine or guanine.

The promoter regions of divergent transcripts were not found to differ in their GC content from the control set of sequences.

4.3.3 Discussion

We analysed two time courses of whole-genome tiling microarray data sets for the periodic expression of transcripts and for the orientation of adjacent transcripts relative to each other.

We identified a high-confidence list of 639 transcripts that showed periodic expression patterns concordant with cell cycle progression. These were mostly ORF transcripts, but included 37 transcripts that were antisense to annotated ORFs and 11 unannotated intergenic transcripts from regions in which no genome features were annotated on either strand.

I clustered the 639 periodically expressed transcripts into eight groups of similar periodic expression patterns (see Figure 4.9). Each cluster was characterised by enrichment for transcription factor binding motifs and by over-representation of Gene Ontology annotations. The TFs with enriched binding motifs are known regulators of cell cycle progression. Hcm1 is a known regulator of S phase expressed genes [45] and the one cluster (yellow) that shows enrichment of the binding motif of Hcm1 contains the histone genes. The expression of this cluster of transcripts peaks in the time points corresponding to the S phase. This cluster also contains 9 antisense transcripts, which may be involved in regulation of expression of other transcripts during the S phase. Another cluster shows enrichment for Mbp1, Swi6 and Swi4. These TFs have been implicated in the regulation of late-G1 transcript expression [43] and the transcripts in this cluster accordingly show elevated expression at time points in the late G1 phase. The GO terms that are enriched in this cluster agree with the specialised role of the genes in this cluster during the end of the G1 phase, as all these GO terms describe functions important for DNA replication. And another cluster that shows peak expression at the G1/S transition shows enrichment in the binding motif for the transcription factor Stb1, which is involved in the G1/S transition [44]. At that stage, the genes required for DNA replication seem to have ceased being transcribed, and the GO annotation that is enriched among genes in this cluster is the synthesis of membrane phospholipids, possibly in relation to the membranes of the daughter cells after mitosis and cytokinesis.

The enriched TF binding motifs and the enriched GO terms suggest specific roles of these clusters of transcripts during cell cycle progression. The assignment of transcripts into eight clusters seemed reasonable considering the hierarchical clustering dendrogram (Figure 4.9). The dendrogram cut was arbitrary, however, and so small transcript groups of specialised function may have been overlooked, as they might have been absorbed into bigger clusters. On the other hand, the two late-G1 clusters (green, orange) only differ in the expression profiles of included transcripts at a few time points, and the heatmap (Figure 4.10) suggests that these could be interpreted as one single cluster. Both clusters also show enrichment for the binding motif of Swi4. The enrichment of TF motifs and GO annotations, however, indicates that there are also functional differences between the two clusters. Nevertheless, this example stresses the fact that the proposed clustering is only one of many possible clusterings of the periodically expressed transcripts, based on a regularised correlation distance between transcript expression profiles.

I have shown that a clustering of this kind is able to group the periodic transcripts into functional modules that are involved in driving the progression of the cell cycle at different stages of the cycle. Spellman *et al.* suggested a similar clustering of periodically expressed ORF transcripts, based on gene expression microarray data [48]. The results presented here extend their findings, as the tiling array data set is of higher resolution and additionally encompasses non-ORF transcripts. The 37 periodic antisense transcripts, as well as the 11 unannotated intergenic transcripts were split between the different clusters. As the samples had been treated with actinomycin D before hybridisation to the microarrays, it can be assumed that a negligible number of transcribed antisense segments are second-strand artifacts from the reverse transcription [129]. The role of antisense transcripts in *S. cerevisiae* is yet unclear, as *S. cerevisiae* lacks DICER and other components of the RNA interference (RNAi) pathway [128]. The expression of four of the periodically expressed antisense transcripts is negatively correlated with the opposite sense transcripts' expression, but the majority of the periodic antisense transcripts does not show any substantial correlation with the expression profile of the respective sense transcript. Thus, their role does not seem to be RISC-mediated decay or, more generally, mRNA degradation mediated through complementary base pairing. At present, we can only speculate on the potential role of the periodically expressed antisense transcripts, and on the role of the periodic unannotated intergenic transcripts.

Nevertheless, our findings indicate that non-coding transcripts are involved in regulatory processes during different stages of the cell cycle. Follow-up studies are required to elucidate the actual role of these transcripts.

Transcript orientation Examples of divergent transcript pairs with correlated expression patterns have been described previously in the *S. cerevisiae* genome [20]. We surveyed the whole transcriptome of *S. cerevisiae* in an unbiased manner and found multiple pairs of adjacent transcripts with regions of length ≤ 400 bp in between them. These pairs were stratified into divergent, convergent and tandem transcript pairs (Figure 1.1). Divergent transcripts were found to show stronger correlated expression levels than convergent transcripts, tandem transcripts and random transcript pairs. Multiple pairs of divergent transcripts have a genomic region of length 170–180 bp between their TSSs, with regions of length 180–220 bp also being common. A bidirectional promoter region of length 170–180 bp could be explained in the light of two recent studies, in which the positioning of nucleosomes in relation to transcriptional start sites was investigated [142, 143]. Many TSSs were found to be preceded by a short region of about 140 bp that lacks any nucleosomes (a nucleosome-free region (NFR)) and contains binding sites for transcription factors [142]. Adjacent to this region is a nucleosome which covers the TSS and starts about 13 bp upstream of the TSS [143, the authors used the TSS positions as we had annotated them in the exponential growth phase transcriptome study (Section 4.2)]. A region of 170–180 bp could roughly correspond to one single NFR that regulates both of the divergent transcripts and is flanked by two nucleosomes that cover the TSSs of the two transcripts (≈ 140 bp $+ 2 \cdot 13$ bp; note that, due to the reporter spacing on the microarray [Section 4.1], our estimate of the exact TSS position can only be accurate within ± 7 bp). A general relation between transcript orientation and nucleosome positioning has also been suggested [142].

It remains to be seen how many of the regions in between observed divergent transcripts are truly bidirectional promoter regions. Such bidirectional promoters have been thoroughly characterised in *H. sapiens* [21, 139], but the described characteristics were not found to be directly applicable to the regions between divergent transcripts in *S. cerevisiae*. However, two out of 125 known binding motifs of *S. cerevisiae* transcription factors were found to be enriched in the promoter regions between divergent transcripts. One is the binding motif of the transcription factor Abf1, a TF with potential chromatin-reorganising activity. The binding motif of Abf1 has previously been observed to be involved in regulating the bidirectional transcription of the gene pair *UGA3–GLT1* [140]. This gene pair is also among the divergent transcript pairs analysed here, and I found the motif of Abf1 in the promoter region between these transcripts and also in 56 other promoter regions between divergent transcripts. As Abf1 was seen to be involved in regulating the bidirectionality of at least one transcript pair (*UGA3–GLT1*), it is plausible that it should regulate further such pairs. The other enriched known binding motif is the one of Rpn4. This TF was reported to stimulate the expression of proteasome genes [144], and as far as I am aware has never been associated with bidirectional promoters. The enrichment of the binding motif of Rpn4 here, however, might be a chance observation and have no functional relevance.

Other transcription factors whose binding modalities are as yet unknown could be responsible for divergent transcription patterns in budding yeast. Lin and co-workers reported sequence motifs that were significantly enriched in bidirectional promoter regions in *H. sapiens* [139]. Of the motifs that they reported, only one that they found *ab initio* using motif-discovery tools was over-represented in the promoter regions between divergent transcripts in *S. cerevisiae*. This motif has a well-defined consensus sequence "RAAATTTTTCA", as ten out of its eleven nucleotide letters are unambiguous. The biological significance of the enrichment of this motif in the promoter regions of divergent transcripts in *S. cerevisiae* and in bidirectional promoters of *H. sapiens* is unclear. Whether this motif happens to be enriched by chance, corresponds to the binding motif of a yet unidentified TF or is a sequence motif that is relevant for nucleosome positioning remains to be determined.

Taken together, all the described observations of divergent transcript pairs and the promoter regions in between them only apply to a fraction of all divergent transcript pairs in the cell-cycle data. The complete set of divergent transcript pairs thus likely consists of two subsets. One subset contains the divergent transcripts that are regulated by a shared bidirectional promoter regions. Further studies are needed to precisely identify this set and to further define the characteristics that bring about the bidirectionality. The other subset consists of pairs of divergent transcripts that are regulated separately by unidirectional promoter regions. Since the genome of *S. cerevisiae* is very densely covered with regions that are transcribed and the average distance between two transcripts is merely 558 bp, having a region of length $\leq$ 400 bp in between any two transcripts does not necessarily indicate any functional connection between the two transcripts. The cutoff of 400 bp was an arbitrary choice based on previous observations that many interesting pairs of adjacent transcripts were less than this distance apart from each other [20]. Due to this cutoff and the restriction to non-overlapping transcripts, some relevant pairs of adjacent transcripts might have been missed.

Nevertheless, the presented findings about transcript orientation in the tiling microarray data, as well as the operon-like transcripts observed earlier (Section 4.2), provide evidence against the existence of a simple "one promoter, one transcript" concept in eukaryotes. Even an elementary model organism, such as *S. cerevisiae*, can provide important insights into the process of transcriptional regulation. As our studies show, high-throughput technologies are an excellent tool for figuring out further pieces of the puzzle of transcription and transcriptional regulation.

