

Inference of gene regulatory interactions
from deep sequencing data

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Piotr Jan Balwierz
aus Polen

Basel, 2012

Original document stored on the publication server of the University of Basel
edoc.unibas.ch



This work is licenced under the agreement Attribution Non-Commercial No Derivatives –
2.5 Switzerland. The complete text may be viewed here:

creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en



Attribution-Noncommercial-No Derivative Works 2.5 Switzerland

You are free:



to Share — to copy, distribute and transmit the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Noncommercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license) available in German:
<http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de>

Disclaimer:

The Commons Deed is not a license. It is simply a handy reference for understanding the Legal Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its contents do not appear in the actual license. Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying of, or linking to this Commons Deed does not create an attorney-client relationship.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf
Antrag von

Prof. Erik van Nimwegen und PD Dr. Dirk Schübeler

Basel, den 14. Dezember 2010

Prof. Dr. Martin Spiess
Dekan

*To the memory of my physics teacher,
Dr. Jerzy Mucha*

Contents

1	Introduction	11
I	Transcriptional regulatory interactions	13
2	Introduction	15
3	Methods for analyzing deep sequencing expression data: Constructing the human and mouse promoterome with deepCAGE data	21
3.1	Background	21
3.2	Results and Discussion	23
3.2.1	Genome Mapping	23
3.2.2	Normalization	23
3.2.3	Noise model	27
3.2.4	Promoterome construction	31
3.2.4.1	Clustering adjacent co-regulated TSSs	33
3.2.4.2	Filtering background transcription	36
3.2.4.3	Proximal promoter extraction and transcription start region construction	36
3.2.5	Promoterome Statistics	37
3.2.5.1	Comparison with known transcription starts	37
3.2.5.2	Hierarchical structure of the proteome	39
3.2.5.3	Comparison with simple single-linkage clustering	40
3.2.6	High and low CpG promoters	44
3.3	Conclusions	46
3.4	Materials and Methods	48
3.4.1	CAGE and RNA-seq expression data	48
3.4.2	Normalization by fitting to a reference distribution	48
3.4.3	Noise model	49
3.4.4	Estimating the multiplicative noise component from the replicate	51

CONTENTS

3.4.5	Estimating the multiplicative noise component by comparing zero and one hour expression in the THP-1 cells PMA time course	52
3.4.6	Likelihood of the expression profile of a single promoter	53
3.4.7	Likelihood for a consecutive pair of promoters	55
3.4.8	Classifying high- and low-CpG promoters	58
3.4.9	Data availability	59
3.5	Supplementary Data	60
3.5.1	Distributions of reads per position for Solexa RNA-seq data	60
3.5.2	Replicate scatter for Solexa RNA-seq data	60
3.5.3	Per ‘exon’ replicate scatter for Solexa RNA-seq data	60
3.5.4	CAGE per TSS replicate scatter	61
3.5.5	CAGE per gene replicate scatter	61
3.5.6	Comparison with FANTOM3 clustering	61
3.5.7	Nearby uncorrelated TSSs	65
3.5.8	Mouse Promoterome Statistics	68
4	ISMARA: Modeling genomic signals as a democracy of regulatory motifs	71
4.1	Introduction	72
4.2	Results	74
4.2.1	An Integrated System for Motif Activity Response Analysis	74
4.2.2	Overview of the results presented by ISMARA	78
4.2.3	Inferring motif activity dynamics: inflammatory response	82
4.2.4	Identifying novel master regulators: Mucociliary differentiation of bronchial epithelial cells	84
4.2.5	Epithelial-Mesenchymal Transition: including microRNAs in core regulatory networks	86
4.2.6	TF activities effecting chromatin state: analysis of ChIP-seq data	88
4.3	Discussion	92
4.4	Methods	95
4.4.1	Materials	97
4.5	Supplementary Methods	98
4.5.1	Human and mouse promoteromes	98
4.5.2	A curated set of regulatory motifs	99
4.5.3	Transcription factor binding site predictions	103
4.5.4	Associating miRNA target sites with each promoter	104
4.5.5	Expression data processing	105
4.5.6	ChIP-seq data processing	107
4.5.7	Motif activity fitting.	108
4.5.7.1	Setting λ through cross-validation	110

4.5.7.2	Error bars on motif activities	110
4.5.8	Processing of replicates	111
4.5.9	Target predictions	113
4.5.9.1	Enriched Gene Ontology categories	115
4.5.10	Principal component analysis of the activities explaining chromatin mark levels	115
4.6	Fraction of variance explained by the fit	119
4.7	Overview of results presented in the web-interface	120
4.8	HNF1a activity in pancreas	132
4.9	Reproducibility of motif activities	132
4.10	Motifs dis-regulated in tumor cells	134
4.11	XBP1 motif activity and mRNA expression	136
4.12	Analysis of the ENCODE ChIP-seq data	136
4.12.1	PCA analysis	140
5	The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.	147
5.1	Introduction	148
5.2	Results	148
5.2.1	Outline of the analysis strategy	148
5.2.2	DeepCAGE quantification of dynamic TSS usage	151
5.2.3	Promoter expression	152
5.2.4	Comprehensive regulatory site prediction	152
5.2.5	Inferring key TFs and their time-dependent activities	152
5.2.6	Core transcriptional regulatory network	156
5.2.7	Validation of edge predictions	158
5.2.8	Single TF knockdowns affect multiple motif activities	159
5.2.9	Many TFs are involved in the differentiation process	162
5.2.10	Web interface to data and analysis results	163
5.3	Discussion	164
5.4	Supplementary Figures	167
5.5	Supplementary Tables	178
5.6	Supplementary Methods	179
5.6.1	DeepCAGE	179
5.6.2	CAGE tag mapping	179
5.6.3	CAGE expression normalization, noise analysis, and promoter construction	179
5.6.4	Gene assignment for CAGE promoters	179
5.6.5	deepCAGE expression Analysis	180
5.6.6	Expression signal versus replicate noise	180
5.6.7	Construction of position specific weight matrices	183

CONTENTS

5.6.8	Binding Site Predictions	185
5.6.9	Motif Activity Inference	186
5.6.10	Combining motif activities from replicates and motif FOV	188
5.6.11	Clustering motifs on activity profiles	189
5.6.12	Permutation and Cross-validation tests	190
5.6.13	Motif target predictions	191
5.6.14	siRNA edge validation and core network construction	191
5.6.15	Motif Activity Analysis of TF knockdowns	192
5.6.16	The data and analysis results available from the FANTOM4 web resource	193
6	Conclusions	195
II	Function and processing of box C/D snoRNAs	199
7	Introduction	201
8	The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing	205
8.1	Introduction	206
8.2	Results	208
8.2.1	New targets for MBII-52	208
8.2.2	MBII-52 changes alternative splicing of targeted pre-mRNAs in reporter gene assays	212
8.2.3	A mouse model of PWS shows changes in the predicted exons	212
8.2.4	MBII-52 is processed into smaller RNAs	215
8.2.5	MBII-52 derived RNAs do not bind to classical snoRNA-associated proteins	217
8.3	Discussion	219
8.3.1	The MBII-52 expression unit generates processed snoRNAs (psnoR- NAs)	219
8.3.2	MBII-52-derived psnoRNAs regulate splicing of several pre- mRNAs	220
8.3.3	Relevance for PWS	223
8.4	Materials and methods	223
9	Expression and Processing of a Small Nucleolar RNA from the Epstein-Barr Virus Genome	225
9.1	Introduction	226
9.2	Results	227

CONTENTS

9.2.1	dentification of v-snoRNA1 by cDNA cloning and expression analysis	227
9.2.2	Co-Immunoprecipitation and FISH analysis of v-snoRNA1 . .	229
9.2.3	v-snoRNA1 expression is strongly stimulated in the lytic cycle	232
9.2.4	Phenotypic traits of a recombinant virus lacking v-snoRNA1 .	232
9.2.5	Computational and functional analysis of v-snoRNA1	236
9.2.6	Processing of v-snoRNA1 into v-snoRNA1 ^{24pp} : potential v-snoRNA1 ^{24pp} targets	236
9.2.7	Conservation of v-snoRNA1 in other viral genomes	237
9.3	Discussion	242
9.4	Parts of Materials and Methods	245
9.4.1	Computational prediction of target sites in rRNAs	245
10	Conclusions	249
III	Acknowledgments	251

CONTENTS

Chapter 1

Introduction

This thesis is organized as follows. In the first part we start with an introduction to the problem of inferring transcription regulatory interactions (Chapter 2). Then we describe detailed methods of promoterome construction in human and mouse genomes (Chapter 3), and show how promoters together with gene expression data can be used to infer transcription regulatory interactions (Chapter 4). Finally, we show an application of this methodology to a human macrophage lineage undergoing differentiation accompanied by detailed experimental validation of the predicted network structure (Chapter 5). Work presented in Chapter 5 comes from the FANTOM4 project.

In the second part we focus on the function of two small nucleolar RNAs (snoRNAs). In Chapter 8 we describe an atypical function of snoRNAs – regulation of the mRNA alternative splicing process by a particular class of mouse snoRNAs (MBII-52 variants). These are of great interest, as the locus encoding MBII-52 is linked to Prader-Willi Syndrome. Methods include *in silico* RNA hybridization screens and experimental confirmation of the predictions. In Chapter 9, we report a discovery of the first virus-encoded snoRNA in Epstein-Barr Virus (EBV). Again, we show that function of this snoRNA (v-snoRNA1) is atypical: it is processed into small, 24 nt long fragments that can function as microRNAs.

Introduction

Part I

Transcriptional regulatory interactions

Chapter 2

Introduction

The process of gene expression and details of its regulation are two of the most studied and still most challenging topics in the molecular biology. All living cells are result of complex gene expression programs. Control takes place at the transcriptional and post-transcriptional level and integrates environmental and developmental signals. The *regulators* responsible for conveying this action need to distinguish a subset of their targets from a pool of genes. Usually this is achieved by recognition of short DNA or RNA motifs within a regulatory region of DNA or mRNA respectively. Regulators change expression of a target gene by many mechanisms, including changing of the rate RNA polymerase binds to the DNA locus, rate of transcription initiation, mRNA export, translation rate and degradation of mRNA templates. Here, we will focus on the first steps of the process of gene expression: regulation by proteins called transcription factors (TFs) which tune the transcription rate of their target genes.

There is a substantial amount of TFs present in eukaryotic genomes; even the yeast genome contains several hundreds of them. They need to control, in a concerted fashion, expression of most of the genes and implement a program necessary in a given environmental condition. Thus it is convenient to talk about transcriptional regulatory networks that govern specific processes (cellular differentiation, response to perturbation, disease, tissue identity maintenance). Such a network consists of a set of transcription factors and a set of target genes which they regulate. Structure of the connections and dynamics of the regulatory interactions is key to understanding the process.

Most of the approaches of inference of regulatory interactions start by identifying the regulatory proteins and their putative binding sites. The fact that regulatory factors recognize specific DNA stretches should, in principle, enable an extensive computational screen for the regulatory sites. In practice, however, there are several major difficulties. The motifs are often short and degenerate, which makes accurate binding sites predictions a hard task. Improvements include: specific and sensitive models of regulatory motifs, reduction of the length of the search regions to experi-

Introduction

mentally validated regions (promoters and/or enhancers), or usage of additional data such as comparative genomics or nucleosome occupancy data.

By far the most popular model of regulatory motifs are position specific weight matrices (WMs). These are tables of size $4 \times n$ (where n is the length of the sites) filled with $\{A, C, G, T\}$ nucleotide frequencies occurring at given positions within binding sites. As such, this model explicitly neglects all the information about nucleotide co-occurrences at different positions, and does not allow for changing the lengths of binding sites. However, due to its simplicity, it requires a relatively low number of known binding sites for parameter estimation and is straightforward to use. It is also important that the most comprehensive regulatory motif data bases contain only WMs. In the course of this work, we will use this model exclusively.

Apart from the primary nucleotide sequence, there are other sources of information to exploit in binding site predictions. One of them are orthologous sequences from related species. Alignments of these sequences provide a source of information about the evolutionary history of a sequence and help estimating the selective pressure (or lack thereof) acting on a locus. On the other hand, it has been shown that promoter sequences of genes belonging to different functional categories evolve with diverse speeds and show substantial turnover of binding sites, often faster than coding regions. Hence, a robust inference requires multiple genomes of species at different evolutionary proximity to the genome under study, and it should incorporate the knowledge of the evolutionary tree topology and branch lengths.

In mammals, many high-scoring sequences are effectively inaccessible to transcription factors due to coverage by nucleosomes. It is often the case that nucleosome-free regions coincide with clusters of TF binding sites (whether or not there is a direct causal dependency in either direction). Thus, computational predictions or experimental assessments of nucleosome occupancy is another indication of TF binding.

However, the most useful piece of information is the knowledge *where* to search for binding sites. Unlike prokaryotes, metazoans contain large fractions of non-coding sequences in their genomes, creating a lot of flexibility for positioning of regulatory regions. The canonical view for the placement of regulatory regions around transcription start sites is that: 1) a single transcription start site is positioned within a *core promoter* – the locus in a physical contact with the RNA polymerase and general transcription factors; 2) the *proximal promoter* spanning several hundred nucleotides around the core promoter, containing specific TF binding sites, mostly upstream of the transcription start site; and 3) a *distal promoter* which might span a couple of kilobases with a much lower density of TF binding sites. Alternative promoters in this view do not share proximal promoters, and usually lead to transcription of a different isoform of a gene (e.g. with a different exon structure). This paradigm is followed by common gene annotation routines – the annotators usually provide a short list of (and often just one) mRNA transcript per gene, which is a “representative” one.

Fortunately, due to development of new protocols and advances in sequencing, we

are now able to observe transcription start usage at single base pair resolution. This is achieved by deep Cap Analysis of Gene Expression (CAGE) (1), which extracts 20bp long sequences from the 5' ends of mRNAs, followed by deep sequencing (see Fig. 2.1). Using this data in Chapter 3, we extend the classical view of promoter architecture by defining of a three-level structure (“promoterome”) which clusters the neighboring transcription start sites. As a result of this procedure, we create a list of overlapping proximal promoters which are the input to the motif search algorithm.

Especially in mammals, distal elements such as enhancers and silencers are known to interact with transcriptional machinery located at the promoter. These regions contain binding sites recognized by the same repertoire of TFs as in promoters. There have been numerous attempts to discover such regions genome-wide. They include computational predictions of groups (“modules”) of binding sites for TFs known to function together, specific nucleosome modification combinations known to occur at regulatory elements (“chromatin signatures”), and physical contact (or proximity) between promoter and distal elements measured by 3C and 4C technologies. None of these, in our opinion, provides quality of annotation close to the one we derive for promoters. Another source of difficulty when working with distal elements is their association with particular genes/transcripts; there are known cases where regulatory regions lie several mega base-pairs away from the regulated TSS or within an intron of another gene.

Neither a presence of a high-scoring DNA sequence implies binding *in vivo*, nor does binding itself imply a function. To cope with this fact, most modern methods for reconstruction of transcriptional regulatory networks utilize expression data massively whenever such data is available for a given organism. The massive amount of expression data allows for machine learning approaches which try to predict expression of a new gene based on its promoter features and expression of many genes in multiple conditions. The observation which led to this approach is the existence of sets (unfortunately called again “modules”) of co-expressed genes. A classic example of this type of method is that of Beer and Tavazoie (2). This approach constructs modules in *Saccharomyces cerevisiae* based solely on expression data using a modified *k-means* algorithm which does not allow to construct clusters which are either too small (< 10 elements) or too wide (Pearson correlation coefficient > 0.65 of expression between centroid and elements). It then performs *de novo* motif discovery in predefined promoter regions of each module. The model is defined as a Bayesian network where features of promoters are at the input nodes and *OR*, *AND* and *NOT* logic gates are used to assign probabilities to 49 clusters. Additional constraints are imposed to ensure sparsity of the resulting network.

The strongest points of this method are its ability to model complex regulatory behavior, and little limitation on available promoter features. These include locations, spacings and orientations of the sequence occurrences. The weak points include the lack of association between discovered motifs and regulatory proteins. There is a

Introduction

limit on a number of possible predictions for a given gene: namely it can be assigned to one cluster (out of 49 in the original publication) with a representative averaged expression level. This method requires a large collection of conditions (255 in the publication) to distinguish modules robustly, and does it not allow for explaining expression of genes not assigned to any module.

Methods mentioned above aim to build complex classifiers that will explain expression levels as accurate as possible. However, in a typical experimental setup a researcher is interested in particular regulatory interactions (i.e. TF binding to promoter sequences) that influence gene expression changes in a limited number of conditions. Usually these are development stages of some tissues, stress conditions, gene knock-outs or over-expressions, disease conditions, etc. Often only two expression level measurements are performed: the perturbation and the background.

Methods of the second main branch take a form of a large regression scheme. The methodology presented in Chapter 4 belongs to this class. The first and most influential work of this kind was REDUCE by Bussemaker, Li and Siggia (3). Over time, the method was updated but the core idea stayed constant. In its most basic shape, the expression changes are modeled as a linear combination of oligonucleotide (presumed to be binding sites) counts in the promoters. The transcription factor activities are the unknown coefficients which need to be inferred from the data. REDUCE possessed a further capability, a motif finder. The motifs did not come predefined, but were found in such a way as to maximize the amount of explained variance in the expression changes. This was done in a “greedy” manner: one motif at a time was selected to maximize the explained fraction of variance and its contribution was then subtracted.

It is worth mentioning that this class of methods aims to explain how TF/motif activities are *changing* across different samples – usually this provides a better clue about the system than the absolute activity values. If the activity changes are small, the linear model is applicable: the expression changes are linear combinations of activity changes.

$$E_{p,s} = E_p^{mean} + \Delta E_{p,s} = E_p^{mean} + \sum_m R_{p,m} \cdot \Delta A_{m,s} + O(\Delta A^2) \quad (2.1)$$

Unfortunately, expression values for many genes change by orders of magnitude, making the linear approximation not applicable. However, similarly to many gene expression studies, we have found that a simple $\log(\cdot)$ transformation of the microarray intensity values or RNA-seq counts greatly improves the approximation. Another reason for such a transformation comes from the fact that in current protocols, cDNA needs to be processed in multiple steps, including PCR amplification, before it can be measured. This multiplicative noise phenomenon is well known in microarray field and we show in Chapter 3 that it also exists in other sequencing approaches

(deepCAGE and RNA-seq; it has been reported before for ChIP-seq too).

The $R_{p,m}$ values are of great interest as they actually define connectivity of the gene regulatory network. A non-zero value for $R_{p,m}$ denotes a regulatory edge between TF m and promoter p . Ultimately we would like to infer these numbers genome-wide, but currently, in order to solve for activities $A_{m,s}$, we need an approximation of these numbers.

Nguyen and D’Haeseleer (4) aimed to decompose the (log-)expression matrix ΔE (of size *promoters* \times *samples*) as in equation 2 into two matrices R and ΔA , using the condition $R \cdot \Delta A = \Delta E$. In the first step they let the R values be the site-counts of particular motifs, and found a least squares estimate for the matrix A . Subsequently, they let all the entries of R vary while keeping A fixed and again solved for best values. They iterated this procedure until convergence was attained. It is worth noting that the equations involving calculation of ΔA are very often overdetermined (number of motifs $<$ number of genes) and yield a unique solution. The calculations of R might become underdetermined if the number TFs binding to a promoter becomes bigger than the number of experimental conditions. Nguyen and D’Haeseleer improve this step by adding a regularization term, ensuring the smallest L2-norm solution is chosen. They also show that when such a term is added, the global decomposition of expression matrix E into a response matrix R and motif activities matrix ΔA is unique. In Chapter 5 we show that, in order to obtain biologically meaningful results and prevent massive overfitting, regularization is necessary, even in calculation of ΔA .

Bussemaker and Das provide lists of different methods published on the topic of prediction of gene expression. They can be found in (5) and under the address: <http://vision.lbl.gov/People/ddas/RegressionPrimer/>

We developed the Motif Activity Response Analysis (MARA) method as a core of the analysis presented in the FANTOM4 project (Chapter 5). It was subsequently improved, largely extended and automated for online use. We made it freely available to an any researcher in the world. The resulting platform, Integrated System for Motif Activity Response Analysis (ISMARA), can handle microarray, RNA-seq and ChIP-seq data. However, in our opinion for the clarity of the presented work it is important to first understand the (IS)MARA method (presented in Chapter 4) and afterwards see how it was applied in the FANTOM4 project (Chapter 5). We therefore present our research non-chronologically, but a reader is of course free to decide upon the order of reading the chapters.

Introduction

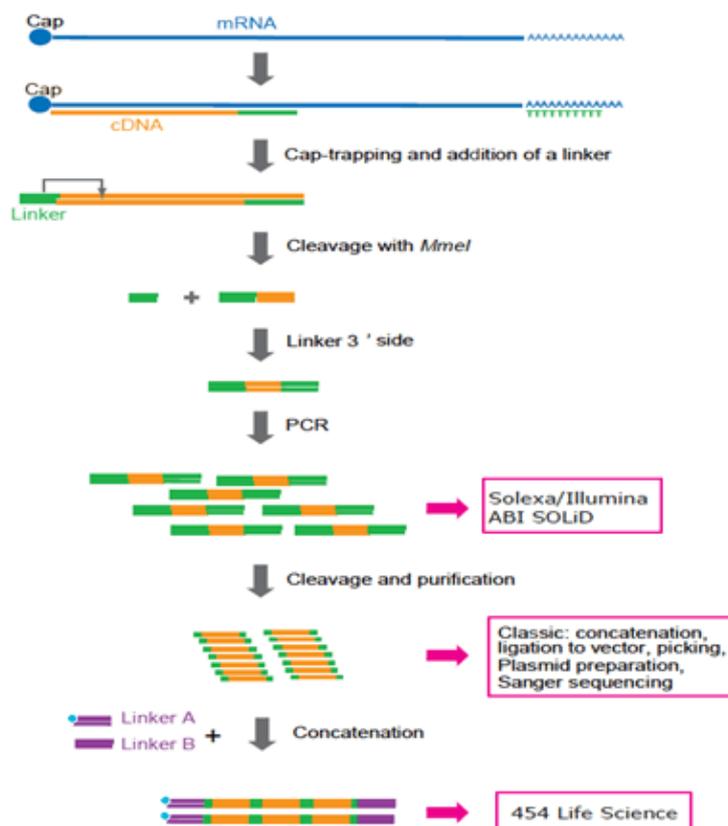


Figure 2.1: The CAGE protocol

Every mature mRNA contains a m^7G cap at its 5' end which, among other functions, protects it from degradation by exonucleases. CAGE protocol uses cap-trapping to select for full-length cDNAs. These are synthesized using random primers to ensure that any yet unknown transcripts can be selected. A ligated linker contains a recognition site for *MmeI* endonuclease, which cleaves double-stranded DNA 20 bp downstream from the linker. The 3' end linker is added and the tags are PCR-amplified. Subsequent steps depend on the sequencing technology in use. Data analyzed in this work comes primarily from 454 Life Sciences sequencing, which requires cleavage of the linkers, concatenation into longer sequences and addition of new linkers.

Chapter 3

Methods for analyzing deep sequencing expression data: Constructing the human and mouse promoterome with deepCAGE data

Piotr J Balwierz, Piero Carninci, Carsten Daub, Jun Kawai, Werner van Belle,
Christian Beisel, and Erik van Nimwegen
Genome Biology, 10(7):R79k 2009, PMID:19624849

With the advent of ultra high-throughput sequencing technologies, increasingly researchers are turning to deep sequencing for gene expression studies and it seems likely that this technology will eventually replace micro-arrays for expression analysis. Here we present a set of rigorous methods for data normalization, quantification of noise, and co-expression analysis. Using these methods on 122 deepCAGE samples of transcription start sites we construct genome-wide ‘promoteromes’ in human and mouse consisting of a three-tiered hierarchy of transcription start sites, transcription start clusters, and transcription start regions.

3.1 Background

In recent years a number of technologies, e.g. 454 and Solexa, have become available that allow DNA sequencing at very high throughput. Although originally these technologies have been used for genomic sequencing, most recently researchers have turned to using these ‘deep sequencing’ or ‘(ultra-)high throughput’ technologies for a number of other applications. For example, recently several researchers have used deep sequencing to map histone modifications genome-wide, or to map the locations

Constructing the human and mouse promoterome with deepCAGE data

at which transcription factors bind the DNA (ChIP-seq). Another application that is rapidly gaining attention is the use of deep sequencing for transcriptome analysis through the mapping of RNA fragments, e.g. (6; 7; 8; 9).

An alternative new high-throughput approach to gene expression analysis is deep CAGE sequencing (10). Cap analysis of gene expression (CAGE) is a relatively new technology introduced by Carninci et al. (1; 11) in which the first 20-21 nucleotides at the 5' ends of capped mRNAs are extracted by a combination of cap trapping and cleavage by restriction enzyme MmeI. The 'CAGE tags' thus obtained can then be sequenced and mapped to the genome. In this way a genome-wide picture of transcription start sites (TSSs) at single base pair resolution can be obtained. In the FANTOM3 project (12) this approach was taken to comprehensively map TSSs in the mouse genome. With the advent of deep sequencing technologies it has now become practical to sequence CAGE tag libraries to much greater depth providing millions of tags from each biological sample. At such sequencing depths significantly expressed TSSs are typically sequenced a large number of times. It thus becomes possible to not only map the locations of TSSs but also quantify the expression level of each individual TSS (10).

There are several advantages that deep-sequencing approaches to gene expression analysis offer over standard micro-array approaches. First, large-scale full length cDNA sequencing efforts have made it clear that most if not all genes are transcribed in different isoforms owing both to splice variation, alternative termination, and alternative transcription start sites (13). One of the drawbacks of micro-array expression measurements has been that the expression measured by hybridization at individual probes is often a combination of expression of different transcript isoforms that may be associated with different promoters and may be regulated in different ways (14). In contrast, because deep sequencing allows measurement of expression along the entire transcript the expression of individual transcript isoforms can in principle be inferred. CAGE-tag based expression measurements directly link the expression to individual transcription start sites, thereby providing a much better guidance for analysis of the regulation of transcription. Other advantages of deep sequencing approaches are that they avoid the cross-hybridization problem that micro-arrays have (15), and that they provide a larger dynamic range.

However, whereas for micro-arrays there has been a large amount of work devoted to the analysis of the data including issues of normalization, noise analysis, sequence-composition biases, background corrections, and so on, deep sequencing based expression analysis is still in its infancy and no standardized analysis protocols have been developed so far. Here we present new mathematical and computational procedures for the analysis of deep sequencing expression data. In particular, we have developed rigorous procedures for normalizing the data, we have developed a quantitative noise model, and we have developed a Bayesian procedure that uses this noise model to join sequence reads into clusters that follow a common expression profile across sam-

ples. The main application that we focus on in this paper is deepCAGE data. We apply our methodology to data from 66 mouse and 56 human CAGE-tag libraries. In particular, we identify transcription start sites (TSSs) genome-wide in mouse and human across a variety of tissues and conditions. In the first part of the results we present the new methods for analysis of deep sequencing expression data, and in the second part we present a statistical analysis of the human and mouse ‘promoteromes’ that we constructed.

3.2 Results and Discussion

3.2.1 Genome Mapping

The first step in the analysis of deep-sequencing expression data is the mapping of the (short) reads to the genome from which they derive. This particular step of the analysis is not the topic of this paper and we only briefly discuss the mapping method that was used for the application to deepCAGE data. CAGE tags were mapped to the human (hg18 assembly) and mouse (mm8 assembly) genomes using a novel alignment algorithm called nexalign(16) that maps tags in multiple passes. In the first pass exactly mapping tags are recorded. Tags that did not match in the first pass were mapped allowing a single base substitution. In the third pass the remaining tags are mapped allowing indels. For the majority of tags there is a unique genome position to which the tag maps with least errors. However, if a tag matched multiple locations at a best match level, a multi-mapping CAGE tag rescue strategy developed by Faulkner et al. (17) was employed. For each tag that maps to multiple positions, a posterior probability is calculated for each of the possible mapping positions which combines the likelihood of the observed error for each mapping with a prior probability for the mapped position. The prior probability for any position is proportional to the total number of tags that map to that position. As shown in (17) this mapping procedure leads to a significant increase in mapping accuracy compared to previous methods.

3.2.2 Normalization

Once the sequence reads or CAGE tags have been mapped to the genome we will have a (typically large) collection of positions from which at least one read/tag was observed. When we have multiple samples we will have, for each position, a read-count or tag-count profile which counts the number of reads/tags from each sample, mapping to that position. These tag-count profiles quantify the ‘expression’ of each position across samples and the simplest assumption would be to assume that the true expression is simply proportional to the tag-count in each sample. Indeed, recent papers dealing with RNA-seq data simply count the number of reads/tags per kilobase

Constructing the human and mouse promoterome with deepCAGE data

per million mapped reads/tags (RPKM)(6). This is, the tags are mapped to the annotated exonic sequences and their *density* is determined directly from the raw data. Similarly, previous efforts in quantifying expression from CAGE data (12) simply defined the *tags per million* of a TSS as the number of CAGE tags observed at the TSS divided by the total number of tags, multiplied by one million. However, such simple approaches assume that there are no systematic variations between samples (which are not controlled by the experimenter) which may cause the absolute tag-counts to vary across experiments. To investigate this issue we considered, for each sample, the distribution of tags per position.

For our CAGE data the mapped tags correspond to TSS positions. Figure 3.1 shows reverse-cumulative distributions of the number of tags per TSS for 6 human CAGE samples. On the horizontal axis is the total number of tags t and on the vertical axis the number of TSS positions to which at least t tags map. As the figure shows, the distributions of tags per TSS are power-laws to a very good approximation, spanning 4 orders of magnitude, and the slopes of the power-laws are also very similar. These samples are all from THP-1 cells both untreated and after 24 hours of PMA treatment. Very similar distributions are observed for essentially all CAGE samples currently available (data not shown).

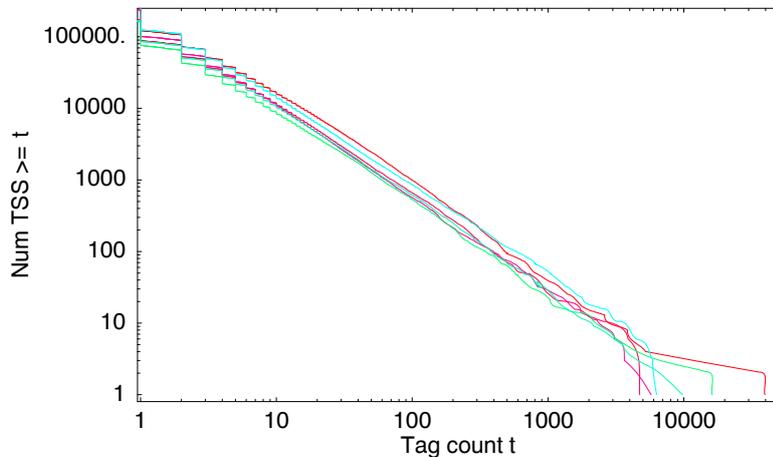


Figure 3.1: Reverse cumulative distributions for the number of different TSS positions that have at least a given number of tags mapping to them. Both axes are shown on a logarithmic scale. The 3 reddish curves correspond to the distributions of the 3 THP-1 cells control samples and the 3 blueish curves to the 3 THP-1 samples after 24 hours of PMA treatment. All other samples show very similar distributions (data not shown).

The large majority of observed TSSs have only a very small number of tags. These TSSs are often observed in only a single sample, and seem to correspond to very low

3.3.2 Results and Discussion

expression ‘background transcription’. On the other end of the scale there are TSSs that have as many as 10^4 tags. Manual inspection confirms that these correspond to TSSs of genes that are likely to be highly expressed, e.g. cytoskeletal or ribosomal proteins. It is quite remarkable in the opinion of these authors that both low expression background transcription, whose occurrence is presumably mostly stochastic, and the expression of the highest expressed TSSs, which is presumably highly regulated, occur at the extremes of a common underlying distribution. That this power-law expression distribution is not an artifact of the measurement technology is suggested by the fact that previous data from high-throughput SAGE studies have also found power-law distributions (18). For CHIP-seq experiments, the number of tags observed per region also appears to follow an approximate power-law distribution (19). In addition, our analysis of RNA-seq dataset from *Drosophila* shows that the number of reads per position follows an approximate power-law distribution as well (supplementary Fig. 3.16). These observations strongly suggest that RNA expression data generally obey power-law distributions. The normalization that we present here should thus generally apply to deep sequencing expression data.

For each sample we fitted (see Methods) the reverse-cumulative distribution of tags per TSS to a power-law of the form

$$n(t) = n_0 t^{-\alpha}, \quad (3.1)$$

with n_0 the inferred number of positions with at least $t = 1$ tag and α the slope of the power-law. Figure 3.2 shows the fitted values of n_0 and α on the horizontal and vertical axis for all 56 human CAGE samples. We see that, as expected, the inferred

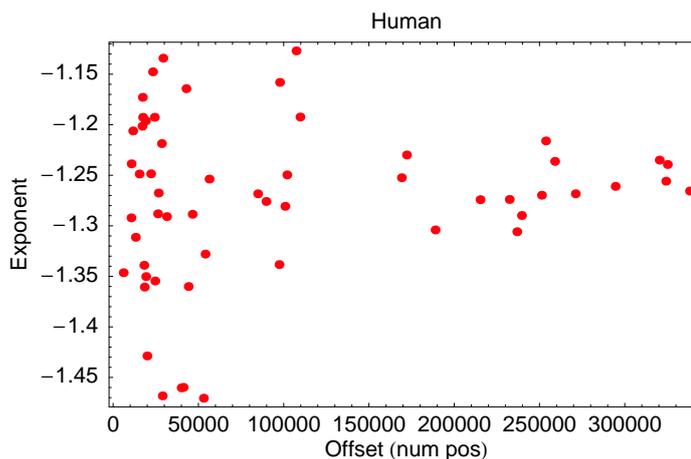


Figure 3.2: Fitted off-sets n_0 (horizontal axis) and fitted exponents α (vertical axis) for the 56 CAGE samples that have at least 100,000 tags.

number of positions n_0 varies significantly with the depth of sequencing, i.e. the dots

Constructing the human and mouse promoterome with deepCAGE data

on the right are from the more recent samples that were sequenced in much greater depth. In contrast, the fitted exponents vary relatively little around an average of about -1.25 , especially for the samples with large numbers of tags.

In the analysis of micro-array data it has become accepted that it is beneficial to use so-called quantile normalization, in which the expression values from different samples are transformed to match a common reference distribution (20). We follow a similar approach here. We make the assumption that the “true” distribution of expression per TSS is really the same in all samples, and that the small differences in the observed reverse-cumulative distributions are the results of experimental biases that are varying across samples. This includes fluctuations in the fraction of tags that maps successfully, variations in sequence-specific linker efficiency, the noise in PCR amplification, etcetera. To normalize our tag count we map all tags to a *reference distribution*. We chose as reference distribution a power-law with an exponent of $\alpha = -1.25$ and we chose the offset n_0 such that the total number of tags is precisely one million. We then used the fits for all samples to transform (see Methods) the tag-counts into normalized *tags per million* (TPM) counts. Figure 3.3 shows the same 6 distributions as above but now after the normalization. Although the

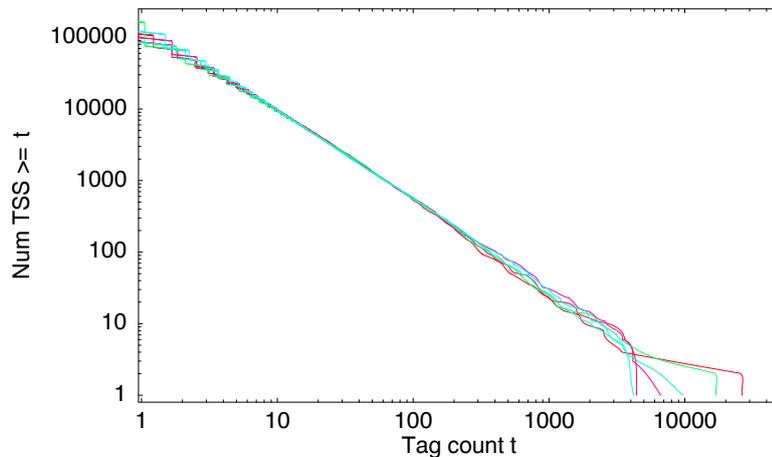


Figure 3.3: Normalized reverse cumulative distributions for the number of different TSS positions that have at least a given number of tags mapping to them. Both axes are shown on a logarithmic scale. The 3 reddish curves correspond to the distributions of the 3 THP-1 control samples and the 3 blueish curves to the 3 THP-1 samples after 24 hours of PMA treatment.

changes that this normalization introduces are generally modest, the nice collaps of the distributions shown in Fig. 3.3 strongly suggests that the normalization improves quantitative comparability of the expression profiles. Indeed, as described below, for a replicate data-set in which two deepCAGE libraries were constructed from a com-

mon mRNA sample, the normalization significantly reduces the apparent variation between the replicates' expression profiles. Finally, we note that normalization to a common power-law distribution has also been proposed for normalizing micro-arrays (21).

In the remainder we will use the normalized tag counts to compare the expression at individual positions in the genome across samples. We also retain the raw tag-counts because, as we will see below, the noise on the observed tag count depends on these raw counts.

3.2.3 Noise model

In order to analyze expression profiles it is necessary to analyze the distribution of the noise on deepCAGE and other deep-sequencing expression measurements and, to our knowledge, such an analysis has not yet been performed. Instead of determining noise on expression measurements, existing work has focused on defining models of the *background* distribution of tags/reads which can be used to identify regions that have significantly more mapped tags/reads than expected from the background model. These background models assume either a simple Poisson distribution, or a Poisson distribution with Gamma-distributed rate (22).

To quantitatively investigate the noise in the expression measurements we compared tag-counts across replicate data-sets. Among the currently available CAGE data-sets there is one pair in which two libraries were prepared from a common mRNA sample and figure 3.4 shows a scatter plot of the normalized tag counts (tags per million, TPM) from the replicate measurements. The figure shows that at high TPM (i.e. for tags with TPMs larger than $e^4 \approx 55$) the scatter has an approximately constant width whereas at low TPM the width of the scatter increases dramatically. This kind of funnel shape is familiar from micro-array expression where the increase in noise at low expression is caused by the contribution of non-specific background hybridization. However, for the deep CAGE data this noise is of an entirely different origin.

In deep sequencing experiments the noise comes from essentially two separate processes. First there is the noise that is introduced in going from the biological input sample to the final library that goes into the sequencer. Second, there is the noise introduced by the sequencing itself. For the CAGE experiments the former includes cap-trapping, linker ligation, cutting by the restriction enzyme, PCR amplification, and concatenation of the tags. In other deep-sequencing experiments, e.g. RNA-seq or ChIP-seq with Solexa sequencing, there will similarly be processes such as the shearing or sonication of the DNA, adding of the linkers, and growing clusters on the surface of the flow cell.

With respect to the noise introduced by the sequencing itself, it seems reasonable to assume that the N tags that are eventually sequenced can be considered a *random*

Constructing the human and mouse promoterome with deepCAGE data

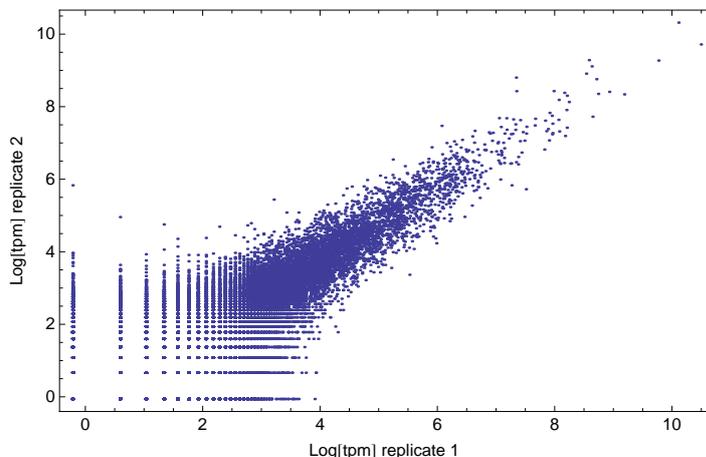


Figure 3.4: CAGE replicate from THP-1 cells after 8 hours of LPS treatment. For each position with mapped tags, the logarithm of the number of tags per million (TPM) in the first replicate is shown on the horizontal axis, and the logarithm of the number of tags per million in the second replicate on the vertical axis. Logarithms are natural logarithms.

sample of size N of the material that went into the sequencer. This will lead to relatively large ‘sampling’ noise for tags that form only a small fraction of the pool. For example, assume that a particular tag has fraction f in the tag pool that went into the sequencer. This tag is expected to be sequenced $n = Nf$ times among the N sequenced tags, and the actual number of times n that it is sequenced will be Poisson distributed according to

$$P(n|f, N) = \frac{(fN)^n}{n!} e^{-Nf}. \quad (3.2)$$

Indeed, recent work (23) shows that the noise in Solexa sequencing itself (i.e. comparing different lanes of the same run) is Poisson distributed. It is clear, however, that the Poisson sampling is not the only source of noise. In Fig. 3.4 there is an approximately fixed width of the scatter even at very high tag-counts, where the sampling noise would cause almost no difference in log-TPM between replicates. We thus conclude that, besides the Poisson sampling, there is an additional noise in the log-TPM whose size is approximately independent of the total log-TPM. Note that noise of a fixed size on the log-TPM corresponds to multiplicative noise on the level of the number of tags. It is most plausible that this multiplicative noise is introduced by the processes that take the original biological samples into the final samples that are sequenced, e.g. linker ligation and PCR amplification may vary from tag to tag and from sample to sample. The simplest, least biased, noise distribution assuming only a fixed size of the noise is a Gaussian distribution (24).

3.3.2 Results and Discussion

We thus model the noise as a convolution of multiplicative noise, specifically a Gaussian distribution of log-TPM with variance σ^2 , and Poisson sampling. As shown in the methods, if f is the original frequency of the TSS in the mRNA pool, and a total of N tags are sequenced, then the probability to obtain the TSS n times is approximately:

$$P(n|\sigma, f, N) = \frac{\exp\left(-\frac{(\log(n/N) - \log(f))^2}{2\sigma^2(n)}\right)}{n\sqrt{2\pi}\sigma(n)}, \quad (3.3)$$

where the variance $\sigma^2(n)$ is given by

$$\sigma^2(n) = \sigma^2 + \frac{1}{n}. \quad (3.4)$$

That is, the measured log-TPM is a Gaussian whose mean matches the log-TPM in the input sample, with a variance equal to the variance of the multiplicative noise (σ^2) plus one over the raw number of measured tags. The approximation (3.3) breaks down for $n = 0$. The probability to obtain $n = 0$ tags is approximately given by (Methods):

$$P(0|\sigma, f, N) = e^{-fN}. \quad (3.5)$$

We used the CAGE technical replicate (Fig. 3.4) to estimate the variance σ^2 of the multiplicative noise (Methods) and find $\sigma^2 = 0.085$. To illustrate the impact of the normalization, determining σ^2 on the same *unnormalized* data-set, we obtained $\sigma^2 = 0.11$, i.e. a 29% increase in the apparent noise between the replicates. In addition to this replicate, among the human CAGE data-sets there is a time course of THP-1 cells after PMA treatment, measured in triplicate, which includes samples before PMA treatment and after only 1 hour of PMA treatment. Manual inspection shows that the correlation of tags per TSS for these two samples is as large as for the technical replicate. This makes sense because on the time scale of one hour the expression of most transcripts can probably not change their expression appreciably (25). Using a procedure (Methods) that takes into account that a small fraction of TSSs may change expression significantly between the two samples, we estimated σ^2 as well for the three zero/one hour sample pairs. The values we estimate are, respectively, $\sigma^2 = 0.048$, $\sigma^2 = 0.116$, and $\sigma^2 = 0.058$.

In summary, using 4 pairs of samples that are (almost) replicates we find estimates of σ^2 ranging from 0.048 to 0.116. Although this analysis provides some evidence that the size of multiplicative noise varies between samples, the range of inferred values is small and we will make the assumption that σ^2 is the same for all samples. As estimate of σ^2 we took an intermediate value of $\sigma^2 = 0.06$ for the rest of our CAGE analysis.

We next validated this noise model as follows. According to our noise model, for

Constructing the human and mouse promoterome with deepCAGE data

TSSs that have nonzero expression in both samples, the z -statistic

$$z = \frac{\log(n') - \log(m')}{\sqrt{2\sigma^2 + \frac{1}{n} + \frac{1}{m}}}, \quad (3.6)$$

with m' the normalized expression at one hour and n' at zero hours, should be Gaussian distributed with standard deviation 1 (Methods). We tested this for the 3 biological replicates at zero/one hour and for the technical replicate. Figure 3.5 shows this theoretical distribution (in black) together with the observed histogram of z -values for the 4 replicates.

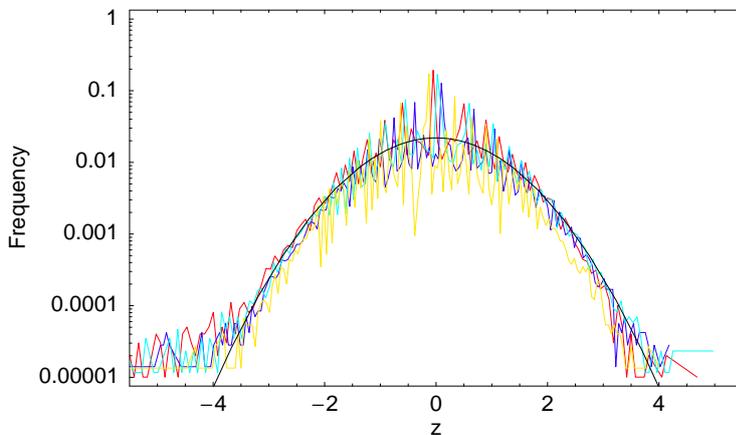


Figure 3.5: Observed histograms of z -statistics for the three zero/one hour (in red, dark blue, and light blue) samples and for the technical replicate (in yellow), compared with the standard unit Gaussian (in black). The vertical axis is shown on a logarithmic scale.

Although the data is noisy it is clear that all three curves obey a roughly Gaussian distribution. Note the deviation from the theoretical curve at very low z , i.e. less than $z < -4$ which appears *only* for the zero/one hour comparisons. These correspond to the the small fraction of positions that are significantly up-regulated at one hour. In summary, Fig. 3.5 clearly shows that the data from the replicate experiments are well described by our noise model.

To verify the applicability of our noise model to RNA-seq data we used two replicate data sets of *Drosophila* mRNA samples that were sequenced using Solexa sequencing and estimated a value of $\sigma^2 = 0.073$ for these replicate samples (Supplementary figure 3.17). This fitted value of σ^2 is similar to those obtained for the CAGE samples.

Finally, the σ^2 values that we infer for the deep sequencing data are somewhat larger than what one typically finds for replicate expression profiles as measured by

micro-arrays. However, it is important to stress that CAGE measures expression of individual TSSs, i.e. single positions on the genome, whereas micro-arrays measure the expression of an entire gene, typically by combining measurements from multiple probes along the gene. Therefore, the size of the ‘noise’ in CAGE and micro-array expression measurements cannot be directly compared. For example, when CAGE measurements from multiple TSSs associated with the same gene are combined, expression profiles become significantly less noisy between replicates ($\sigma^2 = 0.068$ versus $\sigma^2 = 0.085$, see supplementary figures 3.19 and 3.20). This applies also to RNA-seq data ($\sigma^2 = 0.02$ versus $\sigma^2 = 0.074$, see supplementary figure 3.18).

3.2.4 Promoterome construction

Using the methods outlined above on CAGE data we can comprehensively identify TSSs genome-wide, normalize their expression and quantitatively characterize the noise distribution in the expression measurements. This provides the most detailed information on transcription starts and, from the point of view of characterizing the transcriptome, there is in principle no reason to introduce additional analysis.

However, depending on the problem of interest, it may be useful to introduce additional filtering and/or clustering of the TSSs. For example, whereas traditionally it has been assumed that each ‘gene’ has a unique promoter and transcription start site, large-scale sequence analysis, such as performed in the FANTOM3 project (12), have made it clear that most genes are transcribed in different isoforms that use different TSSs. Alternative TSSs not only involve initiation from different areas in the gene locus, e.g. from different starting exons, but TSSs typically come in local clusters spanning regions ranging from a few to over one hundred base pairs wide.

These observations raise the question as to what an appropriate definition of a ‘basal promoter’ is. Should we think of each individual TSS as being driven by an individual ‘promoter’, even for TSSs only a few base pairs apart on the genome? The answer to this question is a matter of choice and depends on the application in question. For example, for the FANTOM3 study the main focus was to characterize all distinct regions containing a significant amount of transcription initiation. To this end the authors simply clustered CAGE tags whose genomic mappings overlapped by at least one base pair (12). Since CAGE-tags are 20-21 bp long, this procedure corresponds to single-linkage clustering of TSSs within 20-21 bp of each other. A more recent publication (26) creates a hierarchical set of promoters by identifying all regions in which the density of CAGE tags is over a given cut-off. This procedure thus allows one to identify all distinct regions with a given total amount of expression for different expression levels and this is clearly an improvement over the *ad hoc* clustering method employed in the FANTOM3 analysis.

Both clustering methods just mentioned cluster CAGE tags based only on the overall density of mapped tags along the genome, i.e. they ignore the expression pro-

Constructing the human and mouse promoterome with deepCAGE data

files of the TSSs *across* the different samples. However, a key question that one often aims to address with transcriptome data is how gene expression is *regulated*. That is, whereas these methods can successfully identify the distinct regions from which transcription initiation is observed, they cannot detect whether the TSSs within a local cluster are similarly expressed across samples or that different TSSs in the cluster have different expression profiles. Manual inspection shows that, whereas there are often several nearby TSSs with essentially identical expression profiles across samples/tissues, one also finds cases in which TSSs that are only a few base pairs apart show clearly distinct expression profiles. We hypothesize that in the case of nearby co-expressed TSSs the regulatory mechanisms recruit the RNA polymerase to the particular area on the DNA but that the final TSS that is used is determined essentially by an essentially stochastic (thermodynamic) process. One could for example imagine that the polymerase locally slides back and forth on the DNA and chooses a TSS based on the affinity of the polymerase for the local sequence, such that different TSSs in the area are used in fixed relative proportions. In contrast, when nearby TSSs show different expression profiles we imagine that there are particular regulatory sites that control initiation at individual TSSs.

Whatever the detailed regulatory mechanisms are, it is clear that for the study of transcription regulation it is important to properly separate local clusters of TSSs that are co-regulated from those that show distinct expression profiles. Below we present a Bayesian methodology that clusters nearby TSSs into *transcription start clusters* (TSCs) that are co-expressed in the sense that their expression profiles are statistically indistinguishable.

A second issue is that, as shown by the power-law distribution of tags per TSS (Fig. 3.1), we find a very large number of different TSSs used in each sample and the large majority of these have very low expression. Many TSSs have only one or a few tags and are often observed in one sample only. From the point of view of studying the regulation of transcription it is clear that one cannot meaningfully speak of ‘expression profiles’ of TSSs that were observed only once or twice and only in one sample. That is, there appears to be a large amount of ‘background transcription’ and it is useful to separate these TSSs that are used very rarely, and presumably largely stochastically, from TSSs that are driven by true ‘promoters’ and that are significantly expressed in at least one sample. Below we also provide a simple method for filtering such ‘background transcription’.

Finally, for each of significantly expressed TSCs there will be a *proximal promoter region* that contains regulatory sites that control the rate of transcription initiation from the TSSs within the TSC. Since TSCs can occur close to each other on the genome, individual regulatory sites may sometimes be controlling multiple nearby TSCs. Therefore, in addition to clustering nearby TSSs that are co-expressed we introduce an additional clustering layer, in which TSCs with overlapping proximal promoter are clustered into *transcription start regions* (TSRs). Thus, whereas differ-

ent TSSs may share regulatory sites, the regulatory sites around a TSR only control the TSSs within the TSR.

Using the normalization method and noise model described above we have constructed comprehensive ‘promoteromes’ of the human and mouse genomes from 122 CAGE samples across different human and mouse tissues and conditions (Materials and Methods) by 1. clustering nearby co-regulated TSSs, 2. filtering out background transcription, 3. extracting proximal promoter regions around each TSS cluster, and 4. merging TSS clusters with overlapping proximal promoters into promoter regions. We now describe each of these steps in the promoterome construction.

3.2.4.1 Clustering adjacent co-regulated TSSs

We define transcription start clusters (TSCs) as sets of contiguous TSSs on the genome, such that each TSS is relatively close to the next TSS in the cluster, and the expression profiles of all TSSs in the cluster are indistinguishable up to measurement noise. To construct TSCs fitting this definition we will use a Bayesian hierarchical clustering procedure that has the following ingredients

1. We start by letting each TSS form a separate, 1 bp wide, TSC.
2. For each pair of neighboring TSCs there is prior probability $\pi(d)$ that these TSCs should be fused, which depends on the the distance d along the genome between the two TSCs.
3. For each pair of TSCs we calculate the likelihoods of two models for the expression profiles of the two TSCs. The first model assumes that the two TSCs have a constant relative expression in all samples (up to noise). The second model assumes that the two expression profiles are independent.
4. Combining prior and likelihoods of the two models we calculate, for each contiguous pair of TSCs, a posterior probability that the two TSCs should be fused.
5. We identify the pair with highest posterior probability and if this posterior is at least 1/2 we fuse this pair and return to step 2. Otherwise the clustering stops.

The details of the clustering procedure are described in the Methods. Here we will briefly outline the key ingredients. The key quantity for the clustering is the likelihood-ratio of the expression profiles of two neighboring TSCs under the assumptions that their expression profiles are the same and independent, respectively. That is if we denote by x_s the logarithm of the TPM in sample s of one TSC, and by y_s the log-TPM in sample s of a neighboring TSC, then we want to calculate the probability $P(\{x_s\}, \{y_s\})$ of the two expression profiles assuming the two TSCs are expressed

Constructing the human and mouse promoterome with deepCAGE data

in the same way, and the probability $P(\{x_s\})P(\{y_s\})$ of the two expression profiles assuming they are independent.

For a single TSS we write x_s as the sum of a mean expression μ , the sample-dependent deviation δ_s from this mean, and a noise term

$$x_s = \text{noise} + \mu + \delta_s. \quad (3.7)$$

The probability $P(x_s|\mu + \delta_s)$ is given by the noise-distribution (3.3). To calculate the probability $P(\{x_s\})$ of the expression profile we assume that the prior probability $P(\mu)$ of μ is uniformly distributed and that the prior probabilities of the δ_s are drawn from a Gaussian with variance α , i.e

$$P(\delta_s|\alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left[-\frac{\alpha}{2}(\delta_s)^2\right]. \quad (3.8)$$

The probability of the expression profile of a TSC is then given by integrating out the unknown ‘nuisance’ variables $\{\delta_s\}$ and μ :

$$P(\{x_s\}) = \int d\mu P(\mu) \prod_s \left[\int d\delta_s P(x_s|\mu + \delta_s) P(\delta_s|\alpha) \right]. \quad (3.9)$$

The parameter α , which quantifies the *a priori* expected amount of expression variance across samples, is determined by maximizing the joint likelihood of all TSS expression profiles (Methods).

To calculate the probability $P(\{x_s\}, \{y_s\})$ we assume that, even though the two TSCs may have different mean expressions, their deviations δ_s are the *same* across all samples. That is, we write

$$x_s = \text{noise} + \mu + \delta_s, \quad (3.10)$$

and

$$y_s = \text{noise} + \tilde{\mu} + \delta_s \quad (3.11)$$

The probability $P(\{x_s\}, \{y_s\})$ is then given by integrating out the nuisance parameters

$$P(\{x_s\}, \{y_s\}) = \int d\mu d\tilde{\mu} P(\mu) P(\tilde{\mu}) \prod_s \left[\int d\delta_s P(x_s|\mu + \delta_s) P(y_s|\tilde{\mu} + \delta_s) P(\delta_s|\alpha) \right]. \quad (3.12)$$

As shown in the Methods section, the integrals (3.9) and (3.12) can be done analytically. For each neighboring pair of TSCs we can thus analytically determine the log-ratio

$$L = \log \left[\frac{P(\{x_s\}, \{y_s\})}{P(\{x_s\})P(\{y_s\})} \right]. \quad (3.13)$$

3.3.2 Results and Discussion

To perform the clustering we also need a prior probability that two neighboring TSCs should be fused and we will assume that this prior probability depends only on the distance between the two TSCs along the genome. That is, for closely-spaced TSC pairs we assume it is a priori more likely that they are driven by a common promoter than for distant pairs of TSCs. To test this we calculated the log-ratio L of equation (3.13) for each consecutive pair of TSSs in the human CAGE data. Figure 3.6 shows the average of L as a function of the distance of the neighboring TSSs. Figure 3.6 shows that, the closer the TSSs the more likely they are to be co-

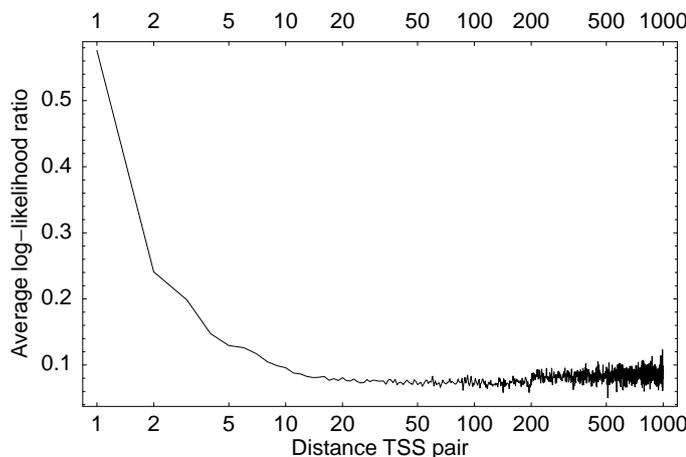


Figure 3.6: Average log-ratio L (equation 3.13) for consecutive pairs of individual TSSs as a function of the distance between the TSSs. The horizontal axis is shown on a logarithmic scale.

expressed. Once TSSs are more than 20 bps or so apart, they are not more likely to be co-expressed than TSSs that are very far apart. To reflect these observations, we will assume that the prior probability $\pi(d)$ that two neighboring TSCs are co-expressed falls exponentially with their distance d , i.e.

$$\pi(d) = e^{-d/l}, \quad (3.14)$$

where l is a length-scale that we set to $l = 10$.

For each consecutive pair of TSCs we calculate L and we calculate a prior log-ratio

$$R = \log \left(\frac{\pi(d)}{1 - \pi(d)} \right), \quad (3.15)$$

where the distance d between two TSCs is defined as the distance between the most highly expressed TSSs in the two TSCs. We iteratively fuse the pair of promoters for which $L + R$ is largest. After each fusion we of course need to update R and L for the neighbors of the fused pair. We keep fusing pairs until there is no longer any pair for which $L + R > 0$ (corresponding to a posterior probability of 0.5 for the fusion).

Constructing the human and mouse promoterome with deepCAGE data

3.2.4.2 Filtering background transcription

If one were principally interested in identifying all transcription initiation sites in the genome, one would of course not filter the set of TSCs obtained using the clustering procedure just described. However, when one is interested in studying regulation of expression then one would want to consider only those TSCs that show a substantial amount of expression in at least 1 sample and remove ‘background transcription’. To this end we have to determine a cut-off on expression level to separate background from significantly expressed TSCs. As the distribution of expression per TSS does not naturally separate into a high expressed and low expressed part, i.e. it is power-law distributed, this filtering is to some extent arbitrary.

According to current estimates there are a few hundred thousand mRNAs per cell in mammals. In our analysis we have made the choice to retain all TSCs such that, in at least 1 sample, at least 10 TPM derive from this TSC, i.e. at least 1 in 100,000 transcripts. With this conservative cut-off we ensure that there is at least 1 mRNA per cell in at least 1 sample. Since for some samples the total number of tags is close to 100,000, a TSC may spuriously pass this threshold by having only 2 tags in a sample with low total tag count. To avoid these we also demand that the TSC has 1 tag in at least 2 different samples. Note that if one were principally interested in ide

3.2.4.3 Proximal promoter extraction and transcription start region construction

Finally, for each of the TSCs we want to extract a *proximal promoter region* that contains regulatory sites that control the expression of the TSC, and in addition we want to cluster TSCs with overlapping proximal promoters. To estimate the typical size of the proximal promoters we investigated conservation statistics in the immediate neighborhood of TSCs. For each human TSC we extracted phastCons (27) scores 2.5 kilobases upstream and downstream of the highest expressed TSS in the TSC and calculated average PhastCons scores as a function of position relative to TSS (Fig. 3.7). We observe a sharp peak in conservation around TSS suggesting that the functional regulatory sites are highly concentrated immediately around the TSS. Upstream of TSS the conservation signal decays within a few hundred base pairs, whereas downstream of TSS the conservation first drops sharply and then more slowly. The longer tail of conservation downstream of TSS is most likely due to selection on the transcript rather than on transcription regulatory sites.