Contributions

The microarray data were provided by Marina Granovskaia and Lars Steinmetz. I performed all the described analyses, except for the identification of periodic transcripts, which was done by Marina Granovskaia, Matt Ritchie and Lars Juhl Jensen.

Chapter 5

Software for reproducible research

Over the last two decades, programming source code (preferably in a widely spoken high-level programming language, such as R [145]) has emerged as an essential medium for the communication and critical academic evaluation of methodology in the biological sciences. Increases in the scale of experimental designs and in the availability of personal computers have fostered the popularity of this medium. Also, demand has increased for functional, well-documented software providing data analysts with generic tools, which they can reuse and extend for their own research needs. This is especially true for the field of bioinformatics, which has seen an exponential growth in data, and yet, in the early days, mostly was based on slightly cryptic and poorly documented software tools that had evolved from individual academic research projects. These tools were not easily reusable, and researchers moving into the field had to rewrite their own solutions to generic problems or to take pains to familiarise themselves with the software [146].

By now, however, the benefits of sharing software have become apparent, and interoperable software tools are publicised in a similar manner to biological research insights. In this chapter, I describe two software packages, which I have (co-)developed. Both packages, *Ringo* [147] and *tilingArray* [148], are integrated in the Bioconductor project [104], which is a collection of packages that extend the statistical environment R [145]. Bioconductor is an open source and open development software project for the analysis and interpretation of genomic data. Bioconductor offers tools that cover a broad range of computational methods, visualisations and experimental data types. The design of these tools allows the construction of scalable, reproducible and interoperable workflows.

5.1 *tilingArray*

5.1.1 Introduction

In 2005, high-resolution tiling microarrays with reporters for representing a whole genome were relatively new. Most available software for microarray analysis focused on expression microarrays, the reporters of which only represent the annotated ORFs of an organism. The unique challenges posed by a whole-genome microarray platform required new data processing and analysis approaches, and suitable extensions to existing methods developed for expression microarrays. We conducted a high-resolution tiling microarray study that aimed to redefine the transcriptome of *S. cerevisiae* ([9], Section 4.2). From the programming code that was written to facilitate the analysis of the *S. cerevisiae* tiling arrays, we constructed the Bioconductor package *tilingArray*. The package provided other users with the means to conduct similar analyses and to reproduce our results. The package was developed based on a transcriptome study using a 25mer oligonucleotide microarray platform. Some functions in the package *tilingArray* are specific to such a study, while other functions are sufficiently general to be applied to other tiling microarray platforms and study objectives. In accordance with Bioconductor's open source, open development philosophy, *tilingArray* is also meant as a starting framework for the analysis of tiling arrays, which bioinformaticians can edit, extend and adapt for their own analysis needs.

5.1.2 Reporter annotation

During the analysis of tiling microarrays, a mapping of the reporters to the genome sequence is required and needs to be made available in the analysis software. Users may decide to use the reporter mapping annotation that is provided by the manufacturer of the microarray. As long as the complete genome sequence is known, however, a custom association of reporters to genomic locations can be established via alignment of the reporter sequences to the genome sequence. Powerful, specialised alignment tools for mapping multiple short sequences to a whole genome sequence are available. These include *Exonerate* [149] and the Bioconductor package *Biostrings*. These tools are flexible with respect to the conditions for a true *match* of a reporter to a genome segment. One consideration is whether to allow mismatches between the reporter sequences and the "matching" genome segment. If mismatches are allowed, different types of mismatches might be treated differently. Mismatches at the ends of the reporter have been shown to impede hybridisation less than mismatches in the middle of the reporter [150].

Once the reporters have been mapped to the genome, the following information needs to be extracted from the output of the alignment tool and stored in a sensible data structure in the analysis environment: Which reporters match the genome, and at what genomic coordinates? How many distinct genomic coordinates are matched by each reporter? Only reporters that uniquely match one single genome position are informative for assessing transcription at that position. The readouts of reporters that match multiple genome positions are harder to interpret as their level is affected by cross-hybridisation. Still, there are cases in which even reporters with more than one genomic match may be of interest. If, for example, a reporter matches two genomic positions that are located in a pair of duplicated genes, the reporter level could provide insights into the duplication event.

The most natural way of representing the reporter-to-genome mapping in R would be to use the `data.frame` class. However, repeatedly extracting subsets of a data.frame for a genomic region of interest is too slow for practical purposes. *tilingArray* employs an object of class environment to store the mapping. Per chromosome strand, the object holds four vectors of equal length and ordering that specify at which genomic positions reporter matches start and end, what identifiers or indices these reporters have in the intensities data, and whether these reporters match uniquely to the genomic positions[1].

5.1.3 Data preprocessing

On tiling microarrays, the whole genome is represented by reporters, in contrast to expression or exon microarrays, which only represent annotated ORFs. The reporters of tiling microarray therefore cannot be expected to meet equal standards of selection as seen with commercial expression microarray platforms. The large number of reporters results in large variations in staining and hybridisation characteristics.

The package *tilingArray* includes functionality for adjusting raw tiling microarray reporter levels measured from RNA samples of interest, based on the output of a genomic DNA hybridisation. We assume that a reporter's intensity level consists of a reporter-specific background level plus a signal level that is proportional to the concentration of the reporter's RNA target:

$$x_i = bg_i + c_i \cdot [\mathrm{RNA}_i] \tag{5.1}$$

where x_i is the intensity measured for reporter i, bg_i is the reporter's background level and c_i is a hybridisation factor denoting how well the reporter measures the target RNA concentration.

Naively, one could assume that both bg_i and c_i depend on the reporter's

[1] The Bioconductor package *Ringo* (see Section 5.2) implements the S4 class probeAnno that extends and solidifies this environment concept by a number of structural assertions.

affinity to the target multiplied by some constant:

$$\begin{aligned} bg_i &= a_i \cdot bg_0 && (5.2) \\ c_i &= a_i \cdot c_0 \quad . && (5.3) \end{aligned}$$

If this assumption holds, dividing the raw intensity x_i by an estimate of this affinity $\widehat{a_i}$ would yield a normalised expression level for reporter i:

$$\frac{x_i}{\widehat{a_i}} = bg_0 + c_0 \cdot [\mathrm{RNA}_i] \quad . \tag{5.4}$$

In practise, this naive approach has not been found sufficient to normalise reporter levels [24], since assumption (5.2) cannot account for cross-hybridisation and other effects that influence the reporter-specific background signal. A better estimate of the reporter-specific background $\widehat{bg_i}$, accounting for non-linear dependencies between the reporter's affinity and its background signal, is required. An excellent way to obtain such reporter-specific background $\widehat{bg_i}$ estimates is to hybridise genomic DNA samples to the same microarray platform that is used to analyse the RNA samples of interest.

Let X_{ij} be the raw intensity of the i-th reporter on the j-th array, then the normalised reporter intensity $\mathbf{X}_{ij}$ is given by

$$\mathbf{X}_{ij} = \frac{X_{ij} - B_j(a_i)}{a_i} \tag{5.5}$$

where a_i is a reporter-specific affinity factor, $B_j(a)$ is a continuous function estimating the background intensity of reporters with affinity a on array j. The affinities a_i are estimated by the geometric mean of reporter i's intensities from the genomic DNA hybridisations. For calculating B_j, the reporters are grouped into strata corresponding to quantiles of a_i. Within each stratum, and for each array j, the midpoint of the shorth is calculated from the distribution of intensities of the reporters whose reporter match positions (RMPs) do not overlap any genome feature[2] on the same

[2] Genome feature is used as a summary term for annotated ORFs, uORFs, pseudo-

or opposite strand. The function B_j is obtained from these values by linear interpolation between them.

The values $\mathbf{X}_{ij}$ are transformed to the *generalised logarithmic* (glog) scale using the *vsn* method [25].

Finally, the function discards reporters with very low reporter-specific affinities a_i, by default the ones corresponding to the 5% lowest affinities a_i. When taking the ratio (5.5), low affinities a_i may result in normalised reporter levels $\mathbf{X}_{ij}$ that are artificially high, even though these reporters are uninformative.

Figure 4.2 (page 80) displays the effect of the normalisation-through-DNA-hybridisations on reporter levels in an example genomic region. The normalisation procedure is implemented in the function `normalizeByReference` of package *tilingArray*.

This normalisation procedure gave the best improvement of the signal-to-noise ratio in the yeast tiling microarray data (see Section 4.2). The package *tilingArray* uses the basic Bioconductor class `ExpressionSet` to store the array data; thus, users can easily apply alternative preprocessing methods from other packages or of their own devising. If, for example, the microarray platform contains mismatch (MM) reporters in addition to the perfect match reporters, users might consider using these for estimating the reporter-specific background intensity bg_i, as done in the MAS5 preprocessing method proposed by Affymetrix [71].

5.1.4 Segmentation

For certain applications of tiling microarrays, such as to distinguish transcribed genomic regions from untranscribed ones, the reporter-wise resolution of the signal may be too fine-grained. Genomic segments transcribed by a polymerase enzyme can measure thousands of base pairs in length and with most tiling microarrays encompass more than one reporter match position (RMP). Moreover, since single reporter measure-

genes, ncRNAs, repeat regions, and transposable elements.

ments are not reliable, segments whose status as *transcribed* is supported by multiple reporters are less likely to be false positives than those supported by few reporters.

Fitting a piece-wise constant function

The package *tilingArray* contains a segmentation function, in which intensity values along the genome are divided into segments of approximately constant hybridisation signal using a dynamic programming algorithm. The segmentation algorithm is based on the structural change model, which is well-established in econometrics [151] and which has already been applied for segmenting array-CGH data [152].

Given m replicate microarrays, with a set of k reporters that match k ordered genomic positions on one chromosome strand, the aim is to find the optimal division of the positions into S segments representing regions of similar reporter levels. To this aim, the algorithm finds an ordered set ζ of $S-1$ segment boundaries, or change points, $\zeta = \{t_2, t_3, \dots, t_S\}$ that minimise the cost function

$$G(\zeta) = \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{i \geq t_s}^{i<t_{s+1}} \left(x_{ij} - \overline{x}_{sj}\right)^2 \tag{5.6}$$

where x_{ij} is the normalised signal of the i-th reporter on the j-th replicate microarray ($j = 1, \dots, m$), $t_1 = 1$, $t_{S+1} = k+1$, and $\overline{x}_{sj}$ is the arithmetic mean of the signal values of array j in segment s. The loss function (5.6) is the sum of squared residuals for all reporters from the mean values of their respective segments. In this notation, two segment boundaries are fixed a priori: t_1 is the first reporter-match position and t_{S+1} is the hypothetical one which would follow the last actual RMP. Hence, for a fixed value of S, there are $\binom{k-1}{S-1}$ different sets of segment boundaries ζ and the algorithm finds the one set that minimises the cost function (5.6).

This minimisation can be reformulated in terms of likelihoods. For the expression levels x_i of all reporters belonging to segment s, we assume the model

$$\forall i \in s \ : \ x_i = \mu_s + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma^2). \tag{5.7}$$

We assume that the residuals ε_i are independently identically distributed. For μ_s and σ, we use the maximum likelihood estimators

$$\hat{\mu}_s = {n_s}^{-1} \sum_{i \in s} x_i \text{ and } \hat{\sigma}^2 = n^{-1} \sum_i (x_i - \hat{\mu})^2 \tag{5.8}$$

where n_s denotes the number of reporters making up segment s, n is the total number of reporters measured, and $\hat{\mu}$ is the mean of all reporter levels.

Setting the parameter S

The single parameter of the segmentation algorithm that needs to be set is S, the number of segments. This parameter controls the sensitivity-specificity trade-off of the algorithm. A setting for the parameter S can be derived by a maximum-likelihood approach. The cost function (5.6), however, has its absolute minimum of 0 if S is equal to k, the number of RMPs on the chromosome strand, i.e. if each RMP makes up an individual segment.

Obviously, this value for S is not useful. An alternative way to derive a setting for S is to employ a penalised maximum-likelihood approach. A penalty term that depends on the number of segments, the model parameters, is subtracted from the (log-)likelihood and the value that maximises the penalised likelihood yields a setting for S. Two such penalised likelihoods are well-known, the Akaike Information Criterion [153] and the Bayesian Information Criterion (BIC) [154]. The package *tilingArray* includes documentation describing how to use compute the penalised likelihood and how to obtain the parameter setting S_{IC} that is deemed optimal in those approaches. The Bayesian Information Criterion, in particular, works well for estimating the number of segments in simulated data. In our study of the yeast transcriptome (see Section 4.2), however, the estimate $\widehat{S}_{BIC}$ largely exceeds a parameter setting deemed suitable, in particular with regard to specificity. The reason may be that the model (5.7) is too simple to represent important aspects of high-resolution tiling microarray data; notably the assumption that the residuals in each segment are independently identically distributed is too naive. The model (5.7) is useful to estimate meaningful segment boundaries once the parameter S is set appropriately, but it might not be sufficiently powerful for inferring S from the data.

We set the parameter S on the basis of data exploration and simulation studies, to guarantee a high sensitivity in the segmentation approach. High specificity was attained by pruning steps following the segmentation (see Section 4.2).

Alternative segmentation approaches

Sliding average thresholding If the aim is to identify segments of a limited number of types, say “transcribed” and “untranscribed”, a combination of a sliding window algorithm and a thresholding algorithm could be considered. In the sliding window algorithm, reporter levels are first smoothed along each strand or chromosome, taking into account the intensities of reporters nearby. In the thresholding algorithm, the RMPs at which smoothed reporter levels exceed a provided threshold determine the segment boundaries.

Such an approach has been used in some early tiling microarray studies for the identification of transcripts (see, e.g., [8, 30]). The major drawback of such a segmentation approach is that a sliding window of fixed size typically leads to biased estimates of the segment boundaries. The amount and direction of this bias is determined by the degree by which signal levels in the transcribed segment exceed the background signal. Figure 5.1 clarifies this problem. Use of such a sliding-window algorithm typically allows the qualitative assessment of whether there is signal above a fixed threshold or not. However, the algorithm fails to provide an accurate estimate of the

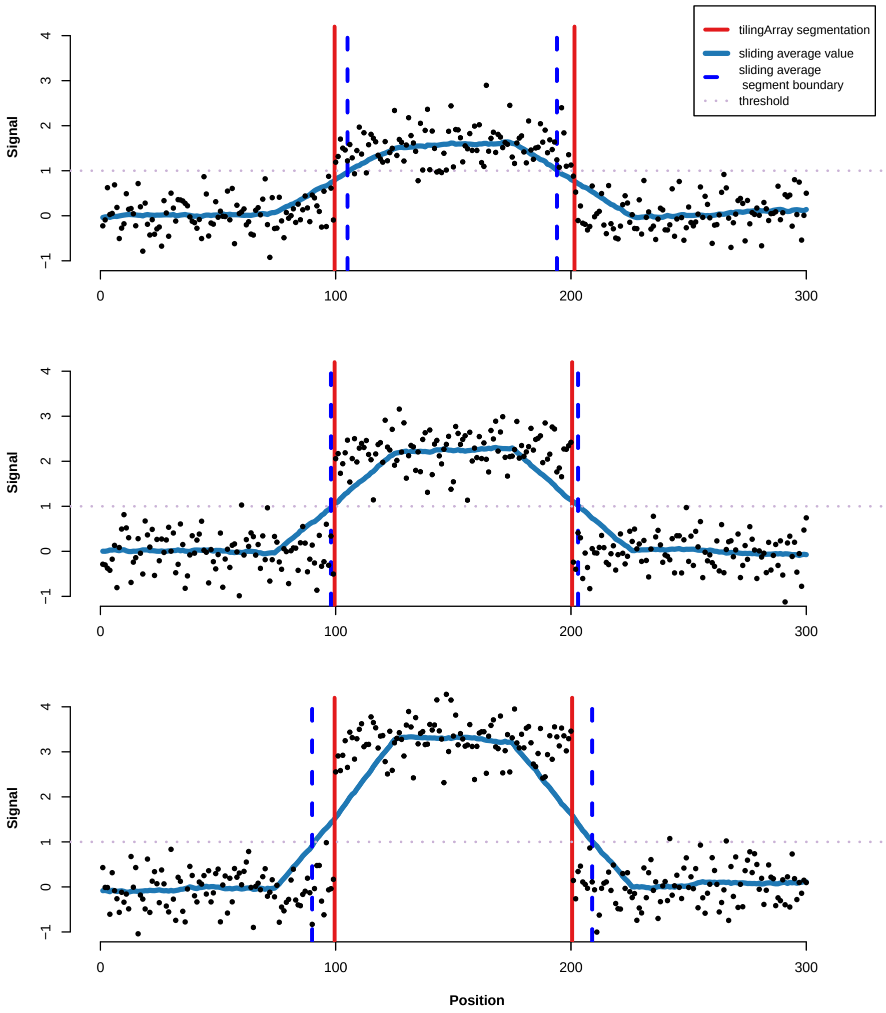


Figure 5.1: *Comparison between the tilingArray segmentation algorithm and sliding average thresholding (SAT).* ***top****: The dots correspond to simulated data for a weakly expressed transcript starting at position 100 and ending at 200. The vertical dashed green lines show the segmentation boundaries set by the segmentation algorithm that is implemented in tilingArray. The blue line shows a sliding average (window width of 50). The vertical dashed light-blue lines show the segment boundaries found by thresholding the sliding average at a threshold of* $y = 1$ *(horizontal dotted line). The SAT results in too short a transcript estimate.* ***middle****: As in the top panel a), for a moderately expressed transcript. The segment boundaries estimated by both algorithms nearly coincide.* ***bottom****: As in the top panel, for a strongly expressed transcript. The SAT estimate of transcript length exceeds the actual length. The tilingArray segmentation produces unbiased estimates for the segment boundaries in all three cases.*

length of the segment that exceeds the threshold. Note that the simulated data used in Figure 5.1 were generated according to the model (5.7) and unsurprisingly the *tilingArray* segmentation algorithm accurately recovers the segment boundaries. In the tiling microarray study of the yeast transcriptome (Section 4.2), we saw abrupt change points in the data, similar to the ones in the simulated data in Figure 5.1. Hence, the drawback of the sliding average thresholding needs to be taken into account with tiling microarray data. An modified SAT algorithm, in which the median instead of the arithmetic mean is used for smoothing, is less affected by the variation in signal levels.