Based on these conservation statistics we conservatively chose the region from -300 to $+100$ with respect to TSS as the proximal promoter region. Although the precise boundaries are to some extent arbitrary it is clear that the conserved region peaks in a narrow region of only a few hundred base pairs wide around TSS. As a final step in the construction of the promoteromes we clustered together all TSCs

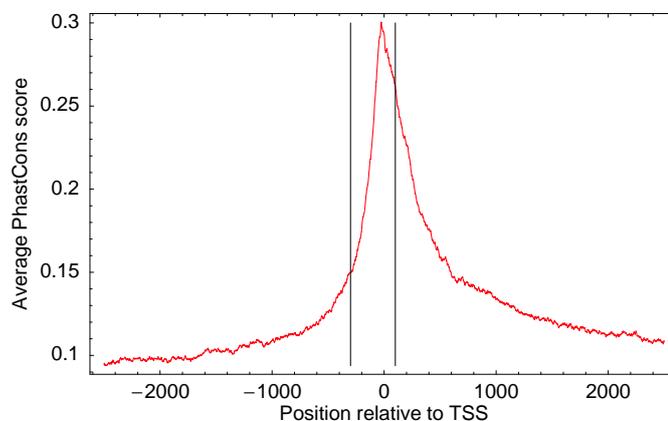


Figure 3.7: Average phastCons (conservation) score relative to TSS of genomic regions upstream and downstream of all human TSCs.

whose proximal promoters regions (i.e. from 300 bps upstream of the first TSS in the TSC to 100 bps downstream of the last TSS in the TSC) overlap into *transcription start regions* (TSRs).

3.2.5 Promoterome Statistics

To characterize the promoteromes that we obtained we compared them with known annotation and we determined a number of key statistics.

3.2.5.1 Comparison with known transcription starts

Using the collection of all human mRNAs from the UCSC database (28) we compared the location of our TSCs with known transcription starts. For each TSC we identified the position of the nearest known transcript start and Fig. 3.8 shows the distribution of the number of TSCs as a function of the relative position of the nearest known start.

By far most common is that there is a known start within a few base pairs of the TSC. We also observe a reasonable fraction of cases where a known start is somewhere between 10 and 100 base pairs either upstream or downstream of the TSC. Known starts more than 100 base pairs from TSC are relatively rare and the frequency drops further with distance, with only a few cases of known starts 1000 base pairs away from the TSC. For 37.7% of all TSCs there is no known start within 1000 base pairs of the TSC and for 27% there is no known start within 5 kilobases. We consider these latter 27% of TSCs novel TSCs. To verify that the observed conservation around TSS shown in Fig. 3.7 is not restricted to known starts we also constructed a profile of average PhastCons scores around novel TSCs (Fig. 3.9). We observe

Constructing the human and mouse promoterome with deepCAGE data

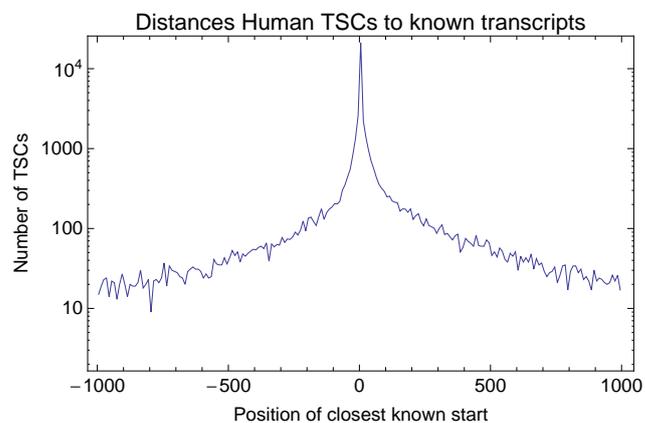


Figure 3.8: Number of TSCs as a function of their position relative to the nearest known transcript start. Negative numbers mean the nearest known start is upstream of the TSC. The vertical axis is shown on a logarithmic scale. The figure shows only the 46,293 TSCs (62.3%) with a known start within 1000 base pairs.

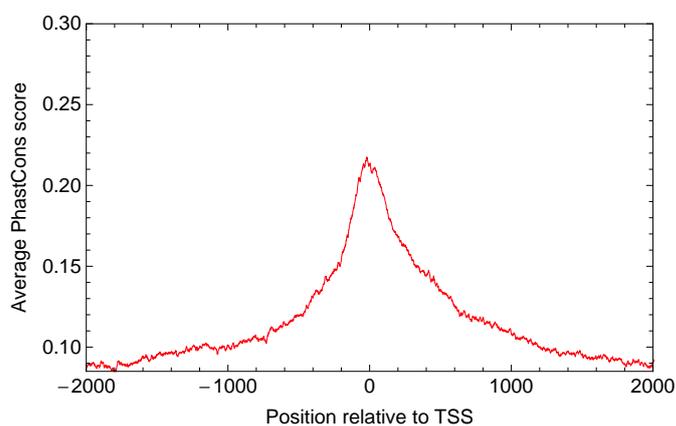


Figure 3.9: Average phastCons (conservation) score relative to TSS of genomic regions upstream and downstream of human TSCs that are more than 5 kilobases away from any known transcript.

a similar peak as the one for all TSCs, although its height is a bit lower and the peak appears a bit more symmetrical, showing only marginally more conservation downstream than upstream of TSS. Although we can only speculate, one possible explanation for the more symmetric conservation profile of novel TSCs is that this class of TSCs might contain transcriptional enhancers that show some transcription activity themselves. In the supplementary material we present analogous figures for the mouse promoterome (figures 3.26 3.25).

3.2.5.2 Hierarchical structure of the proteome

Table 3.1 shows the total numbers of CAGE tags, TSCs, TSRs, and TSSs within TSCs that we find for the human and mouse CAGE data sets. The 56 human CAGE samples

Statistic	Human	Mouse
Number of samples	56	66
Number of mapped CAGE tags	25'469'648	8'104'796
Number of TSSs	6'395'686	1'515'273
Number of TSSs in TSCs	860'823	608'474
Number of TSCs	74'273	77'286
Number of TSRs	43'164	50'915

Table 3.1: Global statistics of the human and mouse ‘promoteromes’ that we constructed from the human and mouse CAGE data. Shown are the number of different samples, the total number of CAGE tags that were mapped to the genome, the total number of different TSSs that were observed at least once, the number of TSSs in transcription start clusters (TSCs), the number of TSCs, and the number of transcription start regions (TSRs).

identify about 74,000 TSCs and the 66 mouse samples identify about 77,000 TSCs. Within these TSCs are about 860,000 and 608,000 individual TSSs, corresponding to about 12 TSSs per TSC in human and about 8 TSSs per TSC in mouse. Note that, while large, this number of TSSs is still much lower than the total numbers of unique TSSs that were observed. This again underscores the fact that the large majority of TSSs are expressed at very low levels.

Next we investigated the hierarchical structure of the human promoter regions (similar results are obtained in mouse, supplementary materials 3.5.8). Figure 3.10 shows the distributions of the number of TSSs per TSC, the number of TSSs per TSR, and the number of TSCs per TSR.

The middle panel shows that the number of TSCs per TSR is essentially exponentially distributed. That is, it is most common to find only a single TSC per TSR, TSRs with a handful of TSCs are not uncommon, and TSRs with more than 10 TSCs

Constructing the human and mouse promoterome with deepCAGE data

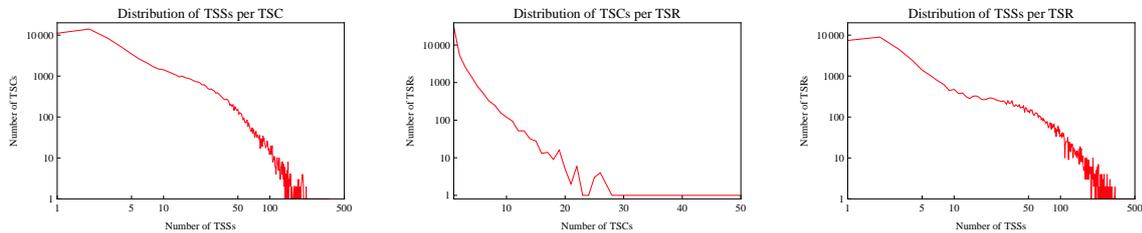


Figure 3.10: Hierarchical structure of the human promoterome. Left: Distribution of the number of transcription start sites (TSSs) per per co-expressed transcription start cluster (TSC). Middle: Distribution of the number of TSCs per transcription start region (TSR). Right: Distribution of the number of TSSs per TSR. The vertical axis is shown on a logarithmic scale in all panels. The horizontal axis is shown on a logarithmic scale in the left and right panels.

are very rare. The number of TSSs per TSC is more widely distributed (left panel). It is most common to find 1 or 2 TSSs in a TSC, and the distribution drops quickly with TSS number. However, there is a significant tail of TSCs with between 10 and 50 or so TSSs. The observation that the distribution of the number of TSSs per TSC has two regimes is even more clear from the right panel of Fig. 3.10 which shows the distribution of the number of TSSs per TSR. Here again we see that it is most common to find 1 or 2 TSSs per TSR, and that TSRs with between 5 and 10 TSSs are relatively rare. There is, however, a fairly wide shoulder in the distribution corresponding to TSRs that have between 10 and 50 TSSs. These distributions suggest that there are two types of promoters: ‘specific’ promoters with at most a handful of TSSs in them, and more ‘fuzzy’ promoters with more than 10 TSSs.

This observation is further supported by the distribution of the lengths of TSCs and TSRs (Fig. 3.11). In particular, the distribution of the length of TSRs (right panel) also shows a clear shoulder involving lengths between 25 and 250 base pairs or so.

3.2.5.3 Comparison with simple single-linkage clustering

In the supplementary materials (3.5.6) we compare the promoteromes obtained with our clustering procedure with those that are obtained with the simple single-linkage clustering procedures used in FANTOM3. The key difference between our clustering and single-linkage clustering employed in FANTOM3 is that in our procedure neighboring TSS with significantly different expression profiles are not clustered. Although TSSs within a few base pairs of each other the genome often show correlated expression profiles, it is also quite common to find nearby TSSs with significantly differing expression profiles. Figure 3.12 shows two examples of regions that contain

3.3.2 Results and Discussion

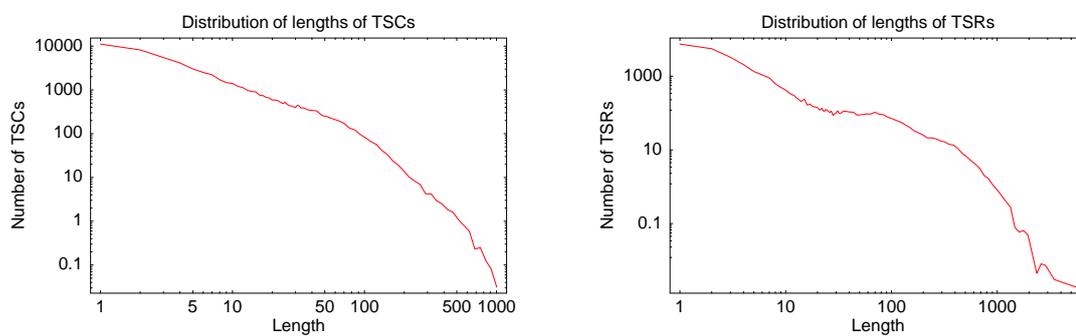


Figure 3.11: Length (base pairs along the genome) distribution of transcription start clusters (left panel) and transcription start regions (right panel). Both axes are shown on logarithmic scales in both panels.

multiple TSSs close to each other on the genome, where some TSSs clearly correlate in expression whereas others do not.

Constructing the human and mouse promoterome with deepCAGE data

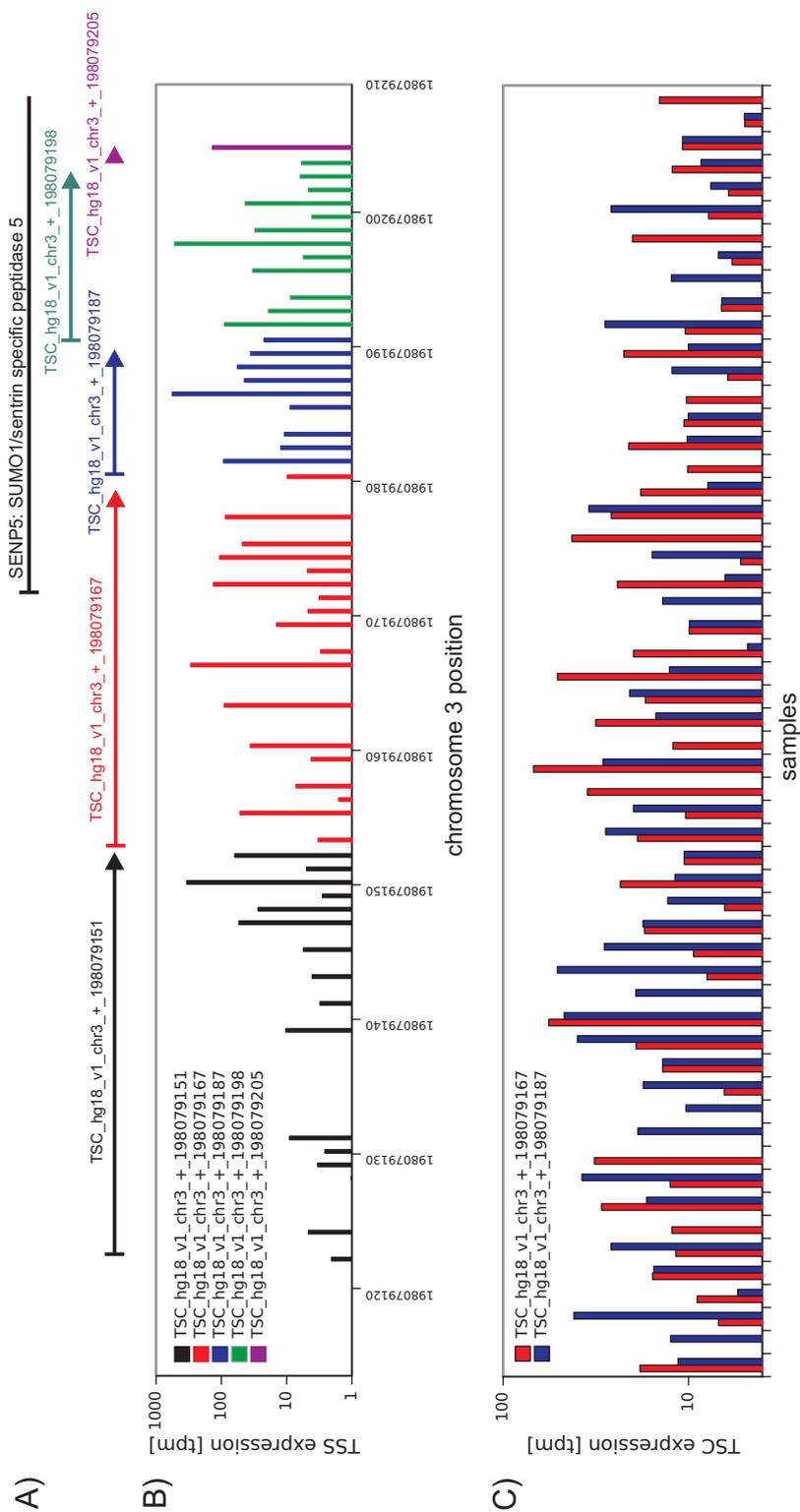


Figure 3.12: Nearby TSCs with significantly differing expression profiles. (A) A 90 base pair region on chromosome 3 containing 5 TSCs (colored segments) and the start of the annotated locus of the SENP5 gene (black segment). (B) Positions of the individual TSSs in the TSC and their total expression, colored by the TSC to which each TSS belongs. (C) expression across the 56 CAGE samples for the red and blue TSCs.

3.3.2 Results and Discussion

Within a region less than 90 base pairs wide our clustering identifies 5 different TSCs that each (except for the downstream most TSC) contain multiple TSSs with similar expression profiles. Any clustering algorithm that ignores expression profiles across samples would cluster all these TSSs into one large TSC. However, as shown in panel Fig. 3.12C for the red and blue colored TSCs, their expression across samples are not correlated at all. A scatter plot of the expression in TPM of the red and blue TSC is shown in the supplementary material (Fig. 3.23), where an additional example analogous to Fig. 3.12 is shown as well.

Since clustering procedures that ignore expression profiles, such as the single-linkage clustering employed in FANTOM3, cluster nearby TSSs with quite dissimilar expression profiles, one would expect that this clustering would tend to ‘average out’ expression differences across samples. To test this we calculated for each TSC the standard deviation in expression (log-tpm) for both our TSCs and those obtained with the FANTOM3 clustering. Figure 3.13 shows the reverse cumulative distributions of the standard deviations for the two sets of TSCs. The figure shows that there

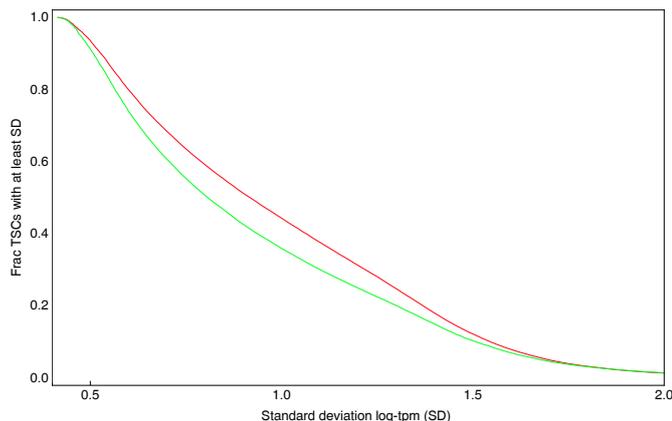


Figure 3.13: Reverse cumulative distributions of the standard deviation in expression across the 56 CAGE samples for the TSCs obtained with our clustering procedure (red) and the FANTOM3 single-linkage clustering procedure.

is a substantial decrease in the expression variation of the TSCs obtained with the FANTOM3 clustering than in the TSCs obtained with our clustering. This illustrates that, as expected, clustering without regard for the expression profiles of neighboring TSSs leads one to averaging out of expression variations. As a consequence, for TSCs obtained with our clustering procedure one is able to detect significant variations in gene expression, and thus potential important regulatory effects, that are undetectable when one uses a clustering procedure that ignores expression profiles.

3.2.6 High and low CpG promoters

Our promoterome statistics above suggest that there are two classes of promoters. That there are two types of promoters in mammals was already suggested in previous CAGE analysis (12) where the wide and fuzzy promoters were suggested to be associated with CpG islands, whereas promoters with a TATA-box tended to be narrow. To investigate this we calculated the CG- and CpG-content of all human promoters. That is, for each transcription start region (TSR) we determined the fraction of all nucleotides that are either C or G (CG-content), and the fraction of all dinucleotides that are CpG (CpG-content). Figure 3.14 shows the two-dimensional histogram of CG-content and CpG-content of all human TSRs. Figure 3.14 clearly shows that

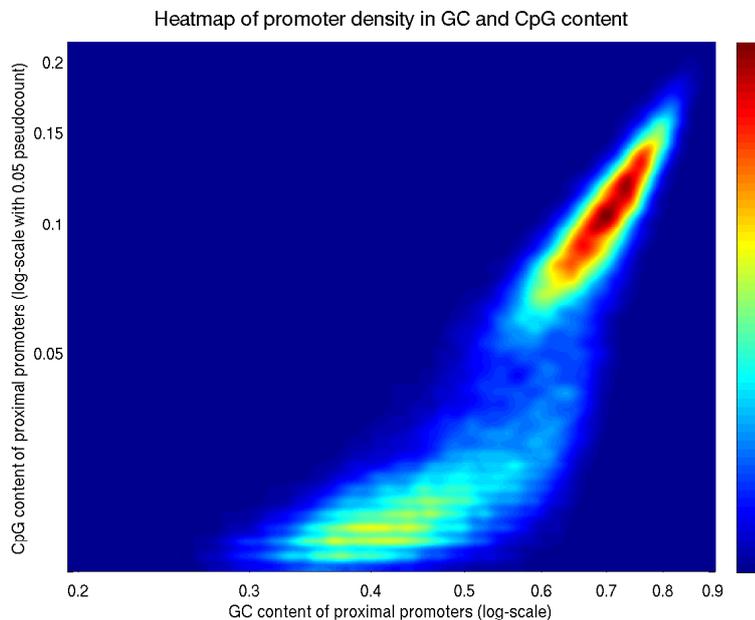


Figure 3.14: Two dimensional histogram (shown as a heatmap) of the CG base content (horizontal axis) and CpG dinucleotide content (vertical axis) of all human transcription start regions (TSRs). Both axes are shown on a logarithmic scales

there are two classes of TSRs with respect to CG- and CpG-content. Although it has been demonstrated previously that CpG-content of promoters shows a bimodal distribution (29) the simultaneous analysis of both CG- and CpG-content allows for a more efficient separation of the two classes, and demonstrates more clearly that there are really only two classes of promoters. We devised a Bayesian procedure to classify each TSR as high-CpG or low-CpG (Methods) that allows us to unambiguously classify the promoters based on their CG- and CpG-content. In particular, for more

than 91% of the promoters the posterior probability of the high-CpG class was either larger than 0.95 or less than 0.05.

To study the association between promoter class and its length distribution we selected all TSRs that with posterior probability 0.95 or higher belong to the high-CpG class, and all TSRs that with probability 0.95 or higher belong to the low CpG class, and separately calculated the length distributions of the two classes of TSRs. Figure 3.15 shows that the length distributions of high-CpG and low-CpG TSRs are

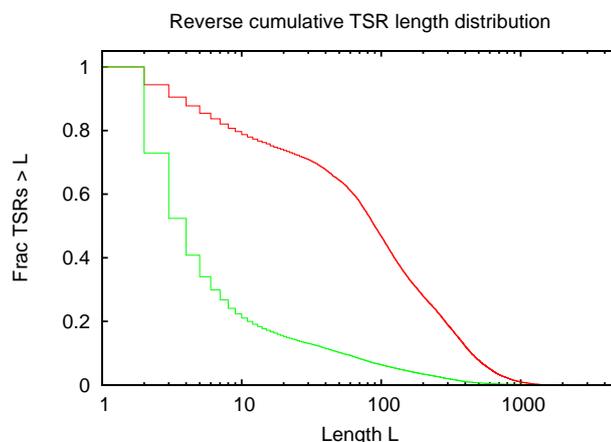


Figure 3.15: Reverse cumulative distribution of the lengths (base pairs along the genome) of transcription start regions for high-CpG (red curve) and low-CpG (green curve) promoters. The horizontal axis is shown on a logarithmic scale.

dramatically different, supporting observations made with previous CAGE data (12). For example, for the high-CpG TSRs only 22% have a width of 10 bps or less. In contrast, for the low-CpG TSRs approximately 80% of the TSRs have a width of 10 bps or less. In summary, our analysis supports that there are two promoter classes in human: one class associated with low CpG-content, low CG-content, and narrow TSRs, and one class associated with high CpG-content, high CG-content, and wide promoters. Similar results were obtained for mouse TSRs (data not shown).

Finally, we compared the promoter classification of known and novel TSRs. Of the 43,164, TSRs 37.7% are novel, i.e. there is no known transcript whose start is within 5 Kb of the TSR. For both known and novel TSRs the classification into high-CpG and low-CpG is ambiguous for about 8% of the TSRs. However, whereas for known TSRs 56% are associated with the high-CpG class, for novel TSRs 76% is associated with the low-CpG class. This is not surprising given that high-CpG promoters tend to be higher and more widely expressed than low-CpG promoters, i.e. they are much less likely to not have been observed previously.

3.3 Conclusions

It is widely accepted that gene expression is regulated to a large extent at the rate of transcription initiation. Currently regulation of gene expression is studied mostly with oligonucleotide micro-array chips. However, most genes initiate transcription from multiple promoters, and while different promoters may be differently regulated, the micro-array will typically only measure the sum of the isoforms transcribed from the different promoters. In order to study gene regulation it is therefore highly beneficial to monitor the expression from individual TSSs genome-wide and the deep CAGE technology now allows to do precisely that. The related RNA-seq technology similarly provides significant benefits over micro-arrays. We therefore expect that, as the costs of deep sequencing continues to come down, deep sequencing technologies will gradually replace micro-arrays for gene expression studies.

Application of deep sequencing technologies for quantifying gene expression is still in its infancy and, not surprisingly, there are a number of technical issues that complicate the interpretation of the data. For example, different platforms exhibit different sequencing errors at different rates and currently these inherent biases are only partly understood. Similarly, it is also clear that the processing of the input samples to prepare the final samples that are sequenced introduces biases that are currently poorly understood. It is likely that many technical improvements will be made over the coming years to reduce these biases.

Apart from the measurement technology as such, an important factor in the quality of the final results is the way in which the raw data are analyzed. The development of analysis methods for micro-array data are very illustrative in this respect. Several years of in-depth study passed before a consensus started to form in the community regarding the appropriate normalization, background subtraction, correction for sequence biases, and noise model. We expect that gene expression analysis using deep sequencing data will undergo a similar development in the coming years. Here we have presented an initial set of procedures for analyzing deep sequencing expression data, with specific application to deep CAGE data.

Our available data suggest that, across all tissues and conditions, the expression distribution of individual TSSs is a universal power-law. Interestingly, this implies that there is no natural expression scale that distinguishes the large number of TSSs which are expressed at very low rates, i.e. so-called background transcription, from the highly regulated expression of the TSSs of highly expressed genes. That is, background transcription and the TSSs of the most highly expressed genes are just the extrema of a scale-free distribution. As we have shown, by assuming that a common universal power-law applies to all samples we can normalize the expression data from different deep sequencing data-sets. The fact that expression profiles from SAGE and from RNA-seq using the Solexa platform also show power-law distributions strongly suggests that this normalization scheme is applicable to deep sequencing expression

data in general. It should be noted that, although all observed distributions are power-laws, there is no *a priori* reason that mammalian cells should have a *common* power-law expression distribution across all tissues and conditions. It is conceivable that, as more extensive data becomes available in the future, we may find significant differences between the expression distributions in different tissues.

The noise in the expression measured across different deep CAGE samples can be accurately modeled by a convolution of multiplicative noise and Poisson sampling and we derived a practical analytical approximation to the resulting noise distribution. Using replicate data-sets we inferred the size of the multiplicative noise for different samples and found it to vary in a small range. In addition, analysis of Solexa RNA-seq data from *Drosophila* showed multiplicative noise of similar size. However, we expect that it is a simplification to assume that the multiplicative really is identical in all experiments, and in the future we will want to apply a more refined analysis that takes into account the differences in the size of the multiplicative noise for different samples. To this end it will be important to design experiments such that at least 1 replicate is available to estimate the size of the multiplicative noise associated with a given experimental procedure.

The noise model allows us to rigorously assess the statistical significance of measured expression differences across different samples. In particular, we developed a Bayesian procedure that calculates the probability that two TSSs have identical expression profiles. Interestingly, we found that TSSs that are less than 10 bps apart on the genome are much more likely to be co-expressed than more distal neighboring TSSs. Using these results we clustered sets of nearby co-expressed TSSs into *transcription start clusters* (TSCs) that we propose are each regulated by a common ‘promoter’. Of course, our ability to detect significant expression differences is limited by the number of available samples and we expect that, as the number of available deep CAGE samples increases, the number of TSCs will increase as well.

Comparative genomic analysis shows a strong peak in sequence conservation restricted to a few hundred base pairs around TSSs. This suggests that the *proximal promoter* associated with each TSC extends a few hundred bps around the TSSs in the TSC. Besides clustering nearby co-expressed TSSs into TSCs we also clustered TSCs whose proximal promoters overlap into transcription start regions (TSRs). Comparing the sequence composition and widths of TSRs we find that there are two classes of promoters in the human and mouse genomes. The first class corresponds to TSRs that are narrow, almost always less than 10 bps wide, and that have low CG-content as well as low CpG-content. The second class corresponds to TSRs that are wide, i.e. anywhere from 25 to 250 bps wide, and that are associated with CpG islands, i.e. having both a high CG-content as well as a high CpG-content. It seems plausible that different mechanisms may be involved in the regulation of these two classes of promoters.

3.4 Materials and Methods

3.4.1 CAGE and RNA-seq expression data

All the samples used in this study were provided by Riken Genomic Sciences Center and come from the FANTOM3, the FANTOM4, and several smaller projects. Each human sample has at least 100,000 mapped tags, and each mouse sample at least 50,000. The lists of all 56 human and 66 mouse samples, with tissue/cell line name, treatment and accession numbers are available from the Genome Biology web page <http://genomebiology.com/2009/10/7/R79/additional> Whenever assigned, accession numbers of the DNA Data Bank of Japan are listed. Raw CAGE data of the FANTOM4 project are available at <http://fantom.gsc.riken.jp/4/>.

The CAGE protocol that was used has been described in (30). The 143 C6 mouse hippocampus and h93, i02, i03 human THP-1 libraries are produced using more recent protocol adapted to 454 Life Sciences (Roche) sequencer as described in methods section of (31). The lengths of the CAGE tags was 20-21bp in all cases.

For the RNA-seq data total RNA was isolated from *Drosophila* Kc cells using Trizol reagent. Purification of mRNA and the generation of cDNA library was performed following the Illumina protocol for mRNA sequencing. Primary sequencing data analysis was done following the Illumina Genome Analyzer software pipeline. ELAND was used for the alignment of short reads to the *Drosophila* genome (Release 5).

3.4.2 Normalization by fitting to a reference distribution

For each CAGE sample we fit the reverse-cumulative distribution $n(t)$ of the number of TSSs with at least t tags to a power-law. To robustly fit these power-laws across different samples with different total numbers of tags we remove the data from the first and last order of magnitude along the vertical axis and apply simple linear regression to the remaining data. As a result, for each sample s there will be a fitted exponent $\alpha(s)$ and a fitted offset $n_0(s)$

For a reference distribution of the form $n_r(t) = r_0 t^{-\alpha}$ the total number of tags is given by

$$T = \sum_{t=1}^{\infty} r_0 t^{-\alpha} = r_0 \zeta(\alpha), \quad (3.16)$$

where $\zeta(x)$ is the Riemann-zeta function. That is, the total number of tags is determined by both r_0 and α . For the reference distribution we chose $\alpha = 1.25$ and $T = \sum_t n_r(t) = 10^6$. Setting $\alpha = 1.25$ in equation (3.16) and solving for r_0 we find

$$r_0 = 217,623. \quad (3.17)$$

To map tag-counts from different samples to this common reference we transform the tag-count t in each sample into a tag-count t' according to

$$t \rightarrow t' = \lambda t^\beta \quad (3.18)$$

such that the distribution $n(t)$ for this sample will match the reference distribution, i.e. $n(t) = n_r(t')$. If the observed distribution has tag-count distribution

$$n(t) = n_0 t^{-\alpha}, \quad (3.19)$$

then in terms of t' this becomes:

$$n(t) = n_0 \left(\frac{t'}{\lambda} \right)^{-\alpha/\beta}. \quad (3.20)$$

Demanding that $n(t) = n_r(t')$ gives:

$$r_0(t')^{-1.25} = n_0 \left(\frac{t'}{\lambda} \right)^{-\alpha/\beta}. \quad (3.21)$$

This equation is satisfied when $\alpha/\beta = 1.25$, that is:

$$\beta = \frac{\alpha}{1.25}. \quad (3.22)$$

Using this and solving for λ we find:

$$\lambda = \left(\frac{r_0}{n_0} \right)^{1.25}. \quad (3.23)$$

3.4.3 Noise model

We model the noise as a convolution of multiplicative Gaussian noise and Poisson sampling noise. Assume that tags from a given TSS position correspond to a fraction f of the tags in the input pool. Let $x = \log(f)$ and let y be the log-frequency of the tag in the final prepared sample that will be sequenced, i.e. for CAGE after capturing, linking, PCR-amplification, and concatenation. We assume that all these steps introduce a Gaussian noise with variance σ^2 so that the probability $P(y|x, \sigma)$ is given by

$$P(y|x, \sigma)dy = \frac{e^{-(y-x)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma}} dy. \quad (3.24)$$

We assume that the only additional noise introduced by the sequencing is simply Poisson sampling noise. That is, the probability to obtain n tags for this position, given y and given that we sequence N tags in total is given by

$$P(n|N, y) = \frac{(e^y N)^n}{n!} e^{-N e^y}. \quad (3.25)$$

Constructing the human and mouse promoterome with deepCAGE data

Combining these two distributions we find that the probability to obtain n tags given that the log-frequency in the input pool was x is given by

$$P(n|\sigma, x, N) = \int_{-\infty}^0 \frac{(e^y N)^n}{n!} e^{-Ne^y} \frac{e^{-(y-x)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma} dy \quad (3.26)$$

This integral can unfortunately not be solved analytically. However, if the log-frequency x is high enough such that the expected number of tags $\langle n \rangle = Ne^x$ is substantially bigger than 1, then the Poisson distribution over y takes on a roughly Gaussian form over the area where $(y-x)^2$ is small enough to contribute substantially to the integral. We thus decided to approximate the Poisson by a Gaussian, i.e. we use

$$\frac{(e^y N)^n}{n!} e^{-Ne^y} \approx \frac{\exp\left(-\frac{n}{2}(y - \log(n/N))^2\right)}{\sqrt{2\pi n}} \quad (3.27)$$

Then the integral over y can be performed analytically. Since the integrand is already close to zero at $y = 0$ (no individual TSS accounts for the entire sample) we can extend the region of integration to $y = \infty$ without loss of accuracy. We then obtain

$$P(n|\sigma, x, N) = \frac{\exp\left(-\frac{(\log(n/N)-x)^2}{2\sigma^2(n)}\right)}{n\sqrt{2\pi}\sigma(n)}, \quad (3.28)$$

where the variance is given by

$$\sigma^2(n) = \sigma^2 + \frac{1}{n}. \quad (3.29)$$

In summary, the expected tag-count is such that the expected log-frequency $\log(n/N)$ matches the input log-frequency x , and has a noise variation of the size σ^2 plus one over the tag-count n .

Although this approximation is strictly only good for large n we find that in practice it is already quite good from $n = 3$ or so onwards and we decided to use this approximation for all tag-counts n . However, it is clear that for $n = 0$ the approximation cannot be used. For the case $n = 0$ we thus have to make an alternative approximation. The probability $P(0|\sigma, x)$ is given by the integral

$$P(0|\sigma, x) = \int_{-\infty}^0 \frac{\exp\left(-Ne^y - \frac{(x-y)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} dy. \quad (3.30)$$

We can again extend the integration range to $y = \infty$ without appreciable error. In addition we introduce a change of variables to

$$z = \frac{y-x}{\sigma} \quad (3.31)$$

and we introduce the variable m which represents the expected number of tags, i.e.

$$m = Ne^x. \quad (3.32)$$

With these definitions the integral becomes

$$P(0|\sigma, x) = \int_{-\infty}^{\infty} e^{-me^{\sigma z} - z^2/2} \frac{dz}{\sqrt{2\pi}}. \quad (3.33)$$

The Gaussian second term in the exponent ensures that the main contribution to the integral comes from the region around $z = 0$. We therefore expand $e^{\sigma z}$ to second order, i.e.

$$e^{\sigma z} \approx 1 + \sigma z + \frac{\sigma^2 z^2}{2}. \quad (3.34)$$

The integral then becomes a Gaussian integral and we obtain the result

$$P(0|\sigma, x) \approx \frac{\exp\left(-\frac{m(2+m\sigma^2)}{2(1+m\sigma^2)}\right)}{\sqrt{1+m\sigma^2}}. \quad (3.35)$$

For small σ this is in fact very close to

$$P(0|0, x) = e^{-m}. \quad (3.36)$$

Both expressions (3.35) and (3.36) are reasonable approximations to the probability of obtaining zero tags given an original log-frequency x .

3.4.4 Estimating the multiplicative noise component from the replicate

Assume a particular TSS position was sequenced n times in the first replicate sample and m times in the second replicate sample. Assume also that both n and m are larger than zero. A little calculation shows that the probability $P(n, m|\sigma)$ is given by

$$P(n, m|\sigma) \propto \frac{1}{\sqrt{2\sigma^2 + \frac{1}{n} + \frac{1}{m}}} \exp\left(-\frac{(\log(n/N) - \log(m/M))^2}{2(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}\right). \quad (3.37)$$

Note that we have not yet specified if by n and m we mean the raw tag-counts or the normalized version. For the comparison of expression levels, i.e. the difference $\log(n/N) - \log(m/M)$ it is clear we want to use the normalized values n' and m' . However, since the normalized values assume a total of one million tags, the normalized values cannot be used in the expression for the variance. Therefore, we use the

Constructing the human and mouse promoterome with deepCAGE data

raw tag-counts n and m in the expression for the variance. That is, the probability takes the form

$$P(n, m|\sigma) \propto \frac{1}{\sqrt{2\sigma^2 + \frac{1}{n} + \frac{1}{m}}} \exp\left(-\frac{\log^2(n'/m')}{2(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}\right). \quad (3.38)$$

We estimate the variance σ^2 by maximizing the probability of the data over all positions for which both n and m are larger than zero. Writing

$$\sigma^2(n, m) = 2\sigma^2 + \frac{1}{n} + \frac{1}{m}, \quad (3.39)$$

the log-probability L of the data can be written as

$$L = -\frac{1}{2} \sum_i \left[\log(\sigma^2(n_i, m_i)) + \frac{\log^2(n'_i/m'_i)}{2\sigma^2(n_i, m_i)} \right]. \quad (3.40)$$

We can now find the maximum of L with respect to σ^2 . Doing this on the replicate CAGE data set we find

$$\sigma = 0.085. \quad (3.41)$$

3.4.5 Estimating the multiplicative noise component by comparing zero and one hour expression in the THP-1 cells PMA time course

Using the assumption that few TSSs change their expression within 1 hour of treatment with PMA, we can also estimate σ^2 by comparing expression across TSSs in the CAGE samples of THP-1 cells before and after 1 hour of PMA treatment. We assume that a large fraction of the TSS positions should be expressed equally in the two experiments but allow for a small fraction of TSS positions to be expressed differently across the two time points.

Let Δ denote the size of the range in log-expression, i.e. the difference between highest and lowest log tag-count, which is about 20,000 in our experiments. We assume a uniform prior distribution $P(x) = 1/\Delta$ over log-frequency x . Assume a TSS position has expression m at zero hours and n at one hour. The probability of this expression given that both are expressed the same is $P(n, m|\sigma)$ that we calculated above. In contrast, if the expression is different between the two time points then the probability is just the prior $1/\Delta$. Let π denote the (unknown) fraction of all positions that is expressed differently between the two time points. Under these assumptions the likelihood of the data is

$$L(D|\sigma) = \prod_i \left[P(n_i, m_i|\sigma)(1 - \pi) + \frac{\pi}{\Delta} \right] \quad (3.42)$$

We now maximize this likelihood with respect to both π and σ^2 . Doing this on zero and one time points of the three replicates gives us estimated σ^2 's of $\sigma^2 = 0.048$, $\sigma^2 = 0.116$, and $\sigma^2 = 0.058$. Note that two of these are a less than the σ^2 's inferred from the replicate.

3.4.6 Likelihood of the expression profile of a single promoter

We want to calculate the likelihoods of two neighboring promoters under the assumption that they have fixed relative expression, and assuming the two profiles are independent. As discussed above, the probability of the observed tag-count n is to a good approximation Gaussian in the log-expression $\log(n)$ with a variance $(\sigma^2 + 1/n)$, where σ^2 is the variance due to replicate noise and $1/n$ is the variance due to the Poisson sampling. However, this Gaussian form breaks down when $n = 0$ and this makes analytic derivations impossible when data-points with zero counts are included. To circumvent this we make two approximations when considering the expression profiles of neighboring promoters. First, we discard all samples s in which both TSSs have zero tag-count $n_s = 0$, i.e. we assume in effect that samples for which both promoters have count zero are equally likely under both models. In addition, for samples s where one of the two TSSs has a zero count we replace the count zero with a pseudo-count of one half of a tag (being intermediate between no tags at all and 1 tag).

We focus first on the probability of the expression profile of a single promoter (considering only the samples in which at least one of the promoters has non-zero tag count). Let s denote a sample, t_s the normalized tag-per-million of a promoter in the sample, and n_s the unnormalized CAGE tag count in the sample. The log-expression values are given by

$$x_s = \log \left(t_s + \delta_{n_s 0} \frac{10^6}{2N_s} \right), \quad (3.43)$$

where the Kronecker delta function is 1 if and only if the tag-count n_s is zero and N_s is the total number of tags in sample s (over all TSSs). We now assume a model of the following form

$$x_s = \text{noise} + \mu + \delta_s, \quad (3.44)$$

where μ is the true average log-expression of this promoter and δ_s is the true deviation from this mean in sample s . Given our noise model we have

$$P(x_s | \mu, \delta_s) = \sqrt{\frac{w_s}{2\pi}} \exp \left[-\frac{w_s}{2} (x_s - \mu - \delta_s)^2 \right], \quad (3.45)$$

where

$$w_s = \frac{1}{\sigma^2 + 1/n_s}, \quad (3.46)$$

Constructing the human and mouse promoterome with deepCAGE data

σ^2 is the variance of the multiplicative noise, and we set $n_s = 1/2$ whenever $n_s = 0$. We need a prior probability distribution for the true expression variation δ_s and we will assume this prior to be Gaussian with mean zero, i.e. we assume

$$P(\delta_s|\alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left[-\frac{\alpha}{2}(\delta_s)^2\right], \quad (3.47)$$

where α sets the scale of the variation that promoters show. As discussed below, we choose α so as to maximize the likelihood of the all the expression profiles from all TSSs (assuming each TSS is independent).

To obtain the marginal probability of x_s given μ and α we perform the integral:

$$P(x_s|\mu, \alpha) = \int_{-\infty}^{\infty} P(x_s|\mu, \delta_s)P(\delta_s|\alpha)d\delta_s. \quad (3.48)$$

This is a Gaussian integral that can be easily performed and we obtain

$$P(x_s|\mu, \alpha) = \sqrt{\frac{\beta_s}{2\pi}} \exp\left[-\frac{\beta_s}{2}(x_s - \mu)^2\right], \quad (3.49)$$

where

$$\beta_s = \frac{w_s \alpha}{w_s + \alpha}. \quad (3.50)$$

Next, to obtain the marginal probability of x_s given only α we integrate over the mean log-expression μ and to do this we need a prior $P(\mu)$. For simplicity we use a uniform prior over some fixed range, i.e.

$$P(\mu) = \frac{1}{\Delta_\mu}. \quad (3.51)$$

when $-\Delta_\mu/2 \leq \mu \leq \Delta_\mu/2$ and zero outside of this range. We then obtain

$$P(x|\alpha) = \int_{-\Delta_\mu/2}^{\Delta_\mu/2} \prod_s P(x_s|\mu, \alpha) \frac{1}{\Delta_\mu} d\mu. \quad (3.52)$$

We will assume that Δ_μ is large compared to the region over which the probability takes on its maximum so that we can let the integral run from minus infinity to infinity without affecting the result. The precise value of Δ_μ is not important since it will eventually cancel out of the calculation. The result of the integral over μ is

$$P(x|\alpha) = \frac{1}{\Delta_\mu} \sqrt{\frac{2\pi}{S\langle\beta\rangle}} \prod_s \frac{\beta_s}{2\pi} \exp\left[-\frac{S}{2}\left(\langle\beta x^2\rangle - \frac{\langle\beta x\rangle^2}{\langle\beta\rangle}\right)\right], \quad (3.53)$$

where S is the number of samples (for which at least one of the two neighboring promoters has non-zero tag count) and the averages are defined as follows

$$\langle \beta \rangle = \frac{1}{S} \sum_s \beta_s, \quad (3.54)$$

$$\langle \beta x \rangle = \frac{1}{S} \sum_s \beta_s x_s, \quad (3.55)$$

and

$$\langle \beta x^2 \rangle = \frac{1}{S} \sum_s \beta_s (x_s)^2. \quad (3.56)$$

To estimate α we extract, for each promoter p , all samples s for which the promoter has non-zero tag count n_s and we calculate $P(x|\alpha)$ for each of the expression profiles of these promoters. The total likelihood of α is then simply the product of $P(x|\alpha)$ over all promoters

$$L = \prod_p P(x^p|\alpha). \quad (3.57)$$

and we maximize this expression with respect to α .

3.4.7 Likelihood for a consecutive pair of promoters

The key quantity that we want to calculate is the probability that the expression profiles of two neighboring promoters are proportional. That is, that the ‘true’ expression of the one promoter is a constant times the expression of the other promoter. Mathematically, we assume that the means of the log-expressions may be different for the two promoters, but the deviations δ_s are the same. That is, we assume

$$x_s = \text{noise} + \mu + \delta_s \quad (3.58)$$

and

$$y_s = \text{noise} + \tilde{\mu} + \delta_s, \quad (3.59)$$

where x_s and y_s are the log-expression values of the neighboring pair of promoters. Again, as described above, we restrict ourselves to those samples for which at least one of the neighbors has non-zero expression, and add a pseudo-count of half a tag whenever $n_s = 0$.

Constructing the human and mouse promoterome with deepCAGE data

For a single sample we have

$$\begin{aligned}
 P(x_s, y_s, \delta_s | \mu, \tilde{\mu}, \alpha) &= \\
 &= \sqrt{\frac{w_s \tilde{w}_s \alpha}{(2\pi)^3}} \cdot \exp \left[-\frac{w_s}{2} (x_s - \mu - \delta_s)^2 - \frac{\tilde{w}_s}{2} (y_s - \tilde{\mu} - \delta_s)^2 - \frac{\alpha}{2} (\delta_s)^2 \right],
 \end{aligned} \tag{3.60}$$

where

$$\tilde{w}_s = \frac{1}{\sigma^2 + 1/m_s}, \tag{3.61}$$

and m_s is the raw count of tags for the promoter with log-expression y_s . The integral over δ_s is still a Gaussian integral but the algebra is quite a bit more tedious in this case. To simplify the expression we write

$$\delta x_s = x_s - \mu \tag{3.62}$$

and

$$\delta y_s = y_s - \tilde{\mu} \tag{3.63}$$

Then we can write

$$\begin{aligned}
 P(x_s, y_s | \mu, \tilde{\mu}, \alpha) &= \\
 &= \sqrt{\frac{w_s \tilde{w}_s \alpha}{(2\pi)^2 (w_s + \tilde{w}_s + \alpha)}} \cdot \exp \left[-\frac{w_s}{2} (\delta x_s)^2 - \frac{\tilde{w}_s}{2} (\delta y_s)^2 + \frac{(w_s \delta x_s + \tilde{w}_s \delta y_s)^2}{2(w_s + \tilde{w}_s + \alpha)} \right].
 \end{aligned} \tag{3.64}$$

Next we want to integrate over μ and $\tilde{\mu}$. That is, we want to calculate the integrals

$$\int \prod_s P(x_s, y_s | \mu, \tilde{\mu}, \alpha) P(\mu) P(\tilde{\mu}) d\mu d\tilde{\mu}, \tag{3.65}$$

where we again use uniform priors

$$P(\mu) = \frac{1}{\Delta_\mu}. \tag{3.66}$$

Although these integrals are still just Gaussian integrals, the algebra is much more involved. To do the integrals we change variables from μ and $\tilde{\mu}$ to $r = (\mu + \tilde{\mu})/2$ and

3.3.4 Materials and Methods

$q = \mu - \tilde{\mu}$ (the Jacobian determinant of this transformation is 1). We integrate r out of the problem first. Furthermore we introduce notation

$$\sigma_s = \frac{x_s + y_s}{2}, \quad (3.67)$$

$$z_s = x_s - y_s, \quad (3.68)$$

$$\rho_s = \frac{w_s - \tilde{w}_s}{2(w_s + \tilde{w}_s)}, \quad (3.69)$$

$$u_s = \sigma_s + \rho_s(z_s - q), \quad (3.70)$$

$$\gamma_s = \frac{\alpha(w_s + \tilde{w}_s)}{\alpha + w_s + \tilde{w}_s}, \quad (3.71)$$

and finally

$$W_s = \frac{w_s \tilde{w}_s}{w_s + \tilde{w}_s + \alpha} \left(1 + \frac{\alpha}{w_s + \tilde{w}_s} \right). \quad (3.72)$$

Using all this notation we can write the integral over r as

$$P(x, y|q, \alpha) = \frac{1}{(\Delta_\mu)^2} \sqrt{\frac{2\pi}{S\langle\gamma\rangle} \prod_s \frac{\alpha w_s \tilde{w}_s}{(2\pi)^2 (w_s + \tilde{w}_s + \alpha)}} \exp \left[-\frac{1}{2} \left(\sum_s W_s (z_s - q)^2 + S\langle\gamma u^2\rangle - S \frac{\langle\gamma u^2\rangle}{\langle\gamma\rangle} \right) \right], \quad (3.73)$$

where the averages are again defined as

$$\langle\gamma\rangle = \frac{1}{S} \sum_s \gamma_s, \quad (3.74)$$

$$\langle\gamma u\rangle = \frac{1}{S} \sum_s \gamma_s u_s, \quad (3.75)$$

and

$$\langle\gamma u^2\rangle = \frac{1}{S} \sum_s \gamma_s (u_s)^2. \quad (3.76)$$

Finally, we integrate over q . The result can be written as

$$P(x, y|\alpha) = \frac{2\pi}{S(\Delta_\mu)^2} \frac{1}{\sqrt{\langle\gamma\rangle\langle W\rangle + \langle\gamma\rangle\langle\gamma\rho^2\rangle - \langle\gamma\rho\rangle^2}} e^{-SQ/2} \prod_s \sqrt{\frac{\alpha w_s \tilde{w}_s}{(2\pi)^2 (w_s + \tilde{w}_s + \alpha)}}, \quad (3.77)$$

Constructing the human and mouse promoterome with deepCAGE data

with

$$Q = \langle Wz^2 \rangle + \langle \gamma(\sigma + \rho z)^2 \rangle - \frac{\langle \gamma(\sigma + \rho z) \rangle^2}{\langle \gamma \rangle} - \frac{\left[\langle Wz \rangle + \langle \gamma\rho(\sigma + \rho z) \rangle - \frac{\langle \gamma\rho \rangle \langle \gamma(\sigma + \rho z) \rangle}{\langle \gamma \rangle} \right]^2}{\langle W \rangle + \langle \gamma\rho^2 \rangle - \frac{\langle \gamma\rho \rangle^2}{\langle \gamma \rangle}}, \quad (3.78)$$

and all the averages are defined as above. For example, we have

$$\langle \gamma\rho(\sigma + \rho z) \rangle = \frac{1}{S} \sum_s \gamma_s \rho_s (\sigma_s + \rho_s z_s), \quad (3.79)$$

and analogously for all the other averages.

3.4.8 Classifying high- and low-CpG promoters

We first log-transformed the CG- and CpG-contents of all promoters. To do this we added a pseudo-count of 0.05 to the fraction of CpG dinucleotides of all TSRs. We fitted (using expectation-maximization) the joint distribution of log-CG and log-CpG contents of all TSRs to a mixture of two two-dimensional Gaussians of the form

$$P(\vec{x}) = \frac{\rho}{2\pi\sigma_{\text{AT}}^2} \exp \left[-\frac{1}{2\sigma_{\text{AT}}^2} |\vec{x} - \vec{\mu}_{\text{AT}}|^2 \right] + \frac{1-\rho}{2\pi\sigma_{\text{CG}}^2} \exp \left[-\frac{1}{2\sigma_{\text{CG}}^2} |\vec{x} - \vec{\mu}_{\text{CG}}|^2 \right], \quad (3.80)$$

where the components of \vec{x} are the logarithms of the fraction of CGs and CpGs respectively. The fitted solution has

$$\rho = 0.55. \quad (3.81)$$

The center of the low-CpG Gaussian is given by

$$\vec{\mu}_{\text{AT}} = (-0.78, -2.74) \quad (3.82)$$

and the center of the high-CpG Gaussian by

$$\vec{\mu}_{\text{CG}} = (-0.39, -1.95). \quad (3.83)$$

The fitted variance of the low-CpG Gaussian is given by

$$\sigma_{\text{AT}}^2 = 0.036, \quad (3.84)$$

and the fitted variance of the high-CpG Gaussian is given by

$$\sigma_{\text{CG}}^2 = 0.026. \quad (3.85)$$

Using the fitted mixture of Gaussians we can calculate, for each TSR at position \vec{x} the posterior probability that it belongs to the low-CpG class as

$$P(\text{low}|\vec{x}) = \frac{G_{\text{AT}}(\vec{x})\rho}{G_{\text{AT}}(\vec{x})\rho + G_{\text{CG}}(\vec{x})(1 - \rho)}, \quad (3.86)$$

where $G_{\text{AT}}(\vec{x})$ and $G_{\text{CG}}(\vec{x})$ are the fitted low-CpG and high-CpG Gaussians, respectively.

3.4.9 Data availability

The complete human and mouse promoteromes, including the locations of all transcription start sites (TSSs), transcription start clusters (TSCs), transcription start regions (TSRs), and their raw and normalized expression profiles across all CAGE samples are available for download from the SwissRegulon web page http://www.swissregulon.unibas.ch/cage_clustering_supplementary/.

3.5 Supplementary Data

3.5.1 Distributions of reads per position for Solexa RNA-seq data

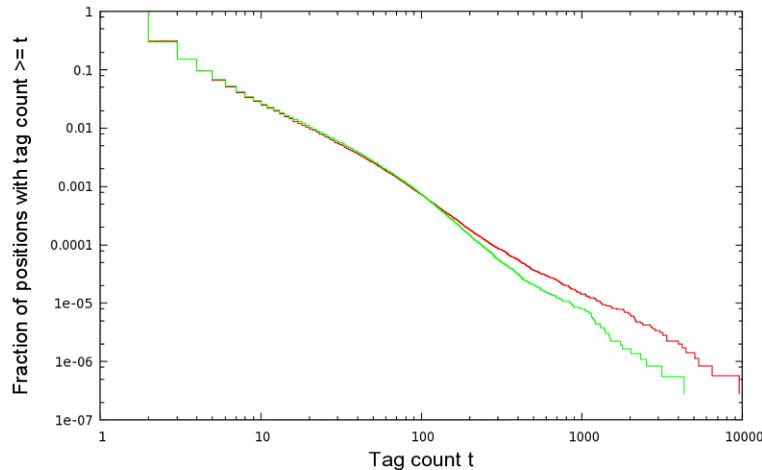


Figure 3.16: Reverse cumulative distributions for the number of reads per position in two RNA-Seq technical replicates of *Drosophila* Kc cells. Both axes are shown on logarithmic scales.

Using Solexa sequencing we obtained two replicate data-sets of RNA-seq data. After mapping the reads to the genome we determined the distribution of the number of reads per position for each replicate. Figure 3.16 shows the reverse cumulative distributions of reads per position that we obtained for these data sets. The figure illustrates that approximately power-law distributions are observed for RNA-seq data as well. This further supports that the roughly power-law distribution of expression levels across individual TSSs is not an artefact of measurement technology but represents the actual distribution of transcript levels in the cells.

3.5.2 Replicate scatter for Solexa RNA-seq data

For the same two RNA-seq samples figure 3.17 shows a scatter-plot of the number of reads per position in the two samples.

3.5.3 Per ‘exon’ replicate scatter for Solexa RNA-seq data

For the same data-set shown in figure 3.17 we used single-linkage clustering to cluster overlapping reads into ‘exons’. Figure 3.18 shows a scatter plot analogous to figure

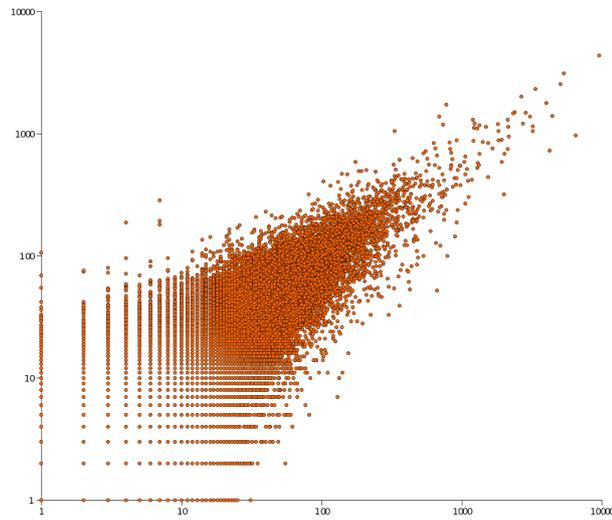


Figure 3.17: Scatter-plot of numbers of reads in the two RNA-seq replicates of *Drosophila Kc* cells obtained with Solexa sequencing. Each data point corresponds to a unique position on the chromosome with the number of reads in the first replicate on the horizontal axis and the number of reads in the second replicate on the vertical axis. Both axes are shown on a logarithmic scale. The size of the multiplicative noise σ^2 estimated from this scatter is $\sigma^2 = 0.073$

3.17 but now for the expression of these ‘exons’ across the two replicates.

3.5.4 CAGE per TSS replicate scatter

Two independent CAGE samples were obtained from a common RNA sample from THP-1 cells after 8 hours of treatment with LPS. Figure 3.19 shows a scatter-plot of the normalized tags-per-million of each TSS for these two replicate samples.

3.5.5 CAGE per gene replicate scatter

For the same two replicate samples shown in figure 3.19 we summed, for each gene, the expression from all TSSs associated with the gene, to obtain a normalized expression per gene. Figure 3.20 shows a scatter-plot of the per gene expression of the CAGE replicates.

3.5.6 Comparison with FANTOM3 clustering

For human our data contained a total of 25,469,648 CAGE tags representing 6,395,686 unique TSS locations in the human genome. Table 3.2 compares the number of TSSs

Constructing the human and mouse promoterome with deepCAGE data

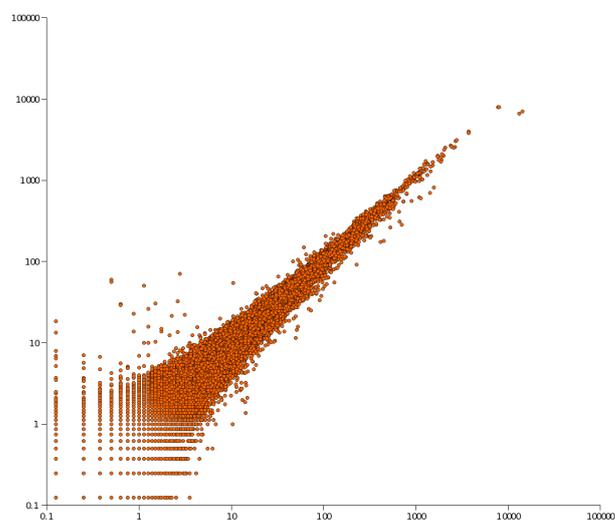


Figure 3.18: Scatter-plot of reads per million in two RNA-Seq replicates of *Drosophila* Kc cells. Each data point corresponds to a cluster of overlapping reads on the chromosome, with horizontal and vertical coordinates given by the number of reads per million for each replicate. The size of the multiplicative noise σ^2 estimated from this data is $\sigma^2 = 0.02$.

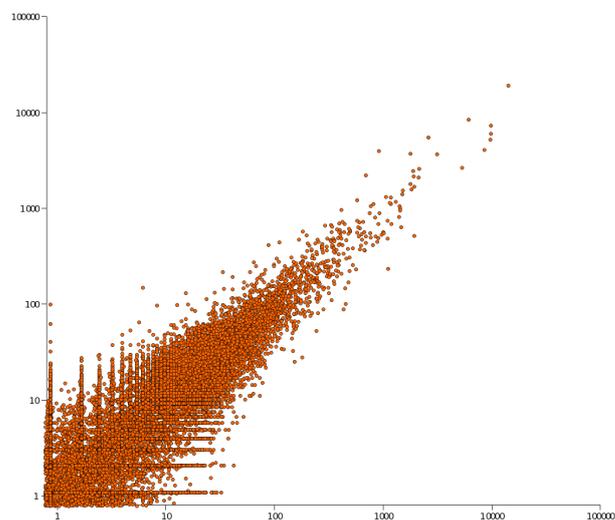


Figure 3.19: Scatter-plot of CAGE expression for two replicate measurements of THP-1 cells after 8 hours of LPS treatment. Each data point corresponds to an individual TSS. Values on the horizontal and vertical axes correspond to normalized tags per million for each TSS. Both axes are shown on a logarithmic scale. The size of the multiplicative noise σ^2 estimated from this data is $\sigma = 0.085$.