Hidden Markov Models Hidden Markov models (HMMs) [98] provide an alternative for segmenting tiling-microarray data. The Bioconductor package *snapCGH* contains an HMM-based algorithm for segmenting array-CGH readouts [155]. An HMM approach is suitable for recovering a finite number of levels, such as DNA copy number variations, and, with appropriate parameter settings, allows accurate identification of segment boundaries. HMMs, however, are limited in the sense that the number of hidden states needs to be fixed. This may be problematic for the segmentation task[3]. Transcribed segments can show various expression levels. For example, the three simulated segments in Figure 5.1 would require three separate hidden states, e.g., "lowly expressed", "expressed" and "highly expressed", and these may only be a small subset of observed transcription rates. Furthermore, alternative splicing, mRNA degradation, and other transcript structures lead to multiple transcribed segments of different levels within the same transcript. Such signal patterns are difficult to interpret with HMMs.

The structural change model (page 124f.) can be seen as a continuous-state model, in which every segment's average intensity would make up a "hidden" state, and the number of states does not need to be fixed in advance (although there is an implicit upper bound, namely the parameter S, the

[3] The same issue arises when HMMs are used for finding ChIP-enriched regions in ChIP-chip data (see Section 3.1.1).

number of segments).

Faster, comparable segmentation algorithms Dynamic programming algorithms tend to be comparatively slow due to the use of nested loops, and the segmentation algorithm of package *tilingArray* is no exception to that rule. Faster segmentation algorithms that are also based on piece-wise constant data models have recently been suggested for segmenting array-CGH data [156, 157]. As both the resolution of tiling microarrays and the number of samples in the individual studies are growing, these faster algorithms are worth considering as methods for segmenting the data.

5.1.5 Visualisation of segments

tilingArray offers extensive functionality for visualising tiling microarray data. The main visualisation function, `plotAlongChrom`, allows for concise visualisation of a genomic region, displaying

- the normalised levels of reporters matching genome positions in this region, separately for both strands if the data are strand-specific
- the fitted segment boundaries within this region
- genome features, such as ORFs, ncRNAs, and transcription factor binding sites, that are annotated in this region (a `data.frame` holding annotated genome features needs to be provided to the function).

Figure 5.2 shows an example visualisation of the Poly-A RNA data from our study of the *S. cerevisiae* transcriptome (see Section 4.2).

5.1.6 Discussion

The Bioconductor package *tilingArray* provides functionalities useful for the analysis of high-density tiling microarray data (such as those generated on Affymetrix whole-genome GeneChip microarrays). The package

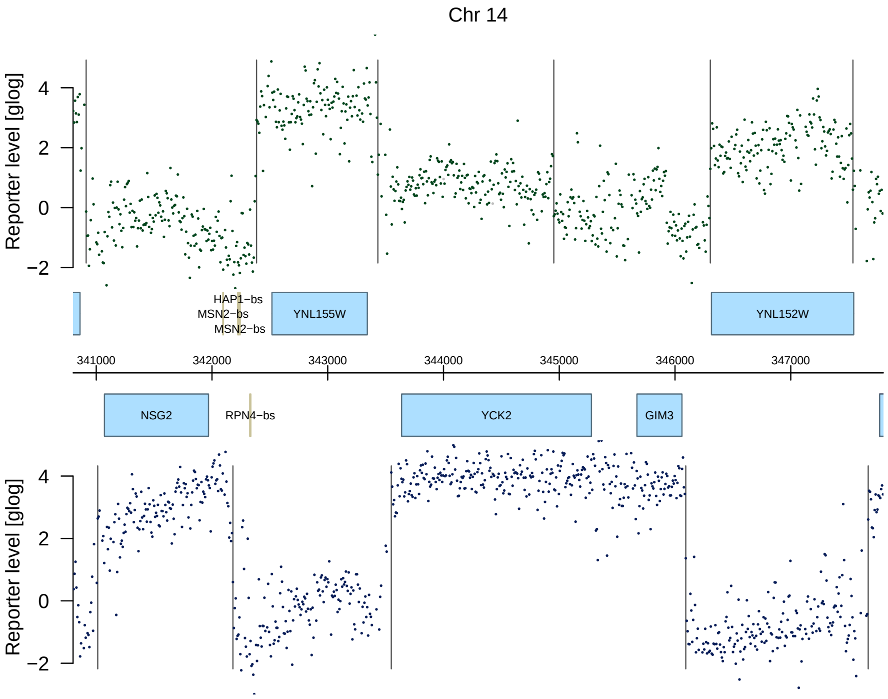


Figure 5.2: *Visualisation of the Poly-A-RNA data from our tiling microarray study of the S. cerevisiae transcriptome (Section 4.2) in a 7 kb region of chromosome 14. The figure was created using the function* `plotAlongChrom` *from the Bioconductor package tilingArray. The green dots in the top panel correspond to the preprocessed reporter intensities on the Watson strand, the blue dots in the bottom panel are the intensities from the Crick strand. The middle panel shows the genomic coordinate and genome features that are annotated in this region. Blue boxes are annotated ORFs; the golden vertical bars are experimentally determined transcription factor binding sites. One can see an operon-like transcript involving the ORFs* GIM3 *and* YCK2 *on the Crick strand.*

was developed during a survey of the complete transcriptome of *S. cerevisiae* in exponential growth phase and includes specific functionality for this objective. Other functionalities of the package, however, are sufficiently generic to be useful in other kinds of tiling microarray studies. *tilingArray* makes use of generic Bioconductor classes, which in combination with Bioconductor's open development philosophy makes the package amenable to being adopted and modified by bioinformaticians with related research objectives. We are continuing to maintain and enhance the package as part of the Bioconductor project.

Contributions to *tilingArray*

The programming source code and documentation that is contained in the package was written by Wolfgang Huber and me, with contributions from Matt Ritchie.

5.2 *Ringo*

5.2.1 Introduction

ChIP-chip, chromatin immunoprecipitation combined with microarray hybridisation, is a widely used assay for protein-DNA interactions and chromatin plasticity. The experimental procedure is described in Section 1.4.1 (page 13f.).

The interpretation of ChIP-chip data poses two computational challenges:

- primary statistical analysis, which is used as a grouping term for quality assessment, data normalisation and transformation, and identification of genomic regions of interest
- integrative bioinformatic analysis, during which the data are interpreted in context of existing genome annotation and related experimental results.

Both tasks rely on visualisation for exploring the data as well as to present the analysis results. For the primary statistical analysis, some degree of standardisation is possible and desirable. Commonly used experimental designs and microarray platforms allow the development of standard workflows and statistical procedures. Most software available for ChIP-chip data analysis can be employed in such standardised experiments [158, 159, 160, 161, 147, 162]. However, the primary analysis steps frequently need to be adapted to specific experiments. The analysis software therefore should offer flexibility in the choice of algorithms for normalisation, visualisation and identification of ChIP-enriched regions. Yet a greater degree of flexibility is required for the second task, integrative bioinformatic analysis, as the data sets, analysis questions and the applicable methods are diverse. A programming environment such as R and Bioconductor offers appropriate flexibility for both tasks.

Ringo is an open-source R/Bioconductor software package for importing raw microarray data, quality assessment, normalisation, visualisation, and for the detection and quantitative assessment of ChIP-enriched regions. The package's basic functionality covers the complete primary statistical analysis for ChIP-chip tiling microarrays. *Ringo* contains data import functions for reading in two-colour tiling microarrays manufactured by NimbleGen Systems [163]. Due to the modular design of *Ringo*, however, data from other microarray platforms can also be processed, once imported using custom functions or functions from other packages. The firm integration of the package with Bioconductor simplifies the construction of sophisticated analysis workflows, which can involve other R and Bioconductor packages.

Ringo is complementary to existing available software for ChIP microarray analysis. For example, `mpeak` [162], `TiMAT` (`http://bdtnp.lbl.gov/TiMAT`), `MAT` [160], `TileMap` [159], `ACME` [101], `HGMM` [161], and `ChIPOTle` [158] provide model-based and non-parametric algorithms for finding ChIP-enriched regions on normalised and quality controlled ChIP-chip data. These softwares commonly provide interfaces to these algorithms, and users may be required to use the softwares in combination with other tools for data import, data preprocessing and follow-up analysis. A unique aspect of *Ringo* is that it facilitates the construction of automated, programmed workflows and offers benefits in the scalability and reproducibility of the analysis.

5.2.2 Implementation

Ringo is an extension package for the statistical environment R. Most of its functionality is implemented in the R programming language, while `C` and `C++` functions are used for performance-critical computations. The package interfaces functions from other Bioconductor packages, most notably from the package *limma* [164]. The object classes used in *Ringo* are standard Bioconductor classes, such as `RGList` and `ExpressionSet`. In addition, the package provides new S4 object classes for representing identified ChIP-enriched regions and reporter-to-genome mappings.

5.2.3 Functionality

Figure 5.3 shows a typical workflow of ChIP-chip data analysis and indicates which steps are facilitated by the Bioconductor package *Ringo*. Key functionalities of the package are import, quality assessment and preprocessing of the raw data, visualisation of raw and processed data and an algorithm for detecting ChIP-enriched regions.

Data I demonstrate the functionalities of the package on a ChIP-chip data set of transcription factor binding events in the cardiomyocyte cell line HL1. The data were generated on 60mer oligonucleotide microarrays that were manufactured by NimbleGen Systems. See Chapter 3 (Section 3.2) for further details about the data.