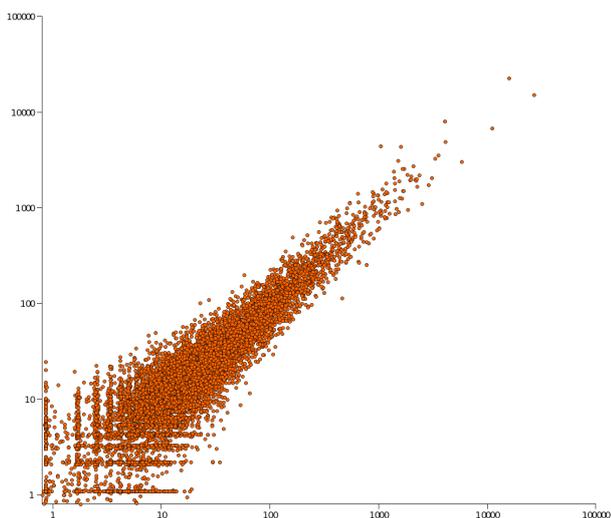


Figure 3.20: Scatter-plot of the normalized tags per million *per gene* for the same two CAGE replicates as shown in Fig. 3.19. Each data point corresponds to a gene. Axes are shown on logarithmic scales. The size of the multiplicative noise σ^2 estimated from this data is $\sigma = 0.068$

Statistic	Our clustering	FANTOM3 clustering
Number of TSSs in TSCs	860'823	1'043'768
Number of TSCs	74'273	64'908
Number of TSRs	43'164	49'461

Table 3.2: Comparison of the number of TSSs, TSCs, and TSRs obtained with our clustering and the FANTOM3 clustering (in which CAGE tags that are 21 bp or less apart are clustered through single-linkage clustering).

Constructing the human and mouse promoterome with deepCAGE data

in TSCs, the number of TSCs, and the number of TSRs between our clustering of CAGE tags and the simple single-linkage clustering employed in the FANTOM3 paper. First of all we see that a significantly larger number of unique TSSs are included in the FANTOM3 clustering. This is a result of the fact that TSSs with expression profiles significantly different from those in the TSC (which may often be low expressed TSSs) are clustered with the TSC in the FANTOM3 clustering, whereas in our clustering these form separate TSCs who are then filtered out owing to their low expression. The total number of TSCs in the FANTOM3 clustering is lower because neighboring TSCs with different expression profiles are all clustered together in the FANTOM3 clustering. Even though the number of TSCs is smaller in the FANTOM3 clustering, the final number of TSRs is a little larger because, owing to the tendency of the FANTOM3 clustering to cluster all nearby TSSs, irrespective of their expression profile, a large number of low expressed TSRs pass the cut-off on minimal expression in the filtering stage.

Figure 3.21 shows a comparison of the distributions of the number of TSSs per TSC, the number of TSCs per TSR, and the number of TSSs per TSR, for our clustering and for the single-linkage clustering that was employed in FANTOM3.

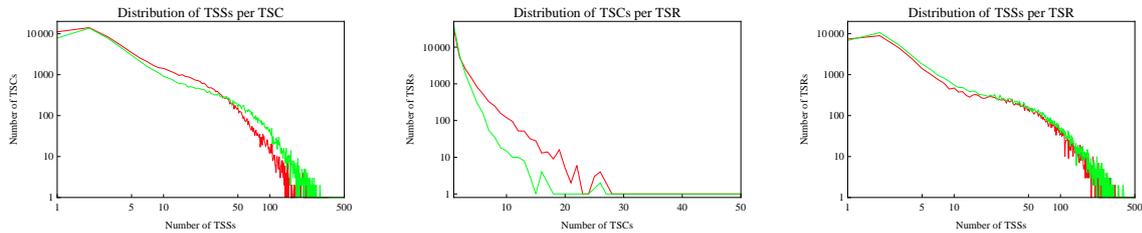


Figure 3.21: Comparison of the hierarchical structure of the human promoterome for our clustering and the FANTOM3 clustering. Left: Distribution of the number of transcription start sites (TSSs) per per co-expressed transcription start cluster (TSC). Middle: Distribution of the number of TSCs per transcription start region (TSR). Right: Distribution of the number of TSSs per TSR. The vertical axis is shown on a logarithmic scale in all panels. The horizontal axis is shown on a logarithmic scale in the left and right panels. The red lines show the distributions obtained using our clustering procedure and the green lines show the distribution obtained using single-linkage clustering employed in FANTOM3.

As illustrated by the left and right panels of figure 3.21, there are in general more TSSs per TSC and more TSSs per TSR for the FANTOM3 clustering. In contrast, there tend to be more TSCs per TSR for our clustering. Both these observations are a result of the fact that in our clustering TSSs with different expression profiles are not clustered together, even if they are near each other, whereas the single-linkage clustering fuses all these TSSs into a single TSC.

3.3.5 Supplementary Data

Figure 3.22 shows the distributions of the lengths TSCs and TSRs for both our clustering and the FANTOM3 clustering. Although on the logarithmic scales the

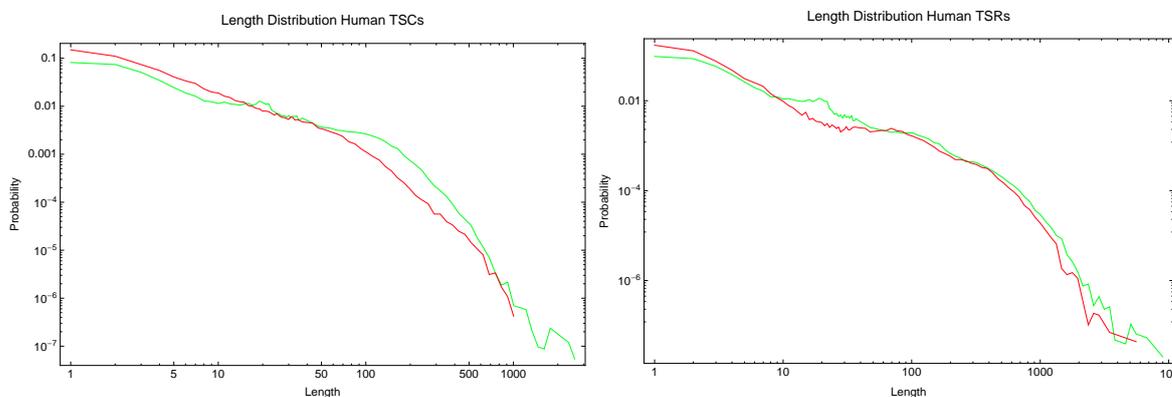


Figure 3.22: Comparison of the length distributions of TSCs and TSRs for the promoteromes obtained using our clustering and using the FANTOM3 clustering. Left: Length distribution of the TSCs. Right: Length distribution of the TSRs. Both axes are shown on logarithmic scales. The red lines show the distributions obtained using our clustering procedure and the green lines show the distribution obtained using single-linkage clustering employed in FANTOM3.

length distributions appear quite similar for the two clustering procedures, the TSCs obtained by the FANTOM3 clustering tend to be significantly wider. More strikingly, for the FANTOM3 clustering there is a pronounced shoulder in the distributions at a width of 21 base pairs, which is almost certainly an artifact of the fact that this distance is exactly the cut-off on the single-linkage clustering.

3.5.7 Nearby uncorrelated TSSs

In figure 12 of the main article we showed an example of neighboring TSCs that have significantly different expression profiles, which were shown in panel C. To further illustrate that these expression profiles are indeed not correlated figure 3.23 shows a scatter plot of the expression of the two TSCs across the 56 CAGE samples. The plot confirms that there is no discernible correlation between the expression profiles of the two TSCs, and they are certainly not tightly co-regulated, which supports that these two TSCs are driven by distinct regulatory sites.

In figure 3.24 below we show another example of a set of nearby TSCs with clearly distinct expression profiles. The interesting feature of this example is that there are two broad TSCs, containing a substantial number of TSSs that all show correlated expression, which are interspersed by a *single* TSS that shows a very different expression profile (the red TSS). The structure of this promoter region suggests that,

Constructing the human and mouse promoterome with deepCAGE data

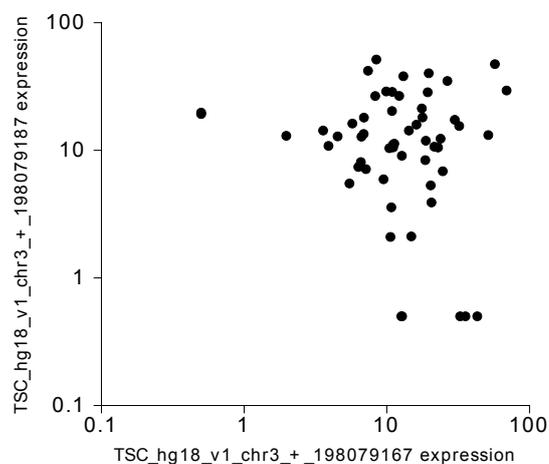


Figure 3.23: Scatter of the expression levels (in TPM) of two nearby TSCs, located on human chromosome 3. Each dot corresponds to one of the 56 human CAGE samples. Both axes are shown on a logarithmic scale.

on the one hand, there is a broad region to which the polymerase is recruited by one set of regulatory mechanisms, while on the other hand there is a single TSS within the same region to which the polymerase is recruited by a distinct regulatory mechanism.

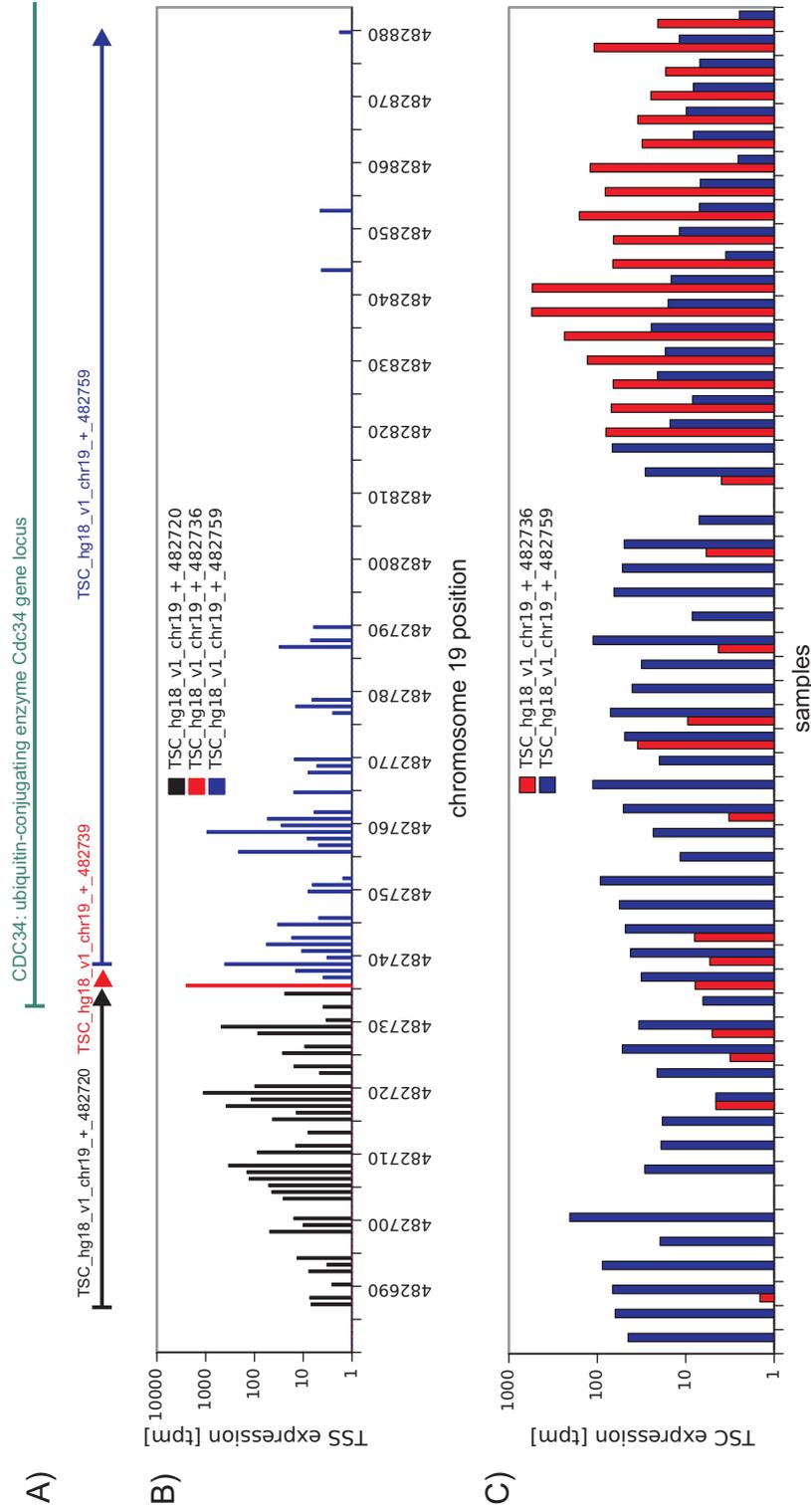


Figure 3.24: Nearby TSCs with significantly differing expression profiles. (A) An approximately 200 base pair region on chromosome 9 containing 3 TSCs (colored segments) and the start of the annotated locus of the CDC34 gene (black segment). (B) Positions of the individual TSSs in the TSC and their total expression, colored by the TSC to which each TSS belongs. (C) expression across the 56 CAGE samples for the red and blue TSCs.

3.5.8 Mouse Promoterome Statistics

For the mouse promoterome, as for the human promoterome, we first calculated the distribution of phastCons conservation scores as a function of position relative to the most expressed TSS in each TSC. Figure 3.25 shows the phastCons conservation profiles that we obtained for both all TSCs (left panel) and the novel TSCs (right panel).

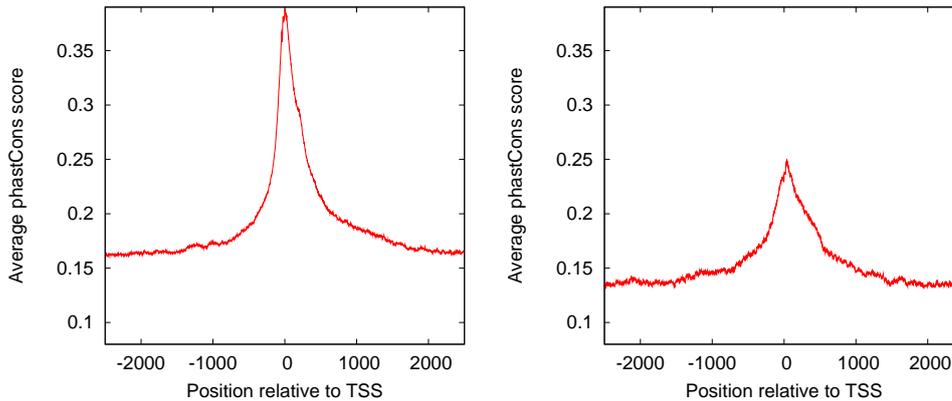


Figure 3.25: Average phastCons (conservation) score relative to TSS of genomic regions upstream and downstream of all mouse TSCs (left panel) and for all mouse TSCs that are more than 5 kilobases away from any known start (right panel).

The conservation profiles for mouse are very similar to the ones that we observed for human. We again see a sharp peak of conservation covering a few hundred base pairs around TSS. The novel promoters show a conservation peak of similar width but with lower height. Interestingly, whereas for human the conservation peak of the novel promoters was close to symmetric, for mouse the novel promoter peak is also clearly asymmetric, although still not as asymmetric as the peak for the known TSSs.

Next we determined the position of the closest start of a known transcript for each mouse TSC. Figure 3.26 shows the distribution of the relative positions of the closest known starts for all mouse TSCs that have a known start within 1000 base pairs of the TSC.

The distribution in figure 3.26 is also very similar to what we observed for the human promoterome. The main difference is that whereas for human 62.2% of all TSCs have a known start within 1000 base pairs, for mouse this is only 59%, which is likely due to the larger amount of data available for human.

Figure 3.27 shows the hierarchical structure of the mouse promoterome that we constructed. In particular, we show the distribution of the number of TSSs per TSC, the number of TSCs per TSR, and the number of TSSs per TSR, as we also showed for the human promoterome in the main article.

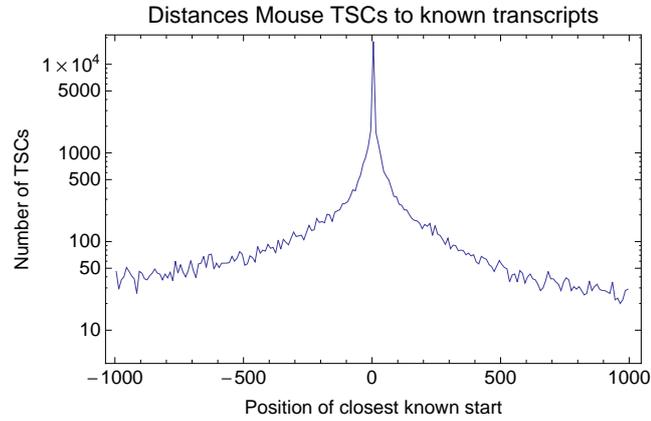


Figure 3.26: Number of TSCs as a function of their position relative to the nearest known transcript start. Negative numbers mean the nearest known start is upstream of the TSC. The vertical axis is shown on a logarithmic scale. The figure shows only the 45,603 TSCs (59%) with a known start within 1000 base pairs.

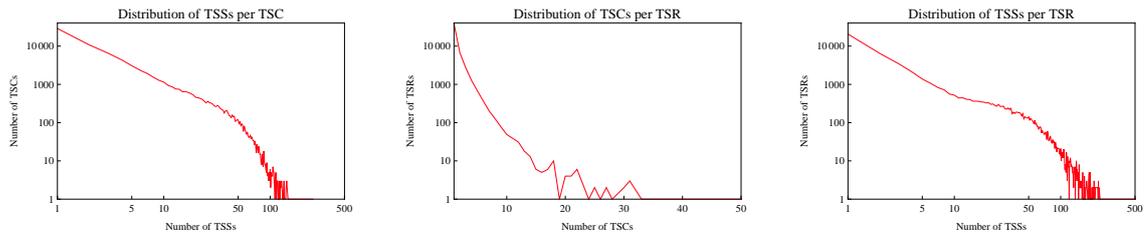


Figure 3.27: Hierarchical structure of the mouse promoterome. Left: Distribution of the number of transcription start sites (TSSs) per per co-expressed transcription start cluster (TSC). Middle: Distribution of the number of TSCs per transcription start region (TSR). Right: Distribution of the number of TSSs per TSR. The vertical axis is shown on a logarithmic scale in all panels. The horizontal axis is shown on a logarithmic scale in the left and right panels.

Constructing the human and mouse promoterome with deepCAGE data

The distributions in Fig. 3.27 are generally very similar to those observed for the human promoterome. The distributions are all a little less wide than for human, which is likely the result of the larger amount of data available for human. Importantly, as in the human data, the distribution of the number of TSSs per TSR also shows the clear ‘shoulder’ corresponding to TSRs with between roughly 10 and 50 TSSs.

Finally, we also calculated the length distributions of mouse TSCs and TSRs, both using our clustering procedure, and using the single-linkage clustering employed in FANTOM3 (figure 3.28). Here too the distributions are very similar to the results that we obtained for the human data. In particular, we clearly see the shoulder in the distribution of TSR lengths for lengths roughly between 25 and 150 base pairs long. We also again see that the single-linkage clustering leads to wider clusters, and leads to an artificial shoulder at 21 base pairs (i.e. the length of the CAGE tags that was chosen as a distance cut-off).

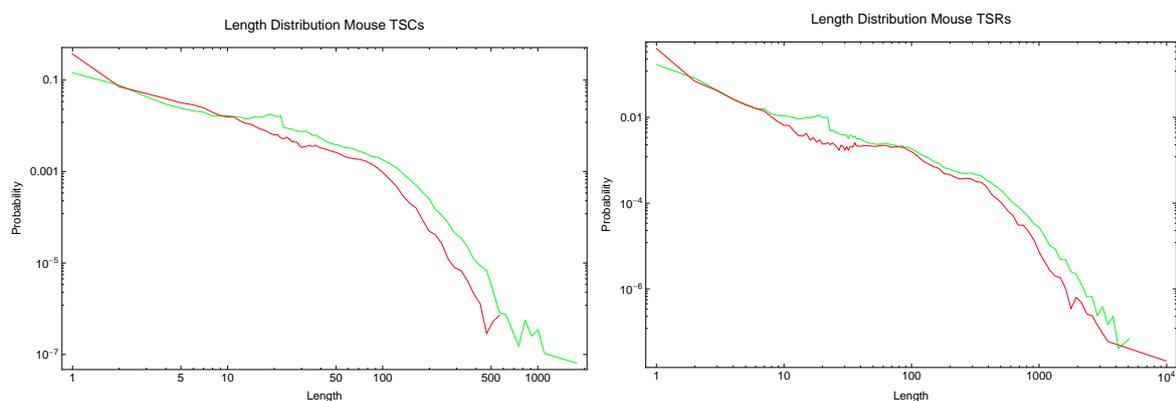


Figure 3.28: Comparison of the length distributions of TSCs and TSRs for the mouse promoteromes obtained using our clustering and using the FANTOM3 clustering. Left: Length distribution of the TSCs. Right: Length distribution of the TSRs. Both axes are shown on logarithmic scales. The red lines show the distributions obtained using our clustering procedure and the green lines show the distribution obtained using single-linkage clustering employed in FANTOM3.

Chapter 4

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan & Erik van Nimwegen
submitted

Accurate reconstruction of the regulatory networks that control gene expression is one of the key current challenges in molecular biology. Although gene expression and chromatin state dynamics are ultimately encoded by constellations of binding sites recognized by regulators such as transcription factors (TFs) and microRNAs (miRNAs), our understanding of this regulatory code and its context-dependent read-out remains very limited. Given that there are thousands of potential regulators in mammals, it is not practical to use direct experimentation to identify which of these play a key role for a particular system of interest.

We developed a methodology that uses genome-wide predictions of TF binding sites and miRNA target sites to model gene expression or chromatin modifications in terms of these sites, and completely automated it into a web-based tool called ISMARA (Integrated System for Motif Activity Response Analysis), located at <http://ismara.unibas.ch>. Given as input only gene expression or chromatin state data across a set of samples, ISMARA identifies the key TFs and miRNAs driving expression/chromatin changes and makes detailed predictions regarding their regulatory roles. These include predicted activities of the regulators across the samples, their genome-wide targets, enriched gene categories among the targets, and direct interactions between the regulators.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Applying ISMARA to data sets from well-studied systems, we show that it consistently identifies known key regulators *ab initio*. We also present a number of novel predictions including regulatory interactions in innate immunity, a master regulator of mucociliary differentiation, TFs consistently upregulated in cancer, and TFs that mediate specific chromatin modifications.

4.1 Introduction

Since the seminal work of Jacob and Monod (32), much has been learned about the molecular mechanisms by which gene expression is regulated, and the molecular components involved. Historically, most work has focused on transcription factors (TFs), arguably the most important regulators of gene expression, which bind to cognate sites in the DNA, frequently in the neighborhood of transcription start sites (TSSs), and regulate the rate of transcription initiation. However, more recently it has become clear that the state of the chromatin, which can be modulated through modifications of the DNA nucleobases and of the histone tails of nucleosomes, also plays a crucial role. For example, the local chromatin state affects the ability of TFs to access their binding sites, and the chromatin state can in turn be modified through TF-guided recruitment of chromatin modifying enzymes. Furthermore, an entirely new layer of post-transcriptional regulation has been uncovered in recent years in the form of microRNAs (miRNAs) (33). These guide RNA-induced silencing complexes to target mRNAs, inhibiting their translation and accelerating their decay (34).

In spite of these many insights, our current understanding of the function of genome-wide gene regulatory networks in mammals is still rudimentary. For example, we only know the sequence specificity of less than 500 (35; 36; 37) of the approximately 1500 (38) TFs in mammalian genomes. Our knowledge of how TF binding is affected by chromatin state, of the combinatorial interactions between TFs and their co-factors, and the impact of post-translational modifications on TF activity, is even more fragmentary. Our understanding of the transcriptome-wide effects of miRNAs on their targets is similarly limited. It is thus clear that we are still far from being able to develop realistic quantitative models of gene regulatory networks in mammals. Consequently, rather than aiming to develop comprehensive computational models of gene regulatory dynamics, the most constructive contribution that computational approaches can currently provide is to develop models that help guide experimental efforts.

Given a particular mammalian subsystem or process, e.g. a particular developmental or cellular differentiation process, or the response of a tissue to a particular perturbation, the initial steps in unraveling its gene regulatory circuitry are to identify the key regulators in the process, and to characterize the rough functional roles

these regulators play. However, for the vast majority of mammalian systems these initial steps have yet to be taken. Given the large number of potential regulators, a direct experimental approach, e.g. through large-scale screening, is typically not feasible. There is thus a strong need for computational methods that, given a system of interest, can predict key regulatory players and make concrete, directly testable, hypotheses about their regulatory roles. We here present an integrated and completely automated computational methodology that accomplishes exactly this task.

Our approach, ISMARA (Integrated System for Motif Activity Response Analysis), capitalizes on a number of recent computational and experimental technological developments. First, whereas large-scale screening of the functions of individual regulators in a particular system is often impractical, it is relatively straight forward to measure gene expression (i.e. with microarray or RNA-seq) or chromatin state (with ChIP-seq) in high-throughput across a set of samples of interest. Second, over the last years sophisticated comparative genomic methods have been developed that allow relatively accurate computational prediction of regulatory sites for hundreds of TFs and miRNAs on a genome-wide scale (39; 40; 41). Third, through extensive experimental efforts, genome-wide annotations of transcript structures (13) and promoters (Chapter 3) have also become available.

Given as input a set of genome-wide gene expression or chromatin state measurements across a number of samples, ISMARA models the gene expression or chromatin state dynamics in terms of a comprehensive set of computationally predicted regulatory sites, using a simple linear modeling approach called Motif Activity Response Analysis (MARA) that we originally proposed in (25). As a result, ISMARA identifies the key regulators (i.e. TFs and miRNAs) driving gene expression/chromatin state changes across the samples, the activity profiles of these regulators, their target genes, and the sites on the genome through which these regulators act. The analysis is carried out within a completely automated system, which combines pre-calculated annotations of regulatory sites for hundreds of regulators across promoters in mammalian genomes with processing of input data, automated tuning of parameters, and post-processing to provide a large collection of auxiliary analysis results. To use ISMARA, all that users need to do is upload their data to the web-server <http://ismara.unibas.ch/> and submit it to the system, after which all results are presented through a user-friendly graphical web-interface. Importantly, in ISMARA the motif activity response analysis has been extended to model not only gene expression data from various platforms (microarray, RNA-seq), but essentially any sequencing data reflecting a genomic mark (ChIP-seq) including chromatin modifications or TF binding. In addition, ISMARA models not only the effect of TFs on mammalian gene expression, but also the effect of miRNAs. Below we will first describe the methodology used by ISMARA and the results that it provides, and then we will demonstrate its power through a number of applications.

4.2 Results

4.2.1 An Integrated System for Motif Activity Response Analysis

The integrated system for motif activity response analysis (ISMARA) that we developed is schematically depicted in Fig. 4.1. Detailed descriptions of all procedures are provided in the supplementary methods. The system capitalizes on two key resources developed in our group (Fig. 4.1A-C). The first is the genome-wide annotation of promoters in human and mouse, i.e. so-called "promoteromes", that we constructed (Chapter 3) from genome-wide transcription start site data (deepCAGE data (42)). We supplement these promoter sets with 5' ends of known RNA transcripts from human and mouse, and associate transcripts with promoters. The second key resource that we employ is a genome-wide annotation of functional transcription factor binding sites (TFBSs) that we obtained with Bayesian probabilistic methods for quantifying evolutionary selection pressure, which we developed previously (39; 41). Briefly, we constructed multiple alignments of orthologous proximal promoter regions across 7 mammalian genomes and curated a collection of approximately 200 non-redundant mammalian regulatory motifs (positional weight matrices) that represent the DNA binding specificities of close to 350 TFs in both human and mouse. We then used our MotEvo algorithm (41) to predict functional TFBSs for all TF regulatory motifs across all promoters in human and mouse (Fig. 4.1A,C). MotEvo is a Bayesian algorithm which explicitly models the evolution of TFBSs across the mammalian phylogeny (Suppl. Fig. 1).

When modeling expression data, ISMARA also integrates the effects of miRNAs that increase decay of transcripts by binding to sites that are generally located in the 3' untranslated regions (UTRs) of transcripts. We used miRNA target site predictions from TargetScan using preferential conservation scoring (P_{CT}) (40), and calculated an overall score for the targeting of a promoter by a particular miRNA by averaging over all transcripts associated with the promoter (Suppl. Methods).

The result of the regulatory site annotation was, for both human and mouse, a large matrix \mathbf{N} , where N_{pm} is the predicted total number of functional binding sites in promoter p for motif m , where m runs over the 190 TF binding motifs as well as the 86 miRNA 'seed' motifs.

The next step in ISMARA consists of the construction of a data matrix \mathbf{E} , where E_{ps} denotes the 'signal' associated with promoter p for sample s . When provided gene expression data in the form of microarrays, ISMARA applies standard normalization procedures and maps the probes on the microarray to the set of known RNA transcripts, which are each in turn associated with promoters. Microarray platforms currently supported by ISMARA are listed in Suppl. Table 1. For a promoter p , the expression E_{ps} is given by the average log-intensity in sample s of the probes asso-

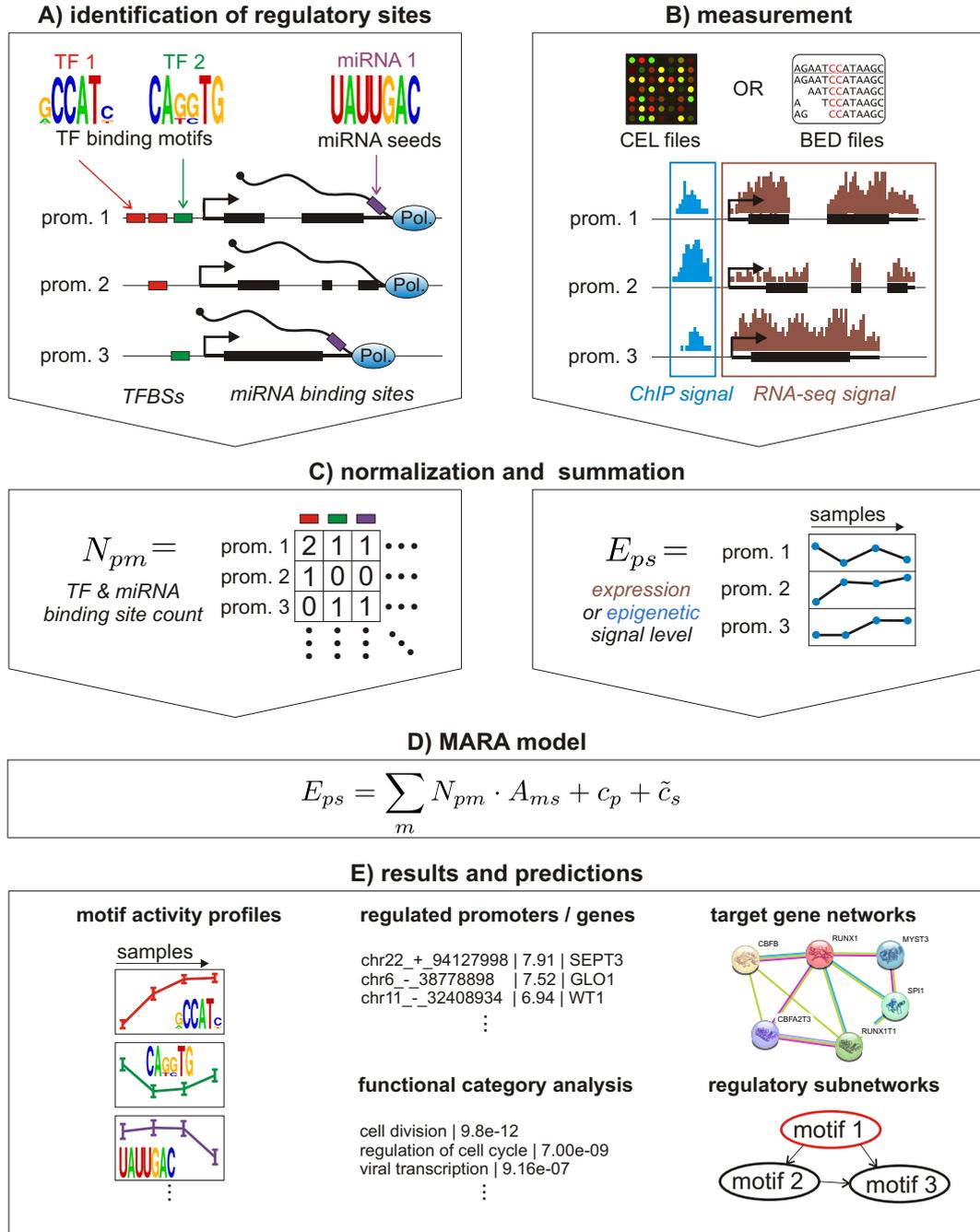


Figure 4.1: Outline of the Integrated System for Motif Activity Response Analysis (ISMARA) (continued on the next page)

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Figure 4.1: Outline of the Integrated System for Motif Activity Response Analysis (ISMARA). **A:** ISMARA starts from a curated genome-wide collection of promoters and their associated transcripts. Using a comparative genomic Bayesian methodology (41), transcription factor binding sites (TFBSs) for ≈ 200 regulatory motifs are predicted in proximal promoters. Similarly, miRNA target sites for ≈ 100 seed families are annotated in the 3' UTRs of transcripts associated with each promoter. **B:** Users provide measurements of gene expression (microarray, RNA-seq) or chromatin state (ChIP-seq). The raw data are processed automatically and, for each promoter and each sample, a signal is calculated. For ChIP-seq data, the signal is calculated from the read density in a region around the transcription start. For gene expression data, the expression signal is calculated from read densities across the associated transcripts (RNA-seq) or intensities of associated probes (microarray). **C:** The site predictions and measured signals are summarized in two large matrices. The components N_{pm} of matrix **N** contain the total number of sites for motif m (TF or miRNA) associated with promoter p . The components E_{ps} of matrix **E** contain the signal associated with promoter p in sample s . **D:** The linear MARA model is used to explain the signal levels E_{ps} in terms of bindings sites N_{pm} and unknown motif activities A_{ms} , which are inferred by the model. The constants c_p and \tilde{c}_s correspond to basal levels for each promoter and sample, respectively. **E:** As output, ISMARA provides the inferred motif activity profiles A_{ms} of all motifs across the samples s , sorted by the significance of the motifs. A sorted list of all predicted target promoters is provided for each motif, together with the network of known interactions between these targets (provided by the String database, <http://string-db.org/>), and a list of Gene Ontology categories that are enriched among the predicted targets. Finally, for each motif, a local network of predicted direct regulatory interactions with other motifs is provided.

ciated with promoter p . Similarly, for RNA-seq data the reads are mapped to the known RNA transcripts and E_{ps} is calculated as the average of the logarithm of the fraction of all reads in the sample that map to transcripts associated with promoter p . When processing ChIP-seq data, the signal E_{ps} is calculated as the logarithm of the fraction of reads in sample s that map to a 2 kilobase region centered on promoter p . Details of the normalization steps involved are again provided in the supplementary methods.

At the core of ISMARA is the MARA model (25) which, similar to previous linear modeling approaches (4; 43), assumes that the ‘signal’ at each promoter p is a linear function of its binding sites N_{pm} :

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \quad (4.1)$$

where c_p is a term reflecting the basal activity of promoter p , \tilde{c}_s reflects the total expression in sample s , and A_{ms} is the (unknown) *activity* of motif m in sample s . That is, using the predicted site counts N_{pm} and the experimentally measured E_{ps} , we use the model (4.1) to infer the activities A_{ms} of all motifs across all samples. To infer the activities, ISMARA uses a Bayesian procedure with a Gaussian likelihood model for the difference between the measured signal E_{ps} and the predicted signal, and a Gaussian prior distribution for the activities (Methods). The latter is used to avoid overfitting and ISMARA uses a cross-validation procedure to set the parameters of the prior (see Suppl. methods). The entire posterior distribution of motif activities is a multi-variate Gaussian which is determined using singular value decomposition (see Suppl. methods).

It is important to note that we do not expect the simple model (4.1) to provide an accurate fit to the signal E_{ps} at individual promoters. As mentioned in the introduction, many factors that influence expression and local chromatin state are not included in our model. Moreover, instead of each binding site contributing linearly to E_{ps} , in reality the expression E_{ps} will likely be a complex combinatorial function of the constellation of binding sites in promoter p . Indeed, we typically find that the simple model (4.1) captures only a small fraction of the variance of E_{ps} across the samples (Suppl. Fig. 2). However, the aim of the model (4.1) is not to fit the signals E_{ps} , but rather to identify which of the motifs m play an important role, and how these motifs contribute to E_{ps} across the samples. Since each motif m targets hundreds to thousands of promoters p , the inferred motif activities A_{ms} are statistical averages of the behaviors of a large number of promoters. This averaging causes the complexities at individual promoters to effectively cancel out and ensures that the overall influence of a motif can still be reliably inferred. To put it differently, if a clear average contribution of a given motif m is detected using the simple linear model (4.1) in spite of it being a poor model at individual promoters, we can be confident that the motif indeed contributes to the signal E_{ps} .

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Apart from inferring motif activities, ISMARA also predicts which individual promoters are regulated by each motif m . As detailed in the Suppl. methods, for each promoter with predicted TFBSs for the motif (i.e. $N_{pm} > 0$) ISMARA estimates the log-likelihood ratio S_{pm} of the entire model with the TFBSs for m in p present, and the model in which the entry N_{pm} has been set to zero. That is, S_{pm} rigorously quantifies how much removal of the sites for m in p decreases the fit of the model to the data.

4.2.2 Overview of the results presented by ISMARA

We have made ISMARA available through a web interface <http://www.ismara.unibas.ch> as part of our SwissRegulon resources (37). Users can directly upload unprocessed microarray (CEL files), RNA-seq, or ChIP-seq data (bed files) which are then analyzed automatically without the need for any additional input from the user (Fig. 4.1B). The results are made available through a web interface and can also be downloaded in flat-file format. To give an overview of the results ISMARA provides, we applied it to the GNF Gene Atlas (44) of mRNA expression profiles across 91 tissues and cell lines in mouse. The results are available at http://ismara.unibas.ch/supp/dataset1/ismara_report/.

The first output of ISMARA is a list of all regulatory motifs sorted by a z -score which summarizes the importance of the motif for explaining the expression variation across the samples. This score roughly corresponds to the average number of standard-deviations the motif activity is away from zero (see Methods and Suppl. Methods). Besides the z -score of each motif, the list also displays the set of associated TFs, a thumbnail of its activity across the input samples, and a sequence logo for each motif (Suppl. Fig. 3). In the Gene Atlas data, the second most significant motif is E2F1..5, corresponding to the E2F1 through E2F5 transcription factors that are known to regulate the cell cycle (45; 46). Following the link from the motif name links leads to a page with additional details regarding the E2F1..5 motif (Suppl. Figs. 4-6), including its inferred activity profile across the samples, once ordered according to the user's input (Suppl. Fig. 4), and once ordered according to the sample-dependent activity z -values (Suppl. Fig. 5). The samples in which the E2F activity is highest are known to be composed of fast dividing cells (bone marrow, hematopoietic stem cells and artificial cell lines), while neural tissues, containing largely non-dividing cells have the lowest E2F activity (Fig. 4.2A). The page also provides a list of predicted target promoters of the motif, sorted by their score S_{pm} (Suppl. Fig. 6). Besides the score, this list includes for each target a link to a genome browser view of the promoter which shows the predicted TFBSs (Suppl. Fig. 7), associated gene and transcripts, and a short description of each target gene. The user can interactively change how many of the top targets are shown, or search for a gene or transcript of interest in the list of all targets. To provide the user with a more intuitive picture

of the predicted list of targets of the motif, a link is provided to a network view of the target genes as provided by the STRING database (47), where network links indicate known functional associations between the genes. For E2F1..5, the STRING network reveals a large, highly connected cluster of predicted targets that are known to be involved in cell cycle, and particularly in DNA replication (Suppl. Fig. 8). The role of the E2F1..5 motif in the cell cycle is further confirmed by Gene Ontology analysis (48) which shows that DNA replication, S phase, and regulation of DNA replication are categories whose genes are most highly enriched among the targets of E2F (Suppl. Fig. 9). Thus, based only on expression data, MARA predicts E2F to be a key regulator of cell proliferation, with E2F activity acting effectively as a marker for proliferation.

For many of the regulatory motifs there are multiple TFs that can bind to the sites of the motif and it is not *a priori* clear which of the TFs is most responsible for the motif activity in a given system. Note that the motif activity is inferred from the behavior of the predicted *targets* of the motif. That is, roughly speaking, an increasing activity is inferred when its targets show on average an increase in expression, that cannot be explained by the presence of other motifs in their promoters. The mRNA expression profiles of the TFs associated with a motif thus provide independent information about the link between the TFs and the motif activities, and ISMARA provides an analysis of the correlation between motif activities and the expression profiles of the associated TFs for each motif. For example, for the case of E2F1..5, the expression of all associated TFs except E2F5 show a very significant positive correlation with the motif activities (Suppl. Figs. 10 and 11). This also shows that these TFs act as *activators*. That is, whenever a negative correlation between motif activity and TF expression is observed, the TF most likely acts as a repressor, e.g. as observed for the known repressor REST (Suppl. Fig. 12). However, it should be noted that motif activity does not need to be a direct function of TF expression, i.e. the effect of a TF on its targets will not only depend on its expression, but possibly on post-translational modifications, on cellular localization, and on the presence of specific co-factors. Therefore, although a strong correlation between TF expression and motif activity is a good indication that the TF is responsible for the motif activity, the absence of such a correlation does not imply that the TF is not involved in the motif's activity.

To gain insight in the transcription regulatory networks that control expression profiles, it is of particular interest to identify direct regulatory connections between the TFs themselves. In ISMARA, a predicted transcription regulatory interaction from motif m to m' occurs when motif m is predicted to target a promoter of one of the TFs associated with m' . To visualize predicted motif-motif regulatory networks ISMARA provides, for each motif m , a local network picture that shows all predicted regulatory connections between m and other regulatory motifs. The user can interactively change the cut-off on the target score S_{pm} to draw this picture. For E2F1..5 we find that

the strongest predicted targets are the promoters of Myb, of TFDP1, and of the E2F2 gene (Suppl. Fig. 13). Indeed, the c-Myb promoter is known to be regulated by an E2F site (49) and the E2F2 promoter has indeed been shown to be bound directly by E2F4 (50). The transcription factor TFDP1 forms hetero-dimers with various members of the E2F family and ISMARA predicts that this co-factor of the E2F family is itself regulated through an E2F site. To our knowledge, this is novel prediction.

An example of a motif with highly condition-specific activity is HNF1A (Fig. 4.2B). The associated transcription factor hepatocyte nuclear factor 1 homeobox A is relatively well-studied and known to be mainly expressed in liver, kidney, stomach and intestine (51; 52), where it is essential for organ function (53). Indeed, ISMARA infers that the HNF1A activity is by far the highest in liver and kidney, followed by intestinal tissues and stomach. In addition to its role in these tissues, HNF1A has also been shown to be important for the function of pancreatic islets, and HNF1A mutations causes monogenic diabetes (52). Indeed, ISMARA predicts high activity for HNF1A in pancreas as well, where its activity ranks 6th and 7th among all motifs in the two replicate samples (Suppl. Fig. 14). Figure 4.2B also illustrates that the inferred motif activities are highly reproducible, in fact more reproducible than the expression profiles from which the motif activities were inferred (Suppl. Fig. 15). The reason for this high reproducibility is that motif activity is inferred from the statistics of all (typically hundreds) of its target promoters.

Experiments are often performed in multiple replicates and one would typically be specifically interested in those motifs that behave reproducibly across the replicates. To this end the ISMARA results page links to a section where users can provide replicate annotation for their samples, which then enables ISMARA to calculate motif activity profiles that are averaged over replicates using a rigorous Bayesian procedure (see Suppl. Methods). As an example, the replicate-averaged results for the mouse GNF atlas are available at http://ismara.unibas.ch/supp/dataset1/averaged_report/.

Apart from averaging over replicates, this procedure can also be used to calculate contrasts between subsets of samples. To illustrate this, we jointly analyzed the human GNF atlas of 79 tissues and cell lines (54) and the NCI-60 reference cancer cell lines (55) (full results at http://ismara.unibas.ch/supp/dataset2/ismara_report/). By treating all non-tumor samples as one condition and all tumor samples as another condition in the averaging, we can identify motifs that are consistently dis-regulated in cancer. Supplementary tables 2 and 3 show the motifs that are most consistently up-regulated or down-regulated in tumors. Among the top up-regulated motifs are several key transcriptional regulators that are well known in cancer biology such as Hif1a (56) (Fig. 4.2C), Myc (57), and E2F (58). ISMARA also identifies a number of miRNAs whose targets are either consistently upregulated, e.g. miR-205 (Fig. 4.2D) and miR-26, or consistently down-regulated, e.g. miR-24 and the miR-17/93/106

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

seed family, in tumors. Indeed, multiple studies have found miR-205 to be down-regulated in a number of different cancers, and miR-205 has been shown to have tumor suppressor function (59; 60; 61; 62; 63). It has also been shown that miR-26a delivery suppresses hepatic tumors in mouse (64), supporting the downregulation of this miRNA in cancer. Similarly, miR-17 is a known oncogene (65), supporting that its targets are down in cancer. The literature on miR-24 function in cancer is more ambiguous (66). Some evidence has been provided that miR-24 acts as repressor of apoptosis and is upregulated in certain cancers (67). On the other hand, another study found that miR-24 can inhibit proliferation (68). Notably, the latter study suggested that miR-24 acts through seedless target sites, which by construction are not detected by TargetScan. In summary, in this system ISMARA successfully identified oncogenes and tumor suppressors *ab initio*.

4.2.3 Inferring motif activity dynamics: inflammatory response

To illustrate ISMARA's analysis of time series data, we applied it to a time series of expression data obtained after activation of human umbilical vein endothelial cells (HUVECs) with tumor necrosis factor (TNF, previously also known as $\text{TNF}\alpha$). Messenger RNA expression was measured every 15 minutes for the first 4 hours after treatment, and every 30 minutes for the next 4 hours (69). Whereas the original study focused solely on nascent transcription, standard application of ISMARA to this data set (http://ismara.unibas.ch/supp/dataset3/ismara_report/) uncovers the transcription regulatory network involved in this inflammatory response in remarkable detail.

The response of endothelial cells to TNF is known to be mediated by the $\text{NF}\kappa\text{B}$, GATA2, IRF1, and AP-1 (70) TFs. $\text{NF}\kappa\text{B}$ in particular is crucial for the resulting inflammatory response (71). Indeed, ISMARA infers that the two most significant motifs are IRF1,2,7 and $\text{NF}\kappa\text{B}$. The activity of $\text{NF}\kappa\text{B}$ increases sharply in the first 45 minutes and slower afterwards, until it reaches a steady activity after 3 hours. The activity of the IRF1,2,7 motif increases steadily starting at 30 to 45 minutes after treatment until the end of the time course (Fig. 4.3A). As shown by $\text{NF}\kappa\text{B}$'s local network figure (Fig. 4.3B and on the ISMARA results website), ISMARA infers that IRF1 is activated directly at the level of transcription by $\text{NF}\kappa\text{B}$, which is indeed known from previous studies (72). Other predicted targets of $\text{NF}\kappa\text{B}$ that are also found to be significantly upregulated in this process are TNF receptor genes, components of the JAK-STAT pathway (note that STAT2,4,6 is the 11th most significant motif, indicating that STAT activity changes, affecting the level of *its* targets) and MHC class I genes. The latter are also predicted to be regulated by IRF1,2,7, which is confirmed by experimental data (73). ISMARA also predicts that both $\text{NF}\kappa\text{B}$ and IRF1,2,7 activate the 5th most significant motif, PRDM1 (BLIMP-1), which is an important developmental regulator in the B-cell and T-cell lineages and is required

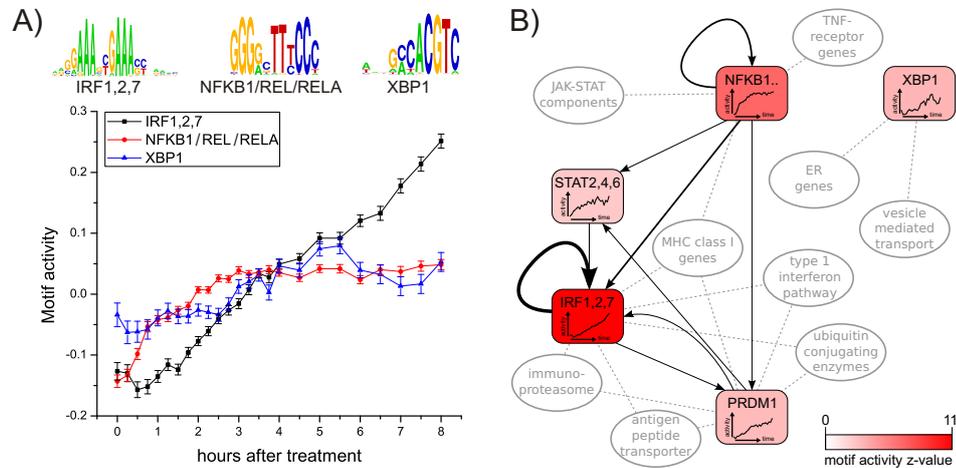


Figure 4.3: Analysis of an inflammatory response time series of human umbilical vein endothelial cells responding to TNF. **A:** Time-dependent activities of the 3 most significant motifs, i.e. $\text{NF}\kappa\text{B}$ (red), IRF1/2 (black), and XBP1 (blue). Error-bars denote uncertainties in the inferred activities. **B:** Summary of the inferred core regulatory network. Selected top motifs are shown together with interactions between them and pathways/functional categories that are enriched among the targets of these motifs. The intensity of the color corresponds to the z -score of the motif, its time-dependent activity is indicated inside the node, and the thickness of each edge corresponds to its target score S_{pm} .

for the secretory pathway in B-cells (74). PRDM1 activity increases, like that of IRF, across the entire time course, and these two TFs appear to share many of their predicted targets, including type 1 interferon pathway genes, the immunoproteasome (75), ubiquitin conjugating enzymes, antigen peptide transporters, and MHC class I genes. All of these targets are consistent with the activation of the antigen presenting pathway by these TFs. Finally, the 3rd most significant motif is XBP1, which is activated only after 2.5 hours. Its predicted targets are highly over-represented for endoplasmic reticulum (ER) genes and genes involved in vesicle-mediated and Golgi transport, consistent with the fact that XBP1 is a major regulator of ER stress and the unfolded protein response (UPR) (76). Moreover, several studies support that the UPR is a general characteristic resulting from inflammation or TNF activation in endothelial cells (77; 78). Interestingly, the induction of XBP1's activity occurs at the same time as the $\text{NF}\kappa\text{B}$ activity stops increasing which is in line with studies showing that the UPR can attenuate $\text{NF}\kappa\text{B}$ induction of inflammation (79; 80; 81). All these predictions of ISMARA, which were made *ab initio* using only the time course expression data, are summarized in the network picture Fig. 4.3B.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Finally, the induction of XBP1's activity is not reflected in the expression of XBP1 itself, which is almost constant across the time course (Suppl. Fig. 16). This underscores that ISMARA infers a motif's activity from the expression of its predicted targets and does not use the regulator's own expression. Indeed, it has been established that XBP1 activity is regulated post-transcriptionally through alternative splicing (82; 83).

4.2.4 Identifying novel master regulators: Mucociliary differentiation of bronchial epithelial cells

Next, we turned to an example system for which much less is known, namely the mucociliary differentiation of bronchial epithelial cells on an air-liquid interface. Aiming to elucidate the regulation of bronchial development, Ross et al. (84) performed differentiation experiments in triplicate over a period of 28 days with cells from three separate donors. This data was then analyzed with commonly used bioinformatic procedures, i.e. genes were clustered into co-expression clusters, and the clusters were analyzed for over-represented gene ontology categories and pathways. This analysis uncovered clusters associated with TGF β pathway genes, extra-cellular adhesion genes, and genes associated with the microtubule cytoskeleton, but no key regulators or regulatory interactions that drive these expression changes were identified.

In contrast, applying ISMARA to this gene expression data set, we obtain the prediction that by far the most important regulatory motif in this system is RFX, whose activity is strongly increasing over the period from roughly day 4 to day 10 in all 3 donors (Fig. 4.4A, http://ismara.unibas.ch/supp/dataset4/ismara_report/). The predicted targets of RFX are highly enriched in genes known to be associated with cilium assembly, axoneme, and the microtubule cytoskeleton genes (Fig. 4.4B) suggesting that RFX directs ciliogenesis in bronchial epithelial cells.

The RFX family of TFs contains 7 members and it is not *a priori* clear which of these are driving the bronchial differentiation. Comparison of the mRNA expression profiles with activity profiles shows that two of the family members, RFX2 and RFX3 exhibit a striking correlation in their expression with the motif activity (Fig. 4.4A and C). Together these results strongly suggest that the TFs RFX2/3 are master regulators of ciliogenesis in this system. This prediction is consistent with previous studies that have shown that RFX3 is necessary for the ciliogenesis of nodal cilia in mouse embryonic development (85) and during ciliogenesis of motile cilia in a mouse cell-culture system (86). More specifically, in the latter study it was found that RFX3 activates the FOXJ1 TF during this process. Interestingly, ISMARA also predicts that RFX directly upregulates FOXJ1 in this system. An interesting novel prediction is that RFX2 is directly regulated by the TF MYB (Fig. 4.4B). This prediction is supported by the observation that the RFX2 promoter is known to contain Myb sites

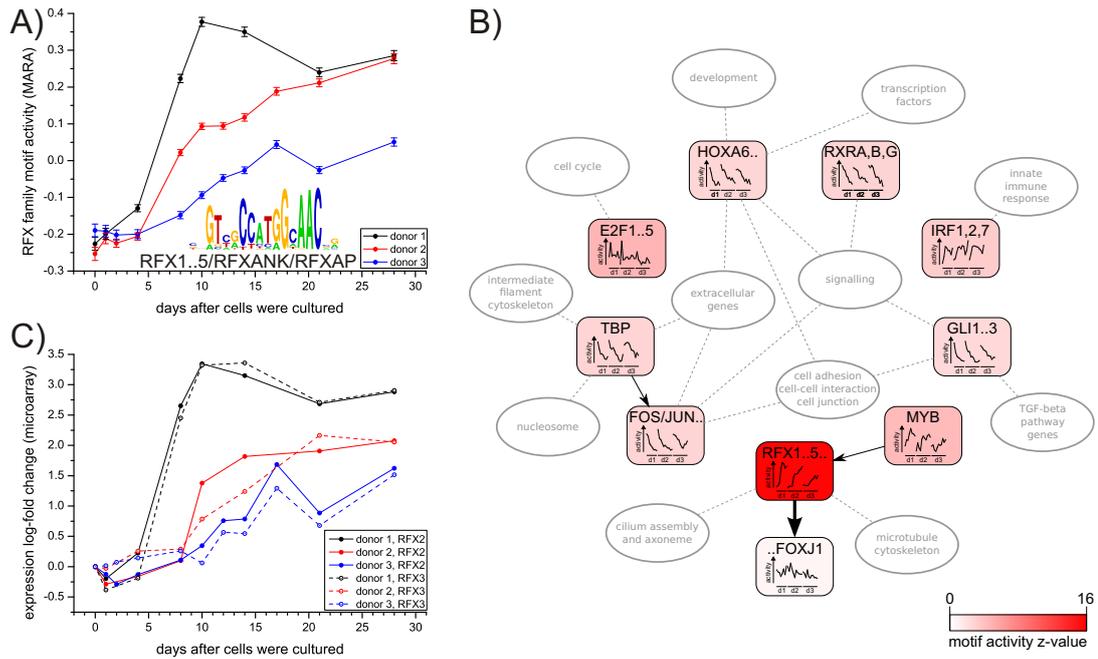


Figure 4.4: Mucociliary differentiation **A**: Inferred RFX motif activity profile in mucociliary differentiation in bronchial epithelial cells from three independent donors (black, red, and blue lines). **B**: Key predicted regulators and their targets in the mucociliary differentiation. Selected top motifs are shown together with predicted interactions between them and pathways/functional categories that are enriched among predicted targets of these motifs. The intensity of the color corresponds to the z -score of the motif, its time-dependent activity for each donor is indicated inside the node, and thickness of the edges corresponds to the target score S_{pm} . **C**: mRNA expression profiles of the RFX2 (solid) and RFX3 (dashed) genes across the differentiation (colors of the donors as in panel A).

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

and is directly regulated by A-myb in spermatogenesis (87).

As indicated in Fig. 4.4B, ISMARA additionally predicts that, in this system, IRF1,2,7 upregulates innate immune response genes, and that a short spike of E2F activity up-regulates cell-cycle genes at day 1. Finally, there is a group of motifs (TBP, FOS_FOS{B,L1}_JUN{B,D}, RXR{A,B,G}, HOX{A6,A7,B6,B7}, and GLI1.3) whose targets are progressively down-regulated across the differentiation time course. The targets of these motifs are generally enriched for extracellular proteins involved in cell adhesion, cell-cell junctions, and signaling. More specifically, targets of GLI1.3 involve genes from the TGF β pathway, targets of TBP involve nucleosomal and intermediate filament cytoskeletal genes, and targets of the homeodomain motif (HOX{A6,A7,B6,B7}) are enriched for developmental genes and transcription factors. The genes in these pathways are most likely involved in the transition of the tissue from squamous to columnar epithelial that occurs during this differentiation. Thus, in contrast to the methods used in the original study (84), ISMARA predicts which regulators are directing various aspects of this differentiation including ciliogenesis, the innate immune response, and the transition from squamous to stratified epithelial.

4.2.5 Epithelial-Mesenchymal Transition: including microRNAs in core regulatory networks

To illustrate ISMARA's ability to integrate the role of both TFs and miRNAs in the gene regulatory network, we took advantage of data from a system in which miRNAs are known to play important regulatory roles: the epithelial-to-mesenchymal transition (EMT). Recently, mRNA expression measurements were performed in duplicate on epithelial and 3 independently-isolated mesenchymal subpopulations within immortalized mammary epithelial cells (88). After running ISMARA on this data (results at http://ismara.unibas.ch/supp/dataset5/ismara_report/), we used replicate-averaging to identify regulators that most consistently and significantly explain the mRNA expression differences between epithelial and mesenchymal cells (results at http://ismara.unibas.ch/supp/dataset5/averaged_report/).

Interestingly, much of what is known about EMT (reviewed by Polyak and Weinberg (89)) is again captured by ISMARA's results. Among the top regulators that ISMARA infers in this system are SNAI1.3, ZEB1, and a family of miRNAs consisting of hsa-miR-141 and hsa-miR-200a (sharing the same seed sequence), that have been shown to form a regulatory network essential for EMT. The predicted activity changes of these regulators are consistent with the extant literature. Namely, the decrease in SNAI1.3 and ZEB1 activity (which indicates a reduced level of their predicted targets) in mesenchymal subpopulations is consistent with the fact that both of them are mainly acting as repressors and are transcriptionally up-regulated

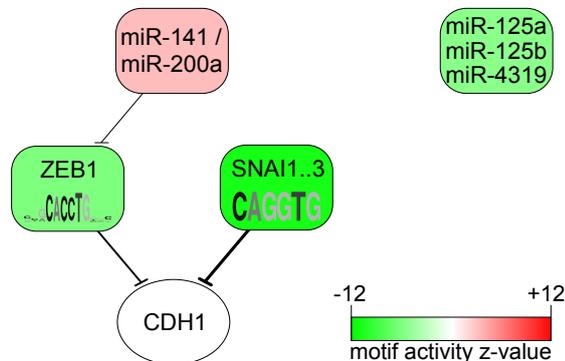


Figure 4.5: Core TF and miRNA regulatory interactions in the epithelial-to-mesenchymal transition, as predicted by ISMARA. Each rectangular node corresponds to a regulatory motif with its color indicating the significance of the change in activity when going from the epithelial to mesenchymal state (z -value defined as $z = (A_{m,\text{mes}} - A_{m,\text{epi}}) / \sqrt{\delta A_{m,\text{mes}}^2 + \delta A_{m,\text{epi}}^2}$). Green/Red indicates targets of the motif are down/up-regulated in the mesenchymal state. Both Zeb1 and Snail are predicted to target the E-cadherin (CDH1) promoter. Note that all interactions shown are repressive.

in the mesenchymal state. The miR-141 and miR-200a miRNAs are known to be down-regulated in the mesenchymal state, causing the mRNA levels of their targets to increase, which is consistent with the positive change in activity predicted by ISMARA. Known regulatory interactions between these factors are also uncovered by ISMARA. For instance, ZEB1 is the top predicted target of the miR-141/200a miRNAs and existing literature confirms that the direct regulation of ZEB1 by miR-200 is critical in EMT (90; 91; 92). Similarly, E-cadherin (or CDH1) is the 3rd and 4th top target gene of the ZEB1 and SNAI1.3 motifs, respectively, and indeed this gene is an epithelial marker known to be targeted by both SNAIL transcription factors (93) and by ZEB1 (94). These key predictions by ISMARA are summarized in Fig. 4.5.