Data import *Ringo* contains functions to read in raw data in NimbleGen file formats, generated when the microarrays are scanned, into an `RGList` object. Users can alternatively supply raw ChIP microarray data in the `RGList` format. The package *limma*, for example, contains a function that can read in most scanner file formats into an `RGList`. Such an object is essentially a `list` class object and contains the raw intensities of the two hybridisations for the Cy5 and Cy3 channel plus information on the reporters on the array and on the analysed samples.

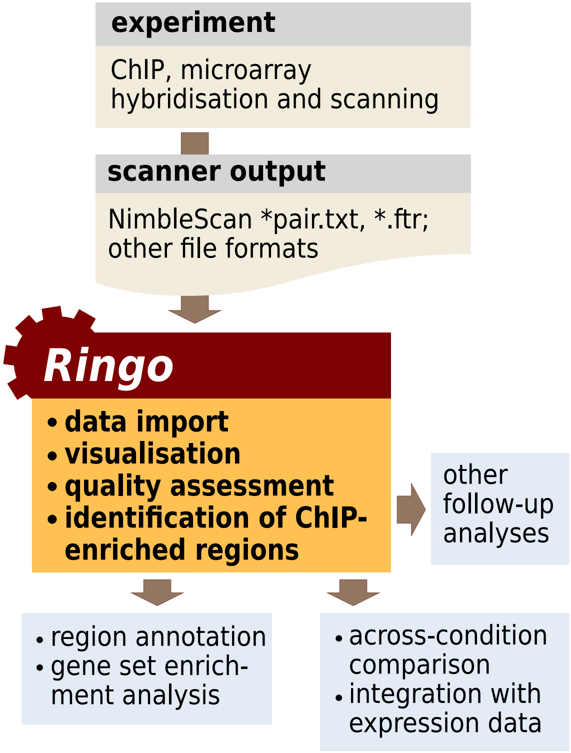

experiment
ChIP, microarray hybridisation and scanning
scanner output
NimbleScan *pair.txt, *.ftr; other file formats
Ringo
• data import
• visualisation
• quality assessment
• identification of ChIP-enriched regions
other follow-up analyses
• region annotation
• gene set enrichment analysis
• across-condition comparison
• integration with expression data

Figure 5.3: *Workflow diagram displaying which steps of the analysis of ChIP-chip experiments are facilitated by the Bioconductor package Ringo.*

Quality assessment *Ringo* contains an extensive set of functions for quality assessment of the data. Its `image` function allows the user to examine the spatial distribution of the feature intensities on the array surface. This can be useful to detect obvious artifacts on the array, such as scratches, bright spots, finger prints etc. that might render parts or all of the readouts invalid (see Figure 3.2 on page 56 for an example).

The *autocorrelation* plot can assist in the assessment of how the reporter tiling across the chromosome affects the levels of reporters. For each base-pair lag d, the correlation coefficient between the intensities of reporters at genomic positions $x + d$ and the reporter intensities at positions x is computed. The correlation coefficient is plotted against the lag d in the autocorrelation plot (see Figure 5.4). Some degree of autocorrelation is to be expected, since the DNA fragments that are hybridised to the microar-

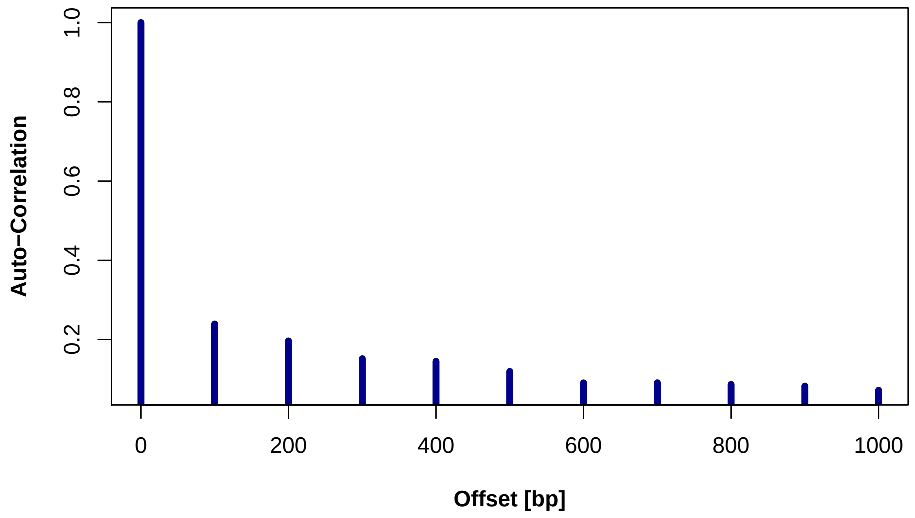


Figure 5.4: *This is an example autocorrelation plot, which was computed and plotted using functions in Ringo. The autocorrelation decreases with increasing lag.*

ray are typically some hundred base pairs long. The visualisation is useful to assess whether the autocorrelation is unusually high or low and up to which distance this is the case.

Furthermore, if the data set contains biological or technical replicates, low correlation between replicate sample intensities may indicate microarrays of questionable quality. *Ringo* therefore contains the function `corPlot` to visualise the correlation between replicate samples' raw and preprocessed reporter intensities.

Normalisation Following quality assessment, microarray data are normalised to increase the signal-to-noise ratio of the data. Then fold changes of normalised reporters intensities of the enriched samples divided by the normalised intensities of the *input* samples are derived. The (generalised) logarithm of these ratios yields the preprocessed reporter levels.

Ringo provides a number of choices for normalisation of ChIP-chip data, interfacing preprocessing methods implemented in the Bioconductor packages *vsn* [25] and *limma* plus the Tukey-biweight scaling of the log-

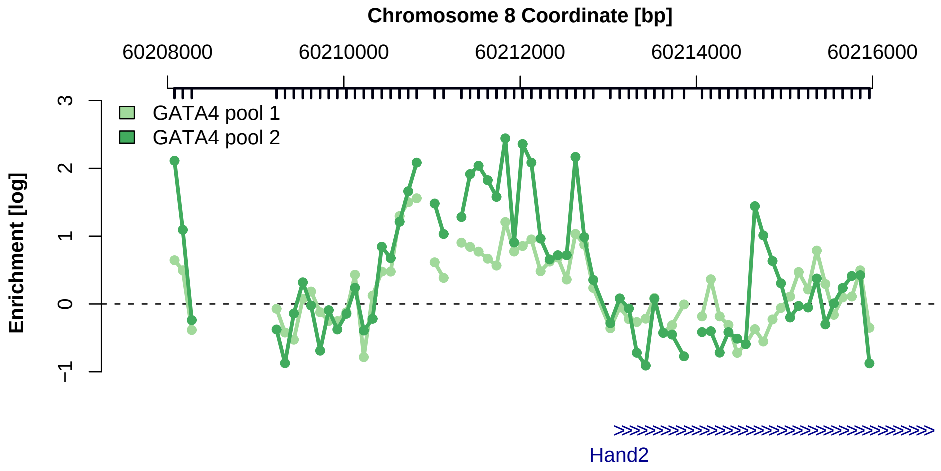


Figure 5.5: *Normalised* GATA4 *ChIP-chip reporter levels around the TSS of the gene* Hand2*. The green lines correspond to the preprocessed levels in the two replicated ChIP samples for the transcription factor* GATA4*. The ticks below the genomic coordinate axis on top indicate genomic positions matched by reporters on the microarray. The blue arrows on the bottom mark the gene* Hand2 *with the arrow direction indicating its transcription direction, i.e. the gene is located on the Watson strand.*

ratios that is suggested by NimbleGen. The preprocessing functions return reporter levels on a $\log_2$ (or glog) scale in an object of class `ExpressionSet`, the basic Bioconductor object class for microarray data.

Reporter mapping A mapping between the identifier of each reporter on the microarray and the genomic positions at which the reporter's sequence matches the genome sequence is required for ChIP-chip analysis. *Ringo* implements a custom S4 class `probeAnno` to store this mapping. The `probeAnno` object corresponds to a set of tables relating chromosomal positions to feature identifiers on the array. The package provides functions that assist in the production of the `probeAnno` object from reporter-to-genome mappings supplied by the array manufacturer or from custom alignments of the reporter sequences to the genome.

Visualising genomic regions An important aspect of genomic data analysis is extensive exploration of the data using different visualisation techniques. In addition to the visualisation functions offered by other R and

Bioconductor packages, *Ringo* provides a function for displaying estimates of reporter-wise log fold enrichment in specified genomic regions. See Figure 5.5 for an example. The plot displays the preprocessed reporter levels, the positions of reporter matches to the genome and genes that are annotated in the specified region.

Smoothing of reporter levels The identification of ChIP-enriched genomic regions, i.e. genomic regions that show enrichment in the immunoprecipitated sample as compared to the untreated input sample, is a key step in ChIP-chip data analysis workflows. A smoothing of reporter levels is suggested to precede the identification of ChIP-enriched regions. Different reporters measure the same target DNA amount with different efficiency. This effect is caused by variable quality of feature synthesis on the array, reporter GC content, target cDNA secondary structure, cross-hybridisation, and other reasons. One way to ameliorate these reporter effects as well as the stochastic noise is to perform a smoothing over individual reporter levels. A window of fixed width is slided along the chromosome, and the reporter level at genomic position x_0 is replaced by the median over the levels of all reporters inside the window that is centred at x_0. Factors to take into account when choosing the width of the sliding window are the size distribution of DNA fragments after sonication, which influences the autocorrelation between reporters (see Figure 5.4), and the spacing between reporter matches on the genome. Note that while sliding-window smoothing potentially allows for clearer identification of ChIP-enriched regions, it will introduce a bias in estimates of the start and end points of such regions (see Section 5.1.4). In the smoothing step, the reporter levels from replicate samples can be combined into a single smoothed sample for ChIP-enrichment of the analysed transcription factor or histone modification. Figure 5.6 shows an example smoothing result. The smoothed reporter levels indicate ChIP enrichment in the displayed genomic region with less variation in the reporter levels than in the two unsmoothed samples. The two replicate samples have been combined into a single smoothed ChIP sample for GATA4 enrichment.

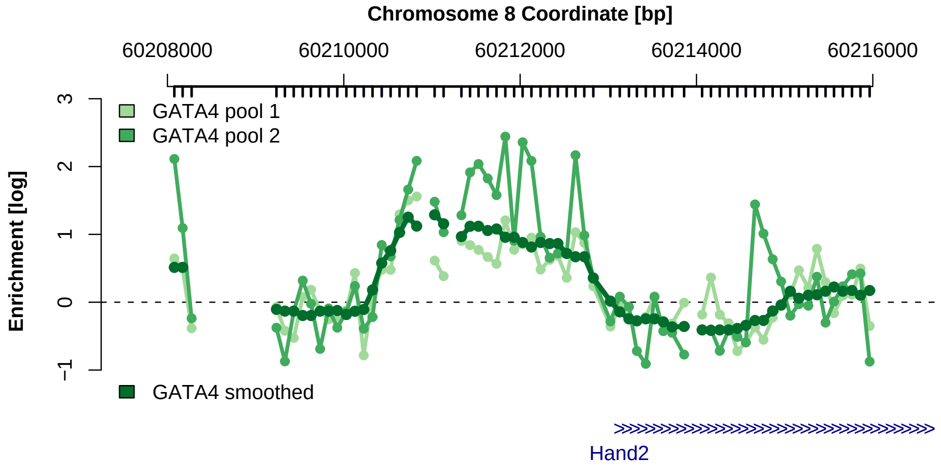


Figure 5.6: *Normalised and smoothed reporter levels around the TSS of the gene Hand2. The pale green lines correspond to the preprocessed levels in the two replicated ChIP samples for the transcription factor* GATA4*. The dark green line is the result of the sliding window median smoothing across the two replicates. For details on the figure annotation, see the legend of Figure 5.5.*

Finding ChIP-enriched regions The aim is to determine a set of genomic regions that appear to be antibody-enriched. Regions can be ranked by a quantitative score of the confidence in the enrichment, or by a score of the degree of enrichment in each region. A *p*-value is a particular case of confidence score and is defined in the context of an appropriate null hypothesis and a probability model. If the simple goal is to find and rank regions in some way that can be reasonably calibrated, such complications are not necessarily required. The confidence in an enrichment should be distinguished from the degree of the enrichment. More profound enrichments tend to result in stronger signals and hence less ambiguous calls. However, the certainty about whether there is enrichment should also be affected by reporter coverage, reporter sequence, and cross-hybridisation.

Which approach is best for identifying enriched regions from ChIP-chip data depends on the microarray design, on the biological context of the experiments, and on the way the regions are going to be used in follow-up analyses. *Ringo* implements one possible approach. For a region to be called *enriched*, the following is required

- the smoothed reporter levels (log ratios) within that region all exceed a certain threshold y_0
- the region contains at least $n_{\min}$ positions matched by reporters
- there are no gaps larger than $d_{\max}$ base pairs between two consecutive match positions in the region.

The minimum number of reporters rule ($n_{\min}$) might seem redundant with the sliding window smoothing, but it becomes important in reporter-sparse regions. For example, if there is only one enriched reporter within a certain genomic 1kb region and no other reporters were mapped to that region, this single reporter arguably does not provide enough evidence for calling this genomic region enriched. Making calls supported only by a few reporters should be avoided. Finally, the $d_{\max}$ rule prevents calling disconnected regions as a single enriched region.

The optimal approach for setting the enrichment threshold y_0 would be to tune it by considering sets of positive and negative control regions. As such control regions are rarely available, a mixture modelling approach is considered.

The distribution of the smoothed reporter levels y is assumed to be a mixture of two underlying distributions. One is the null distribution $\mathcal{L}_0$ of reporter levels in non-enriched regions; the other is the alternative distribution $\mathcal{L}_{\text{alt}}$ of the levels in enriched regions. The challenge is to estimate the null distribution $\mathcal{L}_0$. In *Ringo*, an estimate $\widehat{\mathcal{L}_0}$ is derived based on the empirical distribution of smoothed reporter levels (see Figure 5.7). The null distribution $\mathcal{L}_0$ is assumed to have most of its mass close to its mode m_0, which is close to $y = 0$. $\mathcal{L}_0$ is also assumed to be symmetric about m_0. The alternative distribution $\mathcal{L}_{\text{alt}}$ is assumed to be stochastically larger than $\mathcal{L}_0$ and to contain negligible mass for $y < m_0$. Based on these assumptions, the following estimator of $\mathcal{L}_0$ is derived. First, the position of the mode m_0 is estimated by the midpoint of the shorth of the empirical distribution of those y that fall into the interval $[-1, 1]$ (on a $\log_2$ scale). The distribution $\mathcal{L}_0$ is then estimated from the empirical distribution of $m_0 - |y - m_0|$, i.e.

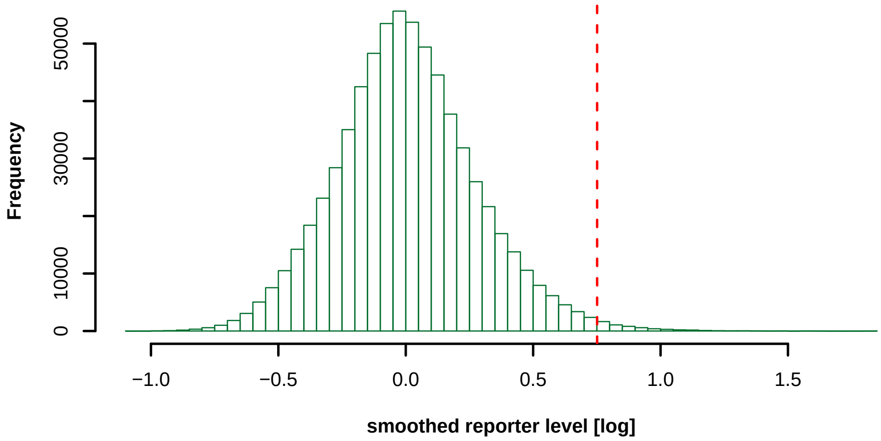


Figure 5.7: *This histogram shows the distribution of smoothed reporter levels for the* GATA4 *ChIP sample. The red dashed line indicates the algorithmically determined threshold y_0, above which smoothed reporter levels are considered to indicate enrichment. See the text for details about the algorithm.*

by reflecting $y < m_0$ onto $y > m_0$.

From the estimated null distribution, an enrichment threshold y_0 can be determined, for example the 99.9% quantile. The red dashed, vertical line in Figure 5.7 denotes one such estimated threshold. Antibodies vary in their efficiency to bind to their target epitope, and the noise level in the data depends on the sample DNA. Thus, y_0 should be computed separately for each antibody and cell type, as the null and alternative distributions, $\mathcal{L}_0$ and $\mathcal{L}_{\text{alt}}$, may vary.

The algorithm described above provides a straightforward estimate for y_0. It has been used in previous studies, for example by Schwartz *et al.* [165]. There are other algorithms that use more complex models of ChIP-chip data [94, 166].

5.2.4 Discussion

The functionality of the software package *Ringo* provides a good starting point for researchers interested in the analysis of ChIP tiling microarrays or of similar data. *Ringo* provides a comprehensive set of functions for quality assessment, data processing, visualisation and ChIP-chip data analysis. The package's close integration with Bioconductor opens up diverse possibilities for subsequent analysis.

In addition to the computational challenges in ChIP-chip data analysis, users need to be aware of experimental design issues, which can lead to false positive and false negative enriched regions. Such experimental issues include lack of antibody specificity or sensitivity, and cross-hybridisation. Although good software can help in identifying them, these issues are the main reason why ChIP-chip, as with most high-throughput approaches, should primarily be seen as a means to generate biological hypotheses that need to be validated in appropriate small-scale studies. Bioconductor provides a integrative framework for the formulation of well-stated hypotheses.

Contributions to *Ringo*

I have written most of the source code included in the package, with contributions from Oleg Sklyar, Tammo Krueger, Matt Ritchie and Wolfgang Huber. I have written the package documentation and most of the manuscript describing the package [147], with contributions from Wolfgang Huber and Oleg Sklyar. Jenny J. Fischer and Silke Sperling provided the ChIP-chip data displayed in this chapter and in the publication [147]. I am still maintaining and enhancing the Bioconductor package *Ringo*.