The activity of the family containing the hsa-miR-125a/b and hsa-miR-4319 miRNAs is the most significantly reduced miRNA family in EMT. This suggests that these miRNAs play a role in mesenchymal cells, consistent with observations that miR-125b promotes invasive tumor characteristics (95).

4.2.6 TF activities effecting chromatin state: analysis of ChIP-seq data

Beyond analyzing gene expression data, motif activity response analysis can be applied to modeling any signal along the genome in terms of the local occurrence of TFBSs. Indeed, in a recent work (96) we applied the MARA approach to ChIP-seq data mapping the dynamics of tri-methylation at lysine 27 of histone 3 (H3K27me3) and identified TFs involved in recruiting this epigenetic mark set by the Polycomb system. In ISMARA the analysis of ChIP-seq data has now been completely automated. In particular, given a ChIP-seq data set, ISMARA quantifies the signal at all promoters across all samples and models this in terms of the TFBSs at each promoter. For the details of ISMARA's processing and normalization of the ChIP-seq data we refer to the Supplementary Methods.

To illustrate ISMARA's results on ChIP-seq data, we make use of data from the ENCODE project in which, besides gene expression, 9 different chromatin modifications were measured across 8 different cell types (97) (all modifications and cell types are listed in Suppl. Tables 4 and 5). We first ran ISMARA separately on each of the 10 data sets, i.e. expression and 9 chromatin modifications (see Suppl. Table 6 for the URLs of the results on all data sets). Exploring these results we observed that motifs that are highly significant for explaining differences in levels of a particular chromatin mark across tissues, were often also highly significant for explaining *mRNA expression* differences. This was particularly the case for methylation of lysine 4 on histone H3 (H3K4me2, H3K4me3), for acetylation of histone H3 (H3K9ac, H3K27ac), and for tri-methylation of lysine 36 on histone H3 (H3K36me3). For example, Fig. 4.6A shows the activity profiles for these marks for the SNAI motif, which is recognized by the Snail TFs. Other examples of activity profiles of motifs with high significance for these marks are shown in Suppl. Fig. 17. As these figures show, for each motif, the activity profile for expression is highly similar to those of all of these histone marks. Indeed, it has been well recognized that these chromatin marks are associated with promoter activity (98), and several recent studies have shown that the levels of these marks can be used to predict gene expression levels (99; 100; 101).

To investigate the correlations between the levels of the different chromatin marks more quantitatively, we performed principal component analysis (PCA) of the distribution of the 10 different marks across all promoters, separately for each sample (Suppl. Methods). Strikingly, we find that in each sample, the first PCA component explains the majority of the variance across promoters, typically explaining around 60% of the total variance (Suppl. Fig. 18). Moreover, we find that the first PCA component looks virtual identical for each sample (Suppl. fig. 18) and Fig. 4.6B shows the first principal component obtained using PCA on the pooled data from all cell types. These findings strongly suggest that there is a single variable which corresponds roughly to 'promoter activity', which captures a large fraction of the

4.4.2 Results

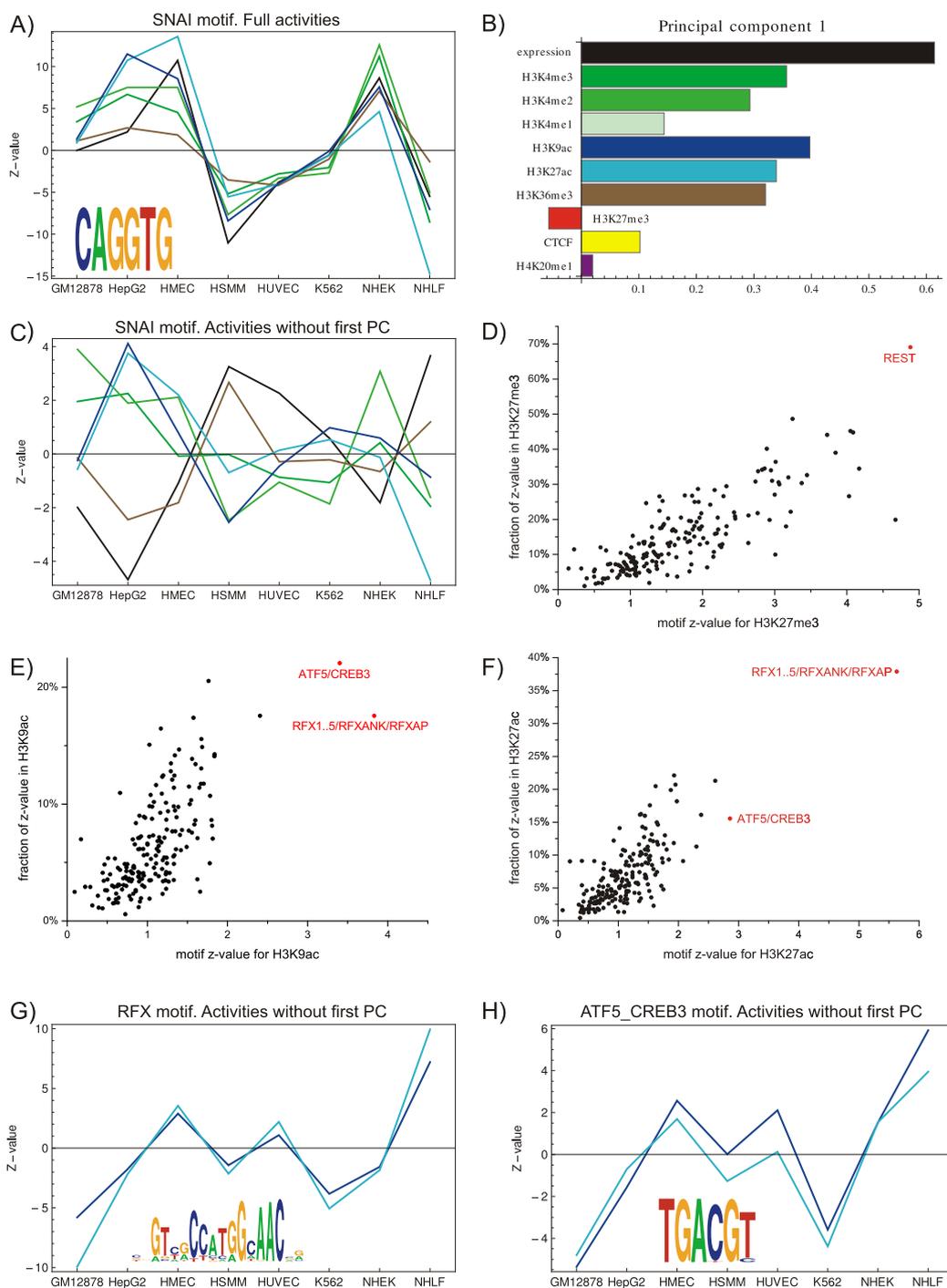


Figure 4.6: ISMARA predicts TFs involved in recruiting specific chromatin marks. (continued on the next page)

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Figure 4.6: ISMARA predicts TFs involved in recruiting specific chromatin marks. **A:** Activity across cell types of the Snail motif for explaining expression (black), and levels of the chromatin marks H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown). **B:** First principal component explaining the majority of variation in chromatin mark levels across all cell types. The bars indicate the relative contributions to the principal component of each mark. **C:** Motif activities of the Snail motif, as in panel A, but after removal of the first principal component. **D:** Z -values and specificities (see text) of motifs for explaining H3K27me3 levels. The REST motif, with both highest z -value and highest specificity, is indicated in red. **E:** As in panel D, for H3K9ac levels. The two most significant motifs are shown in red. **F:** As in panels D and E, for H3K27ac levels. **G:** Activity, after removal of the first principal component, of the RFX motif for explaining H3K9ac (dark blue) and H3K27ac (light blue) levels. **H:** As in panel G, for the ATF5_CREB motif.

variation in all chromatin mark levels at the promoter. In addition, the fact that this first principal vector is identical in all tissues suggests that the relative levels of the different marks in this first principal vectors result not from tissue-specific but from general factors, e.g. conceivably they may result from the general transcription machinery recruiting chromatin modifying enzymes.

The first principal vector has its highest positive component along the expression axis showing that, as expected, the expression level of the gene is most strongly aligned with its ‘promoter activity’. The known activation-associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3, also all have a strong positive component in the ‘promoter activity’ vector. The H3K4me1 mark, which has recently been identified as a mark associated with enhancers when not accompanied by H3K4me3 (102), has a weaker positive component in the ‘promoter activity’ vector, as does the level of binding of the CTCF transcription factor, which is generally associated with open chromatin (103). The known repressive mark H3K27me3, which is set by the developmentally important Polycomb system (104), indeed has a negative component in the promoter activity vector. Finally, the H4K20me1 mark shows little contribution to the first PCA component.

In summary, the PCA analysis has shown that there is a single vector in the 10-dimensional space of expression and chromatin marks that represents the general activity of a promoter and captures almost two-thirds of the variation in the levels of all marks across promoters. As a consequence, whenever a given motif contributes significantly to explaining mRNA expression in a sample, it will also contribute significantly to general promoter activity, and thereby to many of the chromatin marks. This also explains why the activity profiles of motifs that significantly explains ex-

pression are all highly correlated (Fig. 4.6A and Suppl. Fig. 17). Thus, the effect of general promoter activity on all chromatin marks confounds identification of TFs that are involved in affecting specific marks, and it would thus be beneficial to remove it. To this end we separated the activity of each motif into a part along the first PCA component, i.e. the one associated with general promoter activity, and the remaining parts along all other components. We then discarded the part of the activity along the first PCA component. As illustrated in Fig. 4.6C and Suppl. Fig. 17, after removal of the first principal component, there are no longer any obvious correlations in the remaining motif activity profiles for different activating marks. We then analyzed the remaining motif activities to identify motifs that contribute to the levels of a particular chromatin mark, independent of the motif's effect on general promoter activity.

For each motif and each mark, we calculated z -values for the remaining activity with respect to each chromatin mark. In addition, we quantified, for each motif and each chromatin mark, a 'specificity' which measures the fraction of its overall significance that is associated with the mark (Suppl. Methods). Strikingly, we find that for many of the marks, the motifs that most significantly affect the mark are also among the most specific for that mark. For example, REST is the motif with the highest z -value for H3K27me3 levels, and is also by far most specific for H3K27me3 (Fig.4.6D). Indeed, in recent work (96) we showed that REST is involved in recruiting this mark during the differentiation of murine embryonic stem cells into pyramidal neurons, specifically at the neural progenitor state. With respect to the two acetylation marks, i.e. H3K9ac and H3K27a, we find that the same two motifs, i.e. RFX and ATF/CREB, are most significant for both these marks (Fig. 4.6E and F). It is well known that ATF/CREB TFs can recruit histone acetylases (HATs) such as CREB binding protein (CBP) and p300 (105), and for RFX TFs it has also been established that they can recruit HATs at particular promoters (106). Our results thus suggest that recruitment of HATs by TFs bound to ATF/CREB and RFX motifs make an important contribution to genome-wide histone acetylation. Moreover, the activity profiles of these motifs for H3K9ac and H3K27ac are highly similar, suggesting that these two marks may be recruited through a common or highly overlapping pathways. Supplementary Fig. 19 shows the most significant motifs for each of the other marks. Among the additional predictions made by ISMARA is that the PITX motif is associated with both mono- and di-methylation of lysine 4 of histone 3. This prediction is supported by recent biochemical evidence that PITX2 can recruit methyltransferases that methylate H3K4 (107). As expected, CTCF is the most significant motif explaining CTCF binding. ISMARA also makes several predictions that are completely novel, as far as we have been able to determine: It predicts that the hepatocyte nuclear factors HNF1A and HNF4A have the most significant effect on the levels of the H3K36me3 mark, which is known to be set by elongating RNA polymerase (108; 109), and that YY1 and NF-Y most significantly explain variations in H4K20me1 levels.

4.3 Discussion

Just how crucial gene regulatory circuits are in animals is evident when we remind ourselves that every cell in a multi-cellular organism has essentially the same genome, and that the phenotypic differences between cell types largely reflect differences in gene expression. The eventual goal of computational modeling of gene regulatory networks is to have realistic models of the physico-chemical interactions involved on a genome-wide scale, that accurately predict observed expression dynamics. For some very well-characterized systems of moderate size, such explicit biophysical models now appear within reach. For example, for the early antero-posterior body patterning in *Drosophila* relatively realistic models are able to roughly capture the spatial expression patterns of dozens of cis-regulatory modules in terms of the concentration profiles of 5 – 10 TFs (110; 111).

However, for the vast majority of systems our knowledge is far too rudimentary to make such detailed modeling viable. For example, an exciting recent development is the ability to reprogram cells from one differentiated state into either a stem cell state (112) or another differentiated state (113), by over-expressing or silencing specific regulatory factors. Although factors that can trigger the reprogramming cascade are known and increasing amounts of high-throughput data are available for these systems, very little is known about the regulatory networks that ultimately control these differentiation processes. The question faced by a computational biologist when analyzing such systems is how to make progress in identifying the key gene regulatory interactions given little specific knowledge of the system, and the enormous number of components potentially contributing to the system.

The advent of high-throughput technologies now allows the routine measurement of genome-wide mRNA expression across conditions, and such data in principle provide the opportunity to systematically investigate gene regulation on a genome-wide scale. Such investigations require sophisticated computational approaches and, not surprisingly, a vast literature of methods has emerged for analyzing such genome-wide expression data, ranging from explicit regulatory network models to ‘black box’ machine learning methods that mainly aim to capture abstract patterns in the data. Within the computational systems biology community it is sometimes implicitly assumed that the purpose of computational models of gene regulatory networks is to accurately predict gene expression patterns (114). However, for most systems our current knowledge is far too rudimentary to expect that explicit regulatory network models can successfully model genome-wide expression patterns. Moreover, in predicting gene expression, realistic regulatory network models are often outcompeted by *ad hoc* machine learning approaches (115). However, such approaches provide little or no insight into the underlying regulatory networks. In our opinion, the challenge is not so much to develop models that fit gene expression patterns most accurately, but to develop methods that can exploit high-throughput data to gain new insights into

the underlying regulatory processes. To achieve this, the computational methods should help guide subsequent experimental efforts by prioritizing which regulatory factors are likely key players in the system, and making concrete predictions of the regulatory interactions they engage in, i.e. predictions that are directly amenable to experimental follow-up.

The Integrated System for Motif Activity Response Analysis (ISMARA), that we have presented here provides such a computational approach. Using only gene expression or chromatin state (ChIP-seq) data as input, ISMARA makes concrete predictions regarding key regulators and their regulatory interactions. Moreover, in contrast to many computational methods which require dedicated computational experts to apply, ISMARA is completely automated and provides its results in a user-friendly web-interface. In this way, ISMARA empowers experimentalists to infer concrete hypotheses about the genome-wide regulatory interactions acting in their system of interest, and use these to help guide more detailed experimental investigations.

That motif activity response analysis is a powerful method for reconstructing regulatory interactions from high-throughput information was already demonstrated in its original application, i.e. the reconstruction of the core regulatory network of a differentiating human cell line (25). More recently the same approach was applied in several collaborations (116; 117; 118; 119; 120), showing that, in every case, ISMARA successfully inferred key regulators and their regulatory interactions *ab initio*. The applications in this work not only further confirm that, in systems where key regulatory interactions are already known, ISMARA successfully infers them, but also provide a large collection of novel regulatory predictions across different systems in human and mouse, e.g. novel regulators that are dysregulated in cancers, novel regulatory interactions in the inflammatory response, master regulators of the mucociliary differentiation, etcetera. Moreover, the applications highlight several of the advantages of ISMARA. First, the fact that ISMARA infers a regulator's activity from the behavior of its targets means that non-transcriptional activity changes, i.e. due to post-translational modifications, changes in cellular localization, or interactions with co-factors, can also be readily detected. Second, when a regulator's activity is transcriptionally regulated, this can help identify the relevant TF, e.g. as we did in the mucociliary differentiation by identifying RFX2 and RFX3 from the family of RFX TFs, and it can also indicate whether a regulator is acting as a repressor or an activator.

Beyond providing sorted lists of targets of each motif, the Gene Ontology analysis and the automatic visualization of the STRING network of target genes is typically very helpful in identifying the biological functions and pathways that are targeted by a particular regulator in a particular system. The links, for each predicted target, to the individual binding sites on the genome (37) provide precise predictions of the DNA segments through which the regulatory interactions are implemented, allowing for targeted validation experiments. Similarly, ISMARA's predictions of direct reg-

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

ulatory interactions between the key regulatory motifs provide concrete hypotheses regarding the regulatory circuitry that is acting in a given system, e.g. the predicted regulatory feedbacks between NF κ B, IRF, and PRDM1, or the prediction that MYB is an upstream activator of RFX in the mucociliary differentiation.

Apart from the fact that miRNAs form an important separate regulatory layer in gene expression regulation, there are many indications that the actions of miRNAs and TFs are tightly integrated and interlinked (121; 122; 123). By integrating both TF and miRNA regulation within an automated computational inference procedure, this allows researchers to generate hypotheses regarding the interplay of miRNAs and TFs for their system of interest. ISMARA's successful inference of the key regulatory interactions between miRNAs and TFs in EMT demonstrate its capability in this regard.

Finally, it has become clear that, especially in higher eukaryotes, regulation of gene expression involves a tight interplay and feed-back between the actions of TFs and chromatin state, with chromatin state affecting accessibility of the TFs to their sites, and TF binding being an important mechanism for recruiting chromatin modifiers that locally alter the chromatin state. In a recent work (96) we demonstrated that motif activity response analysis, applied to ChIP-seq data measuring histone modifications, can successfully identify key TFs that are involved in dynamic regulation of chromatin state. Here we have applied ISMARA to chromatin state data from the ENCODE project and provided novel predictions for, among other things, regulatory factors involved in recruiting histone acetylations.

There are of course several limitations to our approach. First, we follow Bussemaker and others (4; 43) in using a simple linear model to relate predicted TFBSs to expression patterns. The main advantage of this approach is that the model is very robust, e.g. not sensitive to wrongly predicted TFBSs or to the noise in the microarray and sequencing data. In addition, in contrast to most non-linear models, the linear model can be exactly solved, even for very large numbers of promoters and samples, so that we are guaranteed to have identified the optimal solutions. However, it is clear that it would be desirable to include saturation of the gene expression response to changes in TF activity. A second limitation is that MARA assumes that a given TF acts either mainly as an activator or mainly as a repressor whereas it is clear that some TFs can act as an activator on some targets and as a repressor on other. Indeed, it has been recently shown (124) that allowing such dual function of TFs can significantly increase correlation coefficients between model predictions and measurement. Explicitly considering higher order constellations of TFBSs, e.g. the occurrence of pairs or triplets of TFBSs for particular combinations of TFs, is another obvious extension that we are currently evaluating.

In mammals, sequence-specificities are available for only about 350 of the about 1500 TFs. Thus, it is clear that an important direction for improvement would be to obtain more comprehensive data on the sequence specificity of TFs. Recent de-

velopments in protein array technology (125), and the dramatic decrease in cost of ChIP-seq experiments make it highly likely that significant amounts of such data will become available over the coming years and we plan to use these to expand the set of regulatory motifs in ISMARA on a regular basis. In addition, whenever ChIP-seq data are available for a particular TF in a particular system, these can be used in place of the TFBS predictions to identify target promoters directly. Indeed, we successfully used this approach in our analysis of REST’s role in Polycomb recruitment (96).

ISMARA currently focuses solely on predicted TFBSs in proximal promoters, ignoring the effects of distal enhancers. The main reason for this that, in contrast to promoters, accurate genome-wide maps of distal enhancers have not been available. However, the recent realization that active enhancers exhibit characteristic chromatin modification patterns (126), DNA methylation patterns (127), and more generally DNA accessibility patterns (128), has paved the way for accurate, genome-wide identification of distal enhancers. Once a set of relevant enhancers has been identified, it is straight forward to predict TFBSs within these enhancers and incorporate these into the model.

One of the ultimate goals is to understand how regulatory interactions determine the dynamics of gene expression, and how stable ‘attractors’ corresponding to individual cell types are established. The direct regulatory interactions between motifs that ISMARA predicts provide a first indication of interactions that may be crucial for the observed regulatory dynamics. A key challenge in the coming years is to go beyond analysis at individual time points and develop causal models of the regulatory networks controlling the dynamics of gene expression.

4.4 Methods

Although most of the individual steps in the computational analysis employed in ISMARA are conceptually straight forward, the quality of the final results depends on many details in the computational ‘protocols’, and we have invested large efforts into optimizing these. For space considerations, we provide all detailed methods in the Supplementary methods and here only summarize the key steps.

ISMARA relies on our annotation of promoteromes in human and mouse, which we obtained using a combination of deepCAGE data (Chapter 3) and known transcripts. For each promoter, we extracted 500 bps upstream and downstream of the TSS, and orthologous segments in 6 other mammals. The 7 orthologous sequences were then multiply aligned using T-Coffee (129). We curated a collection of ≈ 200 WMs representing ≈ 350 mammalian TFs using data from the JASPAR (35) and TRANSFAC (36) databases, additional motifs from the literature, and our own analysis of ChIP-chip and ChIP-seq data. Binding sites were predicted using the MotEvo

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

algorithm (39) and were performed separately for CpG and non-CpG promoters. In addition, we estimated a prior probability profile as a function of position relative to TSS for each motif. For miRNA targeting, we used the predictions of TargetScan (40) of target sites in the 3' UTR sequences of transcripts. Using the association between transcripts and promoters the miRNA target sites were associated with promoters. The miRNA predictions encompassed ≈ 100 conserved miRNA seed families. The end result of these regulator-target predictions was a site-count matrix \mathbf{N} , with elements N_{pm} corresponding to the estimated total number of functional binding sites for motif m in promoter p (where motifs include both TF WMs and miRNA seed families). Raw microarray and RNA-seq data were processed using standard normalization procedures. To estimate an expression profile E_{ps} for each promoter p , we use collections of known transcripts from human and mouse. We associate each promoter with all known transcripts starting at or very near the promoter, and intersect RNA-seq reads and microarray probes with the transcripts. ChIP-seq reads are directly intersected with promoter regions, extended to the length of 1000 bp both upstream and downstream of the TSS.

To fit motif activities A_{ms} using equation (4.1) we assume that the deviations between the model and the measurements E_{ps} are Gaussian distributed with unknown variance σ^2 in each sample. To avoid over-fitting we use a Gaussian prior over motif activities A_{ms} and set the variance of this prior so as to maximize generalization accuracy in a cross-validation test. Using SVD to obtain the multi-variate Gaussian posterior of motif activities, we obtain both estimated motif activities A_{ms}^* and associated (marginal) error bars δA_{ms} . The significance of each motif m is summarized by a z -like statistic:

$$z_m = \sqrt{\frac{1}{S} \sum_{s=1}^S \left(\frac{A_{ms}^*}{\delta A_{ms}} \right)^2}, \quad (4.2)$$

where S is the number of samples. To average motif activities over subsets of samples (for example replicates) we use a Bayesian procedure that estimates both the mean activity across a subset of samples, as well as its variation. Using these estimates new error bars δA_{ms} and motif z -scores z_m are calculated.

To predict the targets of a motif m we measure, for each promoter p with predicted binding sites for m , the decrease of the quality of the fit upon removal of motif m from the model, and quantify this by a log-likelihood ratio S_{pm} . Finally, enrichment of targets within particular Gene Ontology categories is done by selecting all targets where inclusion of motif m substantially helps predicting the expression levels ($S_{pm} > 1$) and performing a standard hypergeometric test. Target networks between motifs are constructed by drawing a link from motif m to m' whenever m is predicted to target one of the promoters associated with a TF that is associated with motif m' .

4.4.1 Materials

The publicly available data sets of gene expression profiling were obtained from Gene Expression Omnibus: time course of HUVEC after TNF treatment (GSE9055), mucociliary differentiation of airway epithelial cells (GSE5264), Novartis (GNF) SymAtlas (GSE10246, GSE1133), epithelial and mesenchymal subpopulations within immortalized human mammary epithelial cells (GSE28681), ENCODE ChIP-seq (GSE26386) and expression profiling (GSE26312) in human cell lines. Microarray files from the NCI-60 were downloaded from the project web page (<http://genome-www.stanford.edu/nci60/>).

4.5 Supplementary Methods

4.5.1 Human and mouse promoteromes

The central entities whose regulation is modeled by ISMARA are *promoters*. When analyzing expression data, be they micro-array or RNA-seq, ISMARA estimates and models the expression profiles of individual promoters, and when analyzing ChIP-seq data ISMARA models the chromatin state of genomic regions centered on promoters. Thus, the first step in the analysis consists of the construction of reference sets of promoters in human and mouse. To make a comprehensive list of promoters we used two sources of data: deepCAGE data, i.e. next-generation sequencing data of 5' ends of mRNAs (12; 30), and the 5' ends of all known mRNAs listed in GenBank.

Using CAGE data from a considerable set of human and mouse tissues, we recently constructed genome-wide human and mouse 'promoteromes' (Chapter 3) consisting of a hierarchy of individual transcription start sites (TSSs), transcription start clusters (TSCs) of nearby co-regulated TSSs, and transcription start regions (TSRs), which correspond to clusters of TSCs with overlapping proximal promoter regions. As the basis of our promoter sets we started with the sets of TSCs, i.e. local clusters of TSSs whose expression profiles are proportional to each other to within experimental noise, as identified by deep-CAGE.

As the currently available CAGE data do not yet cover all cell types in human and mouse, a substantial number of cell type-specific promoters are not represented within this set of TSCs. We thus supplemented the TSCs with all 5' ends of mRNAs, using the BLAT (130) mappings from UCSC genome browser web site (131). To avoid transcripts whose 5' ends are badly mapped, we filtered out those for which more than 25 bases at the 5' end of the transcript were unaligned. We then produced reference promoter sets by iteratively clustering the TSCs with the 5' ends of mRNAs as follows: Initially each TSC and each 5' end of an mRNA forms a separate cluster. At each iteration the pair of nearest clusters are clustered, with the constraint that there can be at most one TSC per cluster. That is, we never cluster two TSCs together as our previous analysis in (Chapter 3) has already established that each TSC is independently regulated. This iteration is repeated until the distance between the closest pair of clusters is larger than 150 base pairs. Note that we thus chose the length of sequence wrapped by a single nucleosome, i.e. roughly 150 base pairs, as an *ad hoc* cut-off length for two TSSs to belong to a common promoter. The reasoning behind this choice of cut-off, is that, on the one hand, we have empirically observed that co-expressed TSSs can spread over roughly this length-scale and, on the other hand, that it is not implausible that the ejection of a single nucleosome near the TSS may be responsible for setting this length scale. In any case, the resulting promoters are not sensitive to the precise setting of this cut-off (data not shown). Finally, inspection of the results showed, especially in ubiquitously expressed genes, many apparent TSSs

from Genbank that appear downstream of both the TSSs identified by deep-CAGE and the annotated RefSeq transcripts. It is highly likely that many of these apparent TSSs are due to cDNA sequences that were not full length. Indeed, only a small fraction of the transcripts in the database of mRNAs underwent expert curation, and truncated transcripts are likely common. To avoid such spurious apparent TSSs we removed all clusters which did not contain at least one curated transcript (RefSeq) or a TSC. Finally, since a sequence of at least one associated transcript is necessary to estimate a promoter's expression level from either RNA-seq or micro-array data, we also discarded all promoters that consisted solely of a TSC.

For human, the resulting reference promoter set had 36'383 promoters, of which 13'265 contained both a TSC and at least one RefSeq transcript, 14'538 contained only a TSC together with non-RefSeq transcripts, and 8'580 had at least one RefSeq transcript and potentially non-RefSeq transcripts, but no TSC. For the mouse genome, the corresponding numbers are: 34'050 promoters in total, 8'578 RefSeq-only, 12'303 TSC-only, and 13'169 with both a TSC and at least one RefSeq transcript. These reference promoters sets cover almost all known protein-coding genes in human and mouse.

Finally, as we discussed in Chapter 3, mammalian promoters clearly fall into two classes associated with high and low content of CpG dinucleotides, and these promoter classes have clearly distinct architectures, i.e. different lengths, different numbers of TSSs per promoters, and different distributions of transcript factor binding sites (TFBSs). We classified all promoters into a high-CpG and low-CpG class based on both the CG content and the CpG content in the proximal promoter, as described in Chapter 3. In the TFBS prediction below we perform separate predictions for high-CpG and low-CpG promoters.

4.5.2 A curated set of regulatory motifs

We use standard position dependent weight matrices (WMs) to represent regulatory motifs, i.e. the sequence specificities of TFs. Each WM is named for the TFs that are annotated to bind its site. For some motifs the names correspond to multiple TFs which are all assumed to bind to the same sites. We used a partly manual curation procedure whose details were first described in (25). For completeness, we here also give a description of this curation procedure.

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for mammalian transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory.

Our main objectives were

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

1. To remove redundancy, we aim to have no more than 1 WM representing any given TF. Whenever multiple TFs have WMs that are statistically indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.
2. To associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.
3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consisted of

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM (132).
2. The collection of TRANSFAC vertebrate WMs (version 9.4) and the amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs (35).
3. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).
4. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles (133).

We decided not to include 6 TRANSFAC motifs, which were constructed out of less than 8 sites: M00326 (PAX1, PAX9), M00619 (ALX4), M00632 (GATA4), M00634 (GCM1, GCM2), M00630 (FOXO1), M00672 (TEF). TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and could therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically highly similar and appear redundant. Therefore, we decided to remove this redundancy and for each TF with multiple WMs in TRANSFAC we choose only a single ‘best’ WM based on TRANSFAC’s own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score. The information score of a WM is given by 2 times the length of the WM minus its entropy in bits.