Conclusions

High-throughput studies are powerful tools of the genomic era for further enhancing our knowledge about the fundamental processes that cells employ to utilise the information that is encoded in their DNA. During my time at EMBL-EBI, I had the opportunity to work on three large-scale studies, in which the processes of transcription and transcriptional regulation were investigated in model organisms, using high-throughput technologies such as microarrays. This dissertation gives a comprehensive description of these studies, including their biological and technological background, the development of methods for analysing the data from such studies, as well as the characteristic problems one needs to be aware of during the analysis of the data and the interpretation of the study results. As such, this document may give an impression of the potential of such studies as well as of the challenges that are constantly being faced in current genomics research. In this dissertation, I have also emphasised the fact that biological, statistical, as well as computational aspects need to be considered to obtain useful results out of large-scale genomics studies.

In conclusion, the obtained results exemplify the potential of such high-throughput studies towards the formulation of well-defined hypotheses about transcription and transcriptional regulation in eukaryotic cells.

By now (August 2008), it looks as if the high-throughput sequencing technologies are about to supplant microarrays as the preferred means of investigating transcriptomes and transcriptional regulation in the near future. Even though these sequencing technologies are still immature and only preliminary algorithms for analysing their output are available, these technologies promise to overcome typical problems which are inherent to microarrays, such as reporter-specific noise effects. In addition, in contrast to microarrays, high-throughput sequencing may also allow for reliable quantification of RNAs with few copy numbers, whose signals are difficult to distinguish from background signal on microarrays. However, I expect that microarrays will continue to be used for a few more years at least, because microarrays will likely become more affordable, their output is reasonably well understood and free powerful analysis tools, such as Bioconductor, can be used for their evaluation. Algorithms for the analysis of high-throughput microarray data can be adopted to handle new high-throughput sequencing data. And the next step, integrating the results from an high-throughput experiment with existing annotation and other experimental results, is similar for both microarrays and sequencing studies. Thus, methods, algorithms and software for the analysis of microarray experiments are not going to be dispensable from one day to the other, but rather are being adopted by the scientific community to meet the challenges of these new sequencing technologies in much-needed further studies for investigating transcription and transcriptional regulation.

Appendix A

Publications

These are the publications that I contributed to while researching for this dissertation:

- David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. [9]
- Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. [148]
- Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W, Sperling S (2007) Genomic Organization of Transcriptomes in Mammals: Co-regulation and Co-functionality. [87]
- Toedling J, Sklyar O, Krueger T, Fischer JJ, Sperling S, Huber W (2007) Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. [147]
- Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, Sperling S (2008) Combinatorial Effects of Histone Modifications in Transcription and Differentiation. [88]
- Toedling J, Huber W (2008) Analysis of ChIP-chip data using Bioconductor. [167]

Bibliography

[1] Glass CK, Rosenfeld MG (2008) Transcriptional regulatory machinery and epigenetics at a crossroads. Current Opinion in Cell Biology 20:249–252.

[2] Knippers R (2001) Molekulare Genetik. Georg Thieme Verlag, 8th edition.

[3] Zylber EA, Penman S (1971) Products of RNA polymerases in HeLa cell nuclei. Proceedings of the National Academy of Sciences of the USA 68:2861–2865.

[4] Ro-Choi TS, Raj NB, Pike LM, Busch H (1976) Effects of alpha-amanitin, cycloheximide, and thioacetamide on low molecular weight nuclear RNA. Biochemistry 15:3823–3828.

[5] Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. The EMBO Journal 23:4051–4060.

[6] Willis IM (1993) RNA polymerase III. Genes, factors and transcriptional specificity. European Journal of Biochemistry / FEBS 212:1–11.

[7] Crick FHC (1958) On protein synthesis. Symposia of the Society for Experimental Biology 12:138–163.

[8] Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242–2246.

[9] David L, Huber W, Granovskaia M, Toedling J, Palm CJ, et al. (2006) A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the USA 103:5320–5325.

[10] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349.

[11] Carninci P, Yasuda J, Hayashizaki Y (2008) Multifaceted mammalian transcriptome. Current Opinion in Cell Biology 20:274–280.

[12] Dorigo B, Schalch T, Bystricky K, Richmond TJ (2003) Chromatin fiber folding: requirement for the histone H4 N-terminal tail. Journal of Molecular Biology 327:85–96.

[13] Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705.

[14] Hebbes TR, Thorne AW, Crane-Robinson C (1988) A direct link between core histone acetylation and transcriptionally active chromatin. The EMBO Journal 7:1395–1402.

[15] Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, et al. (2002) Active genes are tri-methylated at K4 of histone H3. Nature 419:407–411.

[16] Grunstein M (1997) Histone acetylation in chromatin structure and transcription. Nature 389:349–52.

[17] Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403:41–45.

[18] Martin C, Zhang Y (2007) Mechanisms of epigenetic inheritance. Current Opinion in Cell Biology 19:266–272.

[19] Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT (2008) The RNA polymerase II core promoter - the gateway to transcription. Current Opinion in Cell Biology 20:253–259.

[20] Kruglyak S, Tang H (2000) Regulation of adjacent yeast genes. Trends in Genetics 16:109–111.

[21] Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, et al. (2004) An abundance of bidirectional promoters in the human genome. Genome Research 14:62–66.

[22] Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470.

[23] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14:1675–1680.

[24] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

[25] Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1:S96–S104.

[26] Dudoit S, Yang YH, Speed TP, Callow MJ (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12:111–139.

[27] Smyth GK, Michaud J, Scott HS (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21:2067–2075.

[28] Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends in Genetics 21:93–102.

[29] Naef F, Magnasco MO (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Physical Review E 68:011906.

[30] Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, et al. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends in Genetics 21:466–475.

[31] Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 122:33–43.

[32] Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, et al. (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proceedings of the National Academy of Sciences of the USA 101:17765–17770.

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

[34] Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An overview of Ensembl. Genome Research 14:925–928. URL `http://www.ensembl.org`.

[35] Okoniewski MJ, Hey Y, Pepper SD, Miller CJ (2007) High correspondence between Affymetrix exon and standard expression arrays. Biotechniques 42:181–185.

[36] Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, et al. (2004) A novel, high-performance random array platform for quantitative gene expression profiling. Genome Research 14:2347–2356.

[37] Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Research 33:5914–5923.

[38] Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F (2004) A model based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99:909–917.

[39] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306–2309.

[40] Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 309:626–630.

[41] Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U, et al. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Research 17:691–707.

[42] Hartwell LH, Unger MW (1977) Unequal division in *Saccharomyces cerevisiae* and its implications for the control of cell division. The Journal of Cell Biology 75:422–435.

[43] Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science 261:1551–1557.

[44] Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ (1999) Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein. Molecular and Cellular Biology 19:5267–5278.

[45] Pramila T, Wu W, Miles S, Noble WS, Breeden LL (2006) The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. Genes & Development 20:2266–2278.

[46] Kumar R, Reynolds DM, Shevchenko A, Shevchenko A, Goldstone SD, et al. (2000) Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. Current Biology 10:896–906.

[47] Mai B, Miles S, Breeden LL (2002) Characterization of the ECB binding complex responsible for the M/G(1)-specific transcription of CLN3 and SWI4. Molecular and Cellular Biology 22:430–441.

[48] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 9:3273–3297.

[49] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2:65–73.

[50] Eisen M, Spellman P, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the USA 95:14863–14868.

[51] de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, et al. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics 21:1164–1171.

[52] Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 8:995–1005.

[53] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25:25–29.

[54] Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genetics 26:183–186.

[55] Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, et al. (2002) A global analysis of *Caenorhabditis elegans* operons. Nature 417:851–854.

[56] Roy PJ, Stuart JM, Lund J, Kim SK (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature 418:975–979.

[57] Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. Journal of Biology 1:5.

[58] Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291:1289–1292.

[59] Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, et al. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Research 13:1998–2004.

[60] Schübeler D, Francastel C, Cimbora D, Reik A, Martin D, et al. (2000) Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. Genes & Development 14:940–950.

[61] Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113:405–417.

[62] Brayton KA, Chen Z, Zhou G, Nagy PL, Gavalas A, et al. (1994) Two genes for de novo purine nucleotide synthesis on human chromosome 4 are closely linked and divergently transcribed. The Journal of Biological Chemistry 269:5313–5321.

[63] Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genetics 31:180–183.

[64] Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nature Reviews Genetics 5:299–310.

[65] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the USA 101:6062–6067.

[66] Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Molecular Biology and Evolution 22:767–775.

[67] Hardin J, Mitani A, Hicks L, VanKoten B (2007) A robust measure of correlation between two genes on a microarray. BMC Bioinformatics 8:220.

[68] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, et al. (2002) Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the USA 99:4465–4470.

[69] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

[70] Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420:563–573.

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

[72] Durinck S, Moreau Y, Kasprzyk A, Davis S, Moor BD, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–3440.

[73] Hamming RW (1986) Coding and Information Theory. Prentice-Hall, Inc.

[74] Marenholz I, Heizmann CW, Fritz G (2004) S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). Biochemical and biophysical research communications 322:1111–1122.

[75] van Driel R, Fransz PF, Verschure PJ (2003) The eukaryotic genome: a system regulated at different hierarchical levels. Journal of Cell Science 116:4067–4075.

[76] Akashi K, He X, Chen J, Iwasaki H, Niu C, et al. (2003) Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis. Blood 101:383–389.

[77] Wolffe AP, Matzke MA (1999) Epigenetics: regulation through repression. Science 286:481–486.

[78] Képès F (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. Journal of Molecular Biology 329:859–865.