We next aimed to obtain, for each human TF, a list of WMs from JASPAR / TRANSFAC, that can potentially be associated to this TF. To do this we aim to

find, for each TF, which motifs from JASPAR/TRANSFAC are associated with a TF that has a highly similar DNA binding domain. To this end we ran Hmmer (134) with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all TFs (E-value cut-off 10^{-9}) associated with either JASPAR or TRANSFAC matrices. We then represented each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR/TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF.

From this data we created a list of ‘necessary WMs’ as follows. For each human TF we obtain the JASPAR WM with the highest percent identity in the DBDs of an associated TF. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations

1. Different WMs for TFs with identical or near identical DBDs.
2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtained three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.
2. The overlap of the binding sites predicted by each WM. We use MotEvo to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the two WMs.
3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the two WMs and calculate the likelihood-ratio of these sites assuming they either derive from a single underlying WM or assuming that the set of sites for each WM derives from an independent WM.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

For each of these three criteria we set a cut-off: 95% identity of the DBDs, 60% overlap of predicted TFBSs, and a likelihood-ratio of e^{40} . Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster so as to optimize the information content of the resulting fused WM, which is obtained by simply summing the counts across each column in the alignment.

Finally, we used MotEvo (41) to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed new WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates position-specific prior probabilities, as described below). The number of top-scoring sites was chosen manually for each motif and was between 100 and 4000 sites, in most cases being 200 or 500 sites.

At this point we excluded one TRANSFAC motif M00395 (HOXA3, HOXB3, HOXD3) which had very low information content and predicted only very low-probability sites. We additionally excluded the motifs M00480 (TOPORS) and M00987 (FOXP1), which were unrealistically specific and (in case of M00987) predicted stretches of poly(T).

For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2, XBP1, PRDM1, and the RXRG dimer) and we used these to replace the corresponding WM in the list.

We improved several motifs by running MotEvo on TF ChIP-seq data: SRF, STAT1/3, REST and ELK1/4/GABPA/GABPB1. Some other motifs were obtained by predicting *de novo* using the Phylogibbs algorithm (135) on ChIP-seq data: SPI1, CTCF, OCT4, SOX2 and NANOG.

For a few motifs JASPAR has recently updated or introduced new motifs which were based on high-throughput data and we included these motifs. This is the case for FOXA2, KLF4, EWSR1-FLI1, FEV, NR4A2. We also removed MA0118, as it had been discarded in JASPAR data base.

Our final list contains 189 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 340 human TFs are associated with our 189 WMs. The corresponding mouse orthologous TFs were selected using the MGI data base (136). The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (<http://www.swissregulon.unibas.ch>).

4.5.3 Transcription factor binding site predictions

After creating reference promoter sets and curating a set of mammalian regulatory motifs we next predicted TFBSs in the proximal promoter regions of each promoter. Analysis of sequence conservation in the neighborhood of TSSs (Chapter 3) and experimentation with TFBS prediction in regions of different lengths around TSSs indicated that a reasonable balance between sensitivity (i.e. including relevant binding sites) and specificity (avoiding too many false positive predictions) can be obtained by predicting TFBSs in a 1 kilobase region around the TSSs of each promoter.

For each promoter, we thus extended the promoter sequence spanned by its cluster of TSSs by 500 bp upstream and 500 bp downstream. We denote this as the *proximal promoter region* of a promoter. We then extracted the sequence of the reference species, i.e. human or mouse and orthologous regions from 6 other mammals (human or mouse, rhesus macaque, cow, dog, horse, and opossum) using pairwise BLASTZ (137) alignments. For each promoter, we multiply aligned the orthologous regions using T-Coffee (129).

To obtain a phylogenetic tree for these mammalian species, with branch lengths corresponding to the expected number of substitutions per neutrally evolving site, we used methods described previously (138). Briefly, we first obtained the topology of the tree from the UCSC genome browser (139). Then, for each pair of species we made pairwise alignments of the coding regions of orthologous genes and extracted all third positions in fourfold-degenerate codons of amino acids that are conserved between the two species. Using these fourfold-degenerate positions we estimated a pairwise distance for each pair of species. Finally, we estimated the lengths of the branches in the phylogenetic tree as those that minimize the square-deviations between the implied pairwise distances and the pairwise distances estimated from the fourfold-degenerate positions. The resulting tree structure is shown in Suppl. Fig. 4.7.

The multiple sequence alignments were then used together with the phylogenetic tree and the collection of WMs as an input for TFBSs predictions using the MotEvo algorithm (41). Given a multiple alignment, MotEvo considers all ways in which the sequence of the reference species can be segmented into ‘background’ positions, ‘binding sites’ for one of the supplied WMs, and ‘unknown functional elements’ (UFEs). The likelihood of alignment columns assigned to background are calculated under a model of neutral evolution along the specified phylogenetic tree. The likelihood of alignment segments assigned to be a site for a given WM are calculated by first estimating which of the species have retained a site for the WM (based on the WM scores of the individual sequences) and then applying an evolutionary model in which substitution rates are set so as to match the sequence preferences of the WM. Finally, segments assigned to be UFEs are assumed to evolve under *unknown* purifying selection constraints on the sequence, which is implemented by treating them as sites

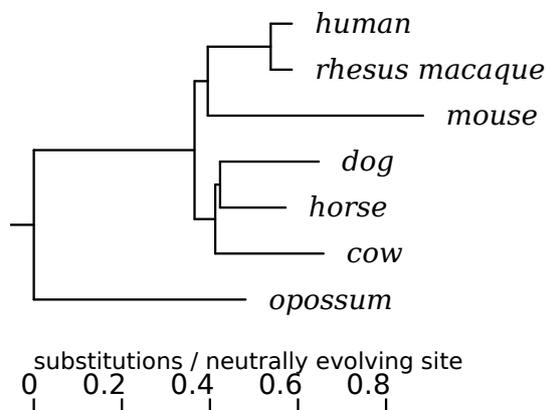


Figure 4.7: The phylogenetic tree used by MotEvo for the transcription factor binding site predictions that are used by ISMARA.

for an unknown WM, which is treated as a nuisance parameter that is integrated out of the likelihood. Finally, MotEvo assigns, at each position of the alignment and for each WM, a posterior probability that a site for the corresponding WM occurs at this position.

Since most motifs show clear positional preferences relative to TSS, we implemented a position-dependent binding prior probability distribution for each motif which we fitted by maximum likelihood using expectation-maximization. Since high-CpG and low-CpG promoters have highly distinct configurations of TFBSs, we estimated the position-dependent prior probability distributions separately for high-CpG and low-CpG promoters.

The final result of this analysis is a matrix \mathbf{N} , with N_{pm} the total number of predicted sites for motif m in promoter p , i.e. the sum of the posterior probabilities of the individual sites. To reduce the probability of spurious predictions, we set $N_{pm} = 0$ whenever the sum of the posteriors of all sites was less than 0.1.

4.5.4 Associating miRNA target sites with each promoter

Apart from incorporating the effects of TFBSs in promoters, ISMARA also integrates the effects of miRNAs in its modeling of expression levels. To this end, we needed to obtain a set of predicted miRNA target sites for each promoter. We base our predictions on the miRNA target predictions of TargetScan using preferential conservation scoring (aggregate P_{CT}) (40) which has shown consistently high performance in various benchmark tests. As opposed to focusing on individual miRNAs, TargetScan groups miRNAs that have identical subsequences at positions 2 through 8 of the miRNA, i.e. the 2-7 seed region plus the 8th nucleotide, and provides predictions

for each such seed motif. We will treat these seed motifs exactly like the regulatory motifs (WMs) for TFs, i.e. a miRNA seed motif can be associated with multiple miRNAs. TargetScan provides predictions for 86 mammalian miRNA seed motifs in total.

TargetScan P_{CT} provides a score for each seed motif and each RefSeq transcript. To obtain a ‘site count’ N_{pm} for the number of sites of miRNA seed motif m associated with promoter p we average the TargetScan P_{CT} scores of all RefSeq transcripts associated with the promoter p . Finally, the miRNA seed motif site counts N_{pm} are simply added as columns to the site count matrix \mathbf{N} with site counts of TFBSs.

4.5.5 Expression data processing

When using expression data from oligonucleotide microarrays, the raw probe intensities are corrected for background and unspecific binding using the Bioconductor packages `affy` (140), `oligo` (141), and `gcrma` (142), depending on the type of the particular microarray used. The micro-arrays that are currently supported by ISMARA are listed in supplementary table 4.1.

For its further analysis, ISMARA uses the logarithms of the probe intensities. For a given sample, the histogram of log-intensities is generally bimodal, with the modes corresponding to probes of non-expressed and expressed genes. The probes are classified as expressed or non-expressed in each sample separately by fitting a two-component Gaussian mixture model to the log-intensity data using the `Mclust R` package (143; 144). Probes that are consistently non-expressed are filtered out from further processing; a probe is considered not to be expressed if in *all* the samples the probability of it belonging to the expressed class is below 0.4. Subsequently, the intensity values are quantile normalized across all input samples.

Micro-array probes can hybridize to multiple transcripts, belonging to different genes, or different isoforms of one gene, and we decided not to rely on transcript annotations of a micro-array producer. Instead, we comprehensively mapped the probe sequences to the set of all transcripts that are associated with our reference set of promoters. Note that we thus also ignore the annotation of probes into probe sets. To calculate the expression of a promoter we average the log-expression levels of all probes that map to one (or more) of the transcripts associated with the promoter (i.e. the start of the transcript is a member of the cluster of starts that defines the promoter). The expression level of the promoter is then a weighted average of the expression levels of these probes, where a probe that maps to n different transcripts obtains a weight $1/n$. That is, in general, a probe can map to multiple transcripts.

When ISMARA uses RNA-seq for input expression data, it expects the RNA-seq data to be provided as genome alignments of the reads to the hg19 or mm9 genome assemblies in BED format. The loci of the mapped reads are then intersected with the genome alignments of all transcripts that are associated with reference promoters. A

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Microarray	Organism	Producer
HG-U133A	<i>Homo sapiens</i>	Affymetrix
HG-U133B	<i>Homo sapiens</i>	Affymetrix
HG-U133_Plus_2	<i>Homo sapiens</i>	Affymetrix
HG-U133A_2	<i>Homo sapiens</i>	Affymetrix
HuGene-1_0-st-v1	<i>Homo sapiens</i>	Affymetrix
HuGene-1_1-st-v1	<i>Homo sapiens</i>	Affymetrix
HT_HG-U133A	<i>Homo sapiens</i>	Affymetrix
HT_HG-U133B	<i>Homo sapiens</i>	Affymetrix
HT_HG-U133_Plus_PM	<i>Homo sapiens</i>	Affymetrix
Mouse430_2	<i>Mus musculus</i>	Affymetrix
Mouse430A_2	<i>Mus musculus</i>	Affymetrix
MOE430A	<i>Mus musculus</i>	Affymetrix
MOE430B	<i>Mus musculus</i>	Affymetrix
MoGene-1_0-st-v1	<i>Mus musculus</i>	Affymetrix
MoGene-1_1-st-v1	<i>Mus musculus</i>	Affymetrix
HT_MG-430A	<i>Mus musculus</i>	Affymetrix
HT_MG-430B	<i>Mus musculus</i>	Affymetrix
MG_U74Av2	<i>Mus musculus</i>	Affymetrix
MG_U74Bv2	<i>Mus musculus</i>	Affymetrix
MG_U74Cv2	<i>Mus musculus</i>	Affymetrix

Table 4.1: Microarrays currently supported by ISMARA

read is associated with a particular transcript if it falls entirely into any of its exons. We thus unfortunately discard a fraction of reads which originated from exon-exon junctions. However, the alternative of using read mappings to transcripts would require the user to map to the exact same set of transcripts as used by ISMARA and this is impractical. In the future ISMARA may be extended to include mapping of raw reads.

To obtain an expression level for each promoter ISMARA calculates a weighted average over all reads mapping to the transcripts associated with the promoter. The weighting results from multiple mappings at two levels. Firstly, a single read can map to multiple genomic loci and, secondly, a single locus may intersect multiple transcripts that are associated with multiple promoters. When a read maps to n genomic loci, we assign a weight of $1/n$ to each locus. If that locus intersects transcripts of m different promoters, then this reads contributes a final weight of $1/(nm)$ to the expression of each promoter. For a given promoter p and sample s , the total weight w_{ps} is the sum of the weights of all the reads that intersect one of the transcripts associated with promoter p . The expression E_{ps} of promoter p in sample s is then given by

$$E_{ps} = \log \left[\frac{w_{ps}}{N_s} \right], \quad (4.3)$$

where N_s is the total number of reads in sample s , which map to any of the transcripts associated with a reference promoter. Note that this weighting procedure is robust to redundancy in the transcript sets. For example, when a promoter is associated with k highly overlapping transcripts, then a read mapping within the exons of these transcripts will get assigned to all these transcripts, with a weight $1/k$ for each. When the total weight w_{ps} of the promoter is calculated, these k are then summed back and will in the end contribute precisely 1 read. Note also that because ISMARA models promoter expression *changes* across conditions rather than absolute levels, there is no need to account for differences in transcript lengths (i.e. these just cause a shift in log-space which cancels out when considering expression changes).

4.5.6 ChIP-seq data processing

Apart from modeling expression dynamics, ISMARA can also process ChIP-seq data to automatically model chromatin state (or TF binding) changes at promoters genome-wide. Examples of such chromatin state data include histone occupancy, histone modifications, TF binding and DNase1 hypersensitivity in promoter regions. After several experiments, we found that integrating the chromatin signal from a region of 2000 bps centered on the TSS of each promoter gives the most robust results. To obtain a chromatin state level E_{ps} of promoter p in sample s , we calculate the sum r_{ps} of the reads that map entirely within this region around promoter p and transform

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

to the log-space after adding a pseudocount:

$$E_{ps} = \log_2 \left(r_{ps} + \frac{N_s l}{L} \right), \quad (4.4)$$

where the second term is a pseudo-count, N_s is the total number of reads mapped to the genome in sample s (the number of lines in the BED file), $l = 2000$ is the length of the regions, and L is the total length of the genome. Note that this pseudo-count is precisely the number of reads that would be expected if all N_s reads were distributed uniformly over the genome. We set to pseudo-count to this value to make the pseudo-count roughly of the same size as the read-count from background reads in regions where the chromatin mark in question does not appear. The rationale is that, in regions where there are only background reads, statistical fluctuations may cause the read-counts r_{ps} to change significantly from sample to sample. By adding a constant pseudo-count of roughly the same size these fluctuations are effectively dampened. More formally, this pseudo-count results within a Bayesian context if we use a Dirichlet prior with an expected density l/L for each region.

4.5.7 Motif activity fitting.

We model log-expression (or ChIP-seq signal) value E_{ps} of a promoter p in sample s as a linear function of the site-counts N_{pm} for all motifs m associated with the promoter, i.e. either TFBSs in the proximal promoter region or miRNA binding sites in the 3' UTRs of associated transcripts. In each sample s , the contribution of the sites N_{pm} to E_{ps} is given by the (unknown) *motif activity* A_{ms} . That is, we fit a model of the form:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \quad (4.5)$$

where \tilde{c}_s and c_p are sample and promoter-dependent constants, and we assume that the noise is Gaussian distributed with an unknown variance σ^2 that is the same for all promoters and in all samples. Under these assumptions we find the following expression for the likelihood of the expression data given the site-counts, motif activities and sample and promoter-dependent constants:

$$P(E | A, c, \tilde{c}, N, \sigma) \propto \prod_{p,s} \frac{1}{\sigma} \exp \left[-\frac{(E_{ps} - \tilde{c}_s - c_p - \sum_m N_{pm} A_{ms})^2}{2\sigma^2} \right] \quad (4.6)$$

We first maximize this expression with respect to all the constants c_p and \tilde{c}_s , and substitute these with their *maximum likelihood* estimates. After doing this we obtain:

$$P(E | A', N, \sigma) \propto \sigma^{-PS} \exp \left[-\frac{\sum_{ps} (E'_{ps} - \sum_m N'_{pm} A'_{ms})^2}{2\sigma^2} \right], \quad (4.7)$$

4.4.5 Supplementary Methods

where P is the number of promoters, S is the number of samples, the N'_{pm} are a motif-normalized site-counts $N'_{pm} = N_{pm} - \langle N_m \rangle$, with $\langle N_m \rangle$ the average site-count per promoter for motif m , the A'_{ms} are sample-normalized activities $A'_{ms} = A_{ms} - \langle A_m \rangle$, i.e. with $\langle A_m \rangle$ the average activity of motif m across the samples, and the E'_{ps} are sample- and promoter-normalized expression values $E'_{ps} = E_{ps} - \langle E_p \rangle - \langle E_s \rangle + \langle \langle E \rangle \rangle$. That is the log-expression matrix E'_{ps} is normalized such that all its rows and columns sum to zero, the activities A'_{ms} are normalized such that the average activity over all samples is zero, i.e. $\sum_s A'_{ms} = 0$, and the site-counts N'_{pm} are normalized such that the average count over all promoters is zero, i.e. $\sum_p N'_{pm} = 0$.

To avoid over-fitting we assign a symmetric Gaussian prior to each motif activity, i.e. the joint prior for all activities is given by:

$$P(A' \mid \lambda, \sigma) \propto \prod_{ps} \exp \left[-\frac{\lambda^2}{2\sigma^2} \sum_m A'^2_{ms} \right], \quad (4.8)$$

where the constant λ^2 sets the width of prior distribution relative to the width of the likelihood function. Using this prior with the likelihood derived above, the posterior distribution of motif activities takes the form:

$$P(A' \mid E, N, \sigma, \tau) \propto \sigma^{-PS} \exp \left[-\frac{\sum_{p,s} \left((E'_{ps} - \sum_m N'_{pm} A'_{ms})^2 + \lambda^2 \sum_m A'^2_{ms} \right)}{2\sigma^2} \right]. \quad (4.9)$$

Since equation (4.9) factorizes into independent expressions for the different samples, it is enough to consider one sample at a time. The posterior distribution for the motif activities in a particular sample takes the general form of a multi-variate Gaussian centered around A'^*_{ms} :

$$P(A'_s \mid E, N, \sigma) \propto \sigma^{-P} \exp \left[-\frac{\sum_{m\tilde{m}} (A'_{ms} - A'^*_{ms}) W_{m\tilde{m}} (A'_{\tilde{m}s} - A'^*_{\tilde{m}s}) + \chi_s^2}{2\sigma^2} \right], \quad (4.10)$$

where the χ_s^2 is the unexplained part of variance in sample s

$$\chi_s^2 = \sum_p \left(E'_{ps} - \sum_m N'_{pm} A'^*_{ms} \right)^2, \quad (4.11)$$

and the matrix W is given by

$$W_{m\tilde{m}} = \sum_p \left(N'_{pm} N'_{p\tilde{m}} + \lambda^2 \delta_{m\tilde{m}} \right). \quad (4.12)$$

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Finally, the *maximum a posteriori* (MAP) estimates A'_{ms*} can be found by minimizing the expression in the numerator of equation (4.9) using standard numerical procedures for ridge regression. ISMARA performs this calculation by singular value decomposition of the N' matrix.

4.5.7.1 Setting λ through cross-validation

Both the MAP estimates A'_{ms*} , and the matrix $W_{m\tilde{m}}$ are functions of λ . The constant λ^2 represents the ratio between the a priori expected variance of activities, to the average squared-deviation of the model from the expression data (which results from both error in the model, noise in the expression measurements, and biological noise). In general λ will depend on the measurement platform used, i.e. microarray, RNA-seq, or ChIP-seq, and also on the samples used, because the true variance in motif activities will depend on the variance in the E_{ps} across the samples. Thus, the appropriate value of λ will generally not be known in advance and ISMARA therefore includes a method for automatically setting λ from the data. To determine the optimal λ ISMARA uses a 80/20 cross-validation scheme. The set of promoters is divided randomly into two sets, with one containing 80% of all promoters (the ‘training set’) and the other the remaining 20% (the ‘test set’). The training set of promoters is used for fitting the motif activities while the quality of the fit is evaluated on the test set. ISMARA then finds the value of λ that minimizes the average squared-deviation of the expression levels in the test set from those predicted by the model. We denote this optimal value of λ by λ^* .

4.5.7.2 Error bars on motif activities

Apart from the MAP estimates A'_{ms*} ISMARA also determines the uncertainties associated with these estimates. Since σ in Eq. 4.10 is not known, we integrate it out with a suitable scale-invariant prior $P(\sigma) \propto \frac{1}{\sigma}$.

$$\begin{aligned}
 P\left(A'_s \mid E, N, \lambda\right) &= \int_{\sigma=0}^{\infty} P\left(A'_s \mid E, N, \sigma, \lambda\right) P(\sigma) d\sigma \\
 &\propto \frac{\Gamma\left(\frac{P}{2}\right)}{\left[\sum_{m\tilde{m}} \left(A'_{ms} - A'_{ms*}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'_{\tilde{m}s*}\right) + \chi_s^2\right]^{\frac{P}{2}}} \quad (4.13) \\
 &\propto \exp\left[-\frac{P \sum_{m\tilde{m}} \left(A'_{ms} - A'_{ms*}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'_{\tilde{m}s*}\right)}{2\chi_s^2}\right],
 \end{aligned}$$

where the last proportionality is a very good approximation when the number of promoters is large. Note that this is again a multi-variate Gaussian distribution. The

covariance matrix of this Gaussian posterior distribution is given by:

$$C_{m\bar{m};s} = \frac{(W^{-1})_{m\bar{m}} \chi_s^2}{P} \quad (4.14)$$

As is well known, given this multi-variation Gaussian form, the marginal distribution for a single motif activity A'_{ms} will be Gaussian distributed with standard-deviations $\delta A'_{ms}$ given by the square root of the corresponding diagonal term of the covariance matrix, i.e.

$$\delta A'_{ms} = \sqrt{C_{mm;s}} \quad (4.15)$$

We define the overall *significance* of a motif m as the average squared ratio between fitted activities and their standard deviations (z -values)

$$z_m = \sqrt{\frac{1}{S} \sum_s \left(\frac{A'_{ms}}{\delta A'_{ms}} \right)^2}. \quad (4.16)$$

4.5.8 Processing of replicates

Careful studies typically involve experimental replicates to account for the part of variability in the readout which is not under direct experimental control. ISMARA allows users to indicate which samples correspond to replicates and will automatically calculate averaged motif activities and error bars across these replicates. To perform this analysis the user should first upload all samples and perform the standard analysis. On the results page ISMARA provides a link to a page where users can interactively annotate which samples are replicates. In addition, if the replicates came in clearly defined batches, for example, when a time-course was performed multiple times, then the user can also indicate this. Once all samples are annotated ISMARA can then perform motif activity averaging across the replicates. Note that this approach can easily be extended beyond replicates, i.e. the user can arbitrarily divide the samples into groups and ISMARA will automatically calculate average motif activities and associated standard-deviations for each group of samples.

Here we describe how activities within a group are averaged. For a given group G of samples and a particular motif, we assume that its activities A_s in samples $s \in G$ are given by a mean activity \bar{A}^g plus some deviation δ_s , i.e

$$A_s = \bar{A}^g + \delta_s, \quad (4.17)$$

where we assume that the prior probability of δ_s is Gaussian distributed with (unknown) standard-deviation σ_g , i.e

$$P(\delta_s|\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp \left[-\frac{1}{2} \frac{\delta_s^2}{\sigma_g^2} \right]. \quad (4.18)$$

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Thus, given the mean activity \bar{A}^g in the group, the probability to have activity A_s in a particular sample s from the group is

$$P(A_s|\bar{A}^g, \sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{1}{2} \frac{(A_s - \bar{A}^g)^2}{\sigma_g^2}\right]. \quad (4.19)$$

Using the input data, ISMARA has inferred the motif activity A_s to have expected value A_s^* with standard-error δA_s for each sample s . That is, once the dependence on all other activities is integrated out, the probability of the expression data D conditioned on the motif activity A_s is a Gaussian with standard-deviation δA_s , i.e.

$$P(D|A_s) = \frac{1}{\sqrt{2\pi}\delta A_s} \exp\left[-\frac{1}{2} \frac{(A_s - A_s^*)^2}{(\delta A_s)^2}\right]. \quad (4.20)$$

Using the expressions for $P(D|A_s)$ and $P(A_s|\bar{A}^g, \sigma_g)$ we can calculate the probability of the data D given the mean activity \bar{A}^g and standard-deviation σ_t by integrating over all unknown A_s :

$$P(D|\bar{A}^g, \sigma_g) = \prod_{s \in G} \left[\int_{-\infty}^{\infty} P(D|A_s) P(A_s|\bar{A}^g, \sigma_g) \mathbf{d}A_s \right]. \quad (4.21)$$

These integrals can be performed analytically and we obtain

$$P(D|\bar{A}^g, \sigma_g) = \prod_{s \in G} \frac{1}{\sqrt{2\pi(\sigma_g^2 + \sigma_s^2)}} \exp\left[-\frac{(A_s^* - \bar{A}^g)^2}{2(\sigma_g^2 + \sigma_s^2)}\right]. \quad (4.22)$$

Although, formally, we should integrate this expression over the unknown standard-deviation σ_g as well, this integral unfortunately cannot be performed analytically. Therefore, we estimate the integral simply by finding the value σ_g^* that maximizes $P(D|\bar{A}^g, \sigma_g)$. Assuming a uniform prior for the mean activity \bar{A}^g of the samples in the group, we then finally obtain an expression for the posterior probability $P(\bar{A}^g|D)$ which we characterize by its mean $\langle \bar{A}^g \rangle$ and standard-deviation $\delta \bar{A}^g$. That is, $\langle \bar{A}^g \rangle$ is the inferred average motif activity for the samples within the group, and $\delta \bar{A}^g$ is the error-bar on this average activity. This mean and error-bar of the activity for the ‘group’ of samples are given by

$$\langle \bar{A}^g \rangle = \frac{\sum_{s \in G} \frac{A_s^*}{(\sigma_g^*)^2 + \sigma_s^2}}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}, \quad (4.23)$$

and

$$\delta \bar{A}^g = \sqrt{\frac{1}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}}. \quad (4.24)$$

Finally, we assign significances z_m to each motif completely analogously as before, but now averaging over all groups, i.e.

$$z_m = \sqrt{\frac{1}{|G|} \sum_g \left(\frac{\langle \bar{A}^g \rangle}{\delta \bar{A}^g} \right)^2}, \quad (4.25)$$

where $|G|$ is the number of groups. A motif will have a high significance z_m when its motif activities in each group vary little relative to their mean in the group, and are large relative to the original error-bars.

4.5.9 Target predictions

In order to infer motif activities A_{ms} , ISMARA assumes that all promoters with predicted target sites for a motif m will respond to changes in motif activity, i.e. in proportion to the predicted number of sites N_{pm} . This is a reasonable assumption when inferring motif activities, as the activities A_{ms} depend on the statistics of all promoters with sites for motif m . However, in a given condition or system, it is likely that only a subset of the promoters with sites for a motif m are in fact regulated by this regulator. This might be due to a limited accessibility, dependence of particular co-factors, weaker affinity of a site, etcetera. Thus, when we aim to predict individual target promoters of a given motif m , we not only use the binding site predictions N_{pm} , but also evaluate at which promoters the activities A_{ms} contribute to explaining the profiles E_{ps} .

To quantify if a given promoter p is targeted by a motif of interest m we first demand that there exists a TFBS prediction, i.e. $N_{pm} > 0$. Second, we quantify the contribution of m to the fit of the expression/chromatin state profile E_{ps} . The most rigorous approach to quantifying the effect of motif m on promoter p is to calculate both the probability of the entire data set, i.e. the profiles E_{ps} across all promoters and samples, with the original site-count matrix \mathbf{N} , and a site-count matrix $\tilde{\mathbf{N}}$ where only the sites for motif m in promoter p are set to zero. To calculate this probability we treat all the unknown motif activities A_{ms} as well as the standard-deviation σ as nuisance parameters that are integrated out of the likelihood. That is, we formally want to calculate the ratio of probabilities

$$R_{pm} = \frac{\int_{-\infty}^{\infty} \mathbf{dA} \int_0^{\infty} \mathbf{d}\sigma P(E|\mathbf{N}, A, \sigma)}{\int_{-\infty}^{\infty} \mathbf{dA} \int_0^{\infty} \mathbf{d}\sigma P(E|\tilde{\mathbf{N}}, A, \sigma)}, \quad (4.26)$$

where the integrals are over all motif activities A_{ms} , and over the standard-deviations σ . Note that, when we set $N_{pm} = 0$ for promoter p and motif m , we make a very small change to the site-count matrix. That is, as there are tens of thousands of promoters and close to 200 motifs, we are changing only one of the millions of entries

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

in the matrix. As a consequence, the inferred motif activities A'_{ms} that result from the mutated matrix \tilde{N} are likely very close to those that result from the original matrix N . Similarly, the inverse covariance matrix W of the mutated matrix is likely also very close to that of the original matrix and, finally, the optimal values of the constants c_p , \tilde{c}_s , and the prior constant λ^* will also change very little under mutation of the matrix. To make the calculation more tractable we will make the approximation that all these quantities are *unchanged* upon mutation of the matrix. Under that approximation we have

$$P(E|A, N, \sigma, \lambda^*) \propto \sigma^{-PS} \exp \left[-\frac{\sum_{s,m,\tilde{m}} (A'_{ms} - A'_{\tilde{m}s}) W_{m\tilde{m}} (A'_{\tilde{m}s} - A'_{ms}) + \sum_{p,s} \chi_{ps}^2}{2\sigma^2} \right], \quad (4.27)$$

where χ_{ps}^2 is the squared-deviation between the observed value E'_{ps} and the predicted value, i.e.

$$\chi_{ps}^2 = \left(E'_{ps} - \sum_m N'_{pm} A'_{ms} \right)^2 \quad (4.28)$$

And for the probability of the data with the mutated site-count matrix we have

$$P(E|A, \tilde{N}, \sigma, \lambda^*) = P(E|A, N, \sigma, \lambda^*) \exp \left[-\frac{\sum_s (\chi_{psm}^2 - \chi_{ps}^2)}{2\sigma^2} \right], \quad (4.29)$$

where χ_{psm}^2 is the squared-deviation for promoter p and sample s when motif m is removed, i.e.

$$\chi_{psm}^2 = \left(E'_{ps} - \sum_{m'} \tilde