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

[80] Johnnidis JB, Venanzi ES, Taxman DJ, Ting JPY, Benoist CO, et al. (2005) Chromosomal clustering of genes controlled by the aire transcription factor. Proceedings of the National Academy of Sciences of the USA 102:7233–7238.

[81] Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature Genetics 39:730–732.

[82] Lee JM, Sonnhammer ELL (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Research 13:875–882.

[83] Teichmann SA, Veitia RA (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. Genetics 167:2121–2125.

[84] Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Research 14:1085–1094.

[85] Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Research 14:1870–1879.

[86] Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

[87] Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, et al. (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics 89:580–587.

[88] Fischer JJ, Toedling J, Krueger T, Schueler M, Huber W, et al. (2008) Combinatorial effects of four histone modifications in transcription and differentiation. Genomics 91:41–51.

[89] Agalioti T, Chen G, Thanos D (2002) Deciphering the transcriptional histone acetylation code for a human gene. Cell 111:381–392.

[90] Kurdistani SK, Grunstein M (2003) Histone acetylation and deacetylation in yeast. Nature Reviews Molecular Cell Biology 4:276–284.

[91] Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in *S. cerevisiae*. PLoS Biology 3:e328.

[92] Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125:315–326.

[93] Ng HH, Robert F, Young RA, Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Molecular Cell 11:709–719.

[94] Bourgon RW (2006) Chromatin-immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a ”gold standard”. Ph.D. thesis, University of California, Berkley,

USA. URL http://www.ebi.ac.uk/~bourgon/papers/bourgon_dissertation_public.pdf.

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

[96] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57:289–300.

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

[98] Rabiner LR, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Magazine 3:4–16.

[99] Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 21 Supplement 1:i274–i282.

[100] Olson EN (2006) Gene regulatory networks in the evolution and development of the heart. Science 313:1922–1927.

[101] Scacheri PC, Crawford GE, Davis S (2006) Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. Methods in Enzymology 411:270–282.

[102] Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Research 12:656–664.

[103] Royce TE, Rozowsky JS, Gerstein MB (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics 23:988–997.

[104] Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5:R80.

[105] Srivastava D, Thomas T, Lin Q, Kirby ML, Brown D, et al. (1997) Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. Nature Genetics 16:154–160.

[106] Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436:876–880.

[107] Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62.

[108] Mata J, Marguerat S, Bähler J (2005) Post-transcriptional control of gene expression: a genome-wide perspective. Trends in Biochemical Sciences 30:506–514.

[109] Chuikov S, Kurash JK, Wilson JR, Xiao B, Justin N, et al. (2004) Regulation of p53 activity through lysine methylation. Nature 432:353–360.

[110] Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, et al. (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature 442:86–90.

[111] Shi X, Hong T, Walter KL, Ewalt M, Michishita E, et al. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature 442:96–99.

[112] Mito Y, Henikoff JG, Henikoff S (2005) Genome-scale profiling of histone H3.3 replacement patterns. Nature Genetics 37:1090–1097.

[113] Ng RK, Gurdon JB (2008) Epigenetic inheritance of cell differentiation status. Cell Cycle 7:1173–1177.

[114] Niu Z, Yu W, Zhang SX, Barron M, Belaguli NS, et al. (2005) Conditional mutagenesis of the murine serum response factor gene blocks cardiogenesis and the transcription of downstream gene targets. The Journal of Biological Chemistry 280:32531–32538.

[115] Shalizi A, Gaudilliere B, Yuan Z, Stegmüller J, Shirogane T, et al. (2006) A calcium-regulated MEF2 sumoylation switch controls postsynaptic differentiation. Science 311:1012–1017.

[116] Jay PY, Harris BS, Maguire CT, Buerger A, Wakimoto H, et al. (2004) Nkx2-5 mutation causes anatomic hypoplasia of the cardiac conduction system. Journal of Clinical Investigation 113:1130–1137.

[117] Arceci RJ, King AA, Simon MC, Orkin SH, Wilson DB (1993) Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. Molecular and Cellular Biology 13:2235–2246.

[118] Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Research 24:238–241.

[119] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15:1034–1050.

[120] Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, et al. (2005) Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer 44:305–319.

[121] Barrera LO, Ren B (2006) The transcriptional regulatory code of eukaryotic cells–Insights from genome-wide analysis of chromatin organization and transcription factor binding. Current Opinion in Cell Biology 18:291–298.

[122] Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29:1165–1188.

[123] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104.

[124] Bofkin LNM (2006) The Causes and Consequences of Variation in Evolutionary Processes Acting on DNA Sequences. Ph.D. thesis, University of Cambridge, UK. URL `http://www.embl-ebi.ac.uk/training/ftp/PhDtheses/LeeBofkinThesis.pdf`.

[125] Sylvestre J, Margeot A, Jacq C, Dujardin G, Corral-Debrinski M (2003) The role of the 3′ untranslated region in mRNA sorting to the vicinity of mitochondria is conserved from yeast to human cells. Molecular Biology of the Cell 14:3848–3856.

[126] Jackson JS, Houshmandi SS, Leban FL, Olivas WM (2004) Recruitment of the Puf3 protein to its mRNA target for regulation of mRNA decay in yeast. RNA 10:1625–1636.

[127] Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nature Reviews Genetics 8:413–423.

[128] Aravind L, Watanabe H, Lipman DJ, Koonin EV (2000) Lineage-specific loss and divergence of functionally linked genes in eurkaryotes. Proceedings of the National Academy of Sciences of the USA 97:11319–11324.

[129] Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Research 35:e128.

[130] Ahdesmäki M, Lähdesmäki H, Pearson R, Huttunen H, Yli-Harja O (2005) Robust detection of periodic time series measured from biological systems. BMC Bioinformatics 6:117.

[131] Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the USA 98:5116–5121.

[132] MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. BMC Bioinformatics 7:113.

[133] Jeffery IB, Madden SF, McGettigan PA, Perriere G, Culhane AC, et al. (2007) Integrating transcription factor binding site information with gene expression datasets. Bioinformatics 23:298–305.

[134] Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. Journal of Molecular Biology 296:1205–1214.

[135] Alexa A, Rahnenführer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22:1600–1607.

[136] Gordon DB, Nekludova L, McCallum S, Fraenkel E (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics 21:3164–3165.

[137] Hereford LM, Osley MA, Ludwig TR, McLaughlin CS (1981) Cell-cycle regulation of yeast histone mRNA. Cell 24:367–375.

[138] Hermsen R, ten Wolde PR, Teichmann S (2008) Chance and necessity in chromosomal gene distributions. Trends in Genetics 24:216–219.

[139] Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, et al. (2007) Transcription factor binding and modified histones in human bidirectional promoters. Genome Research 17:818–827.

[140] Ishida C, Aranda C, Valenzuela L, Riego L, Deluna A, et al. (2006) The UGA3-GLT1 intergenic region constitutes a promoter whose bidirectional nature is determined by chromatin organization in *Saccharomyces cerevisiae*. Molecular Microbiology 59:1790–1806.

[141] Zhang X, Smith TF (1998) Yeast ”operons”. Microbial & Comparative Genomics 3:133–140.

[142] Whitehouse I, Rando OJ, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription. Nature 450:1031–1035.

[143] Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research 18:1073–1083.

[144] Xie Y, Varshavsky A (2001) RPN4 is a ligand, substrate, and transcriptional regulator of the 26S proteasome: a negative feedback circuit. Proceedings of the National Academy of Sciences of the USA 98:3056–3061.

[145] Gentleman R, Ihaka R (1996) R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299–314. URL http://www.r-project.org.

[146] Miller CJ, Attwood TK (2003) Bioinformatics goes back to the future. Nature Reviews Molecular Cell Biology 4:157–162.

[147] Toedling J, Sklyar O, Krueger T, Fischer JJ, Sperling S, et al. (2007) Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics 8:221.

[148] Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics 22:1963–1970.

[149] Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

[150] Rennie C, Noyes HA, Kemp SJ, Hulme H, Brass A, et al. (2008) Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays. BMC Genomics 9:317.

[151] Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66:47–78.

[152] Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ (2005) A statistical approach for array CGH data analysis. BMC Bioinformatics 6:27.

[153] Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

[154] Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461–464.

[155] Marioni JC, Thorne NP, Tavaré S (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22:1144–1146.

[156] Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657–663.

[157] Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, et al. (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24:309–318.

[158] Buck MJ, Nobel AB, Lieb JD (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biology 6:R97.

[159] Ji H, Wong WH (2005) TileMap: create chromosomal map of tiling array hybridizations. Bioinformatics 21:3629–3636.

[160] Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, et al. (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proceedings of the National Academy of Sciences of the USA 103:12457–12462.

[161] Keleş S (2007) Mixture modeling for genome-wide localization of transcription factors. Biometrics 63:10–21.

[162] Zheng M, Barrera LO, Ren B, Wu YN (2007) ChIP-chip: data, model, and analysis. Biometrics 63:787–796.

[163] Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, et al. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Research 12:1749–1755.

[164] Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, editors, Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, pp. 397–420.

[165] Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. Nature Genetics 38:700–705.

[166] Kuan PF, Chun H, Keleş S (2008) CMARRT: a tool for the analysis of ChIP-chip data from tiling arrays by incorporating the correlation structure. In: Proceedings of the Pacific Symposium on Biocomputing. pp. 515–526.

[167] Toedling J, Huber W (2008) Analyzing ChIP-chip Data Using Bioconductor. PLoS Computational Biology 4:e1000227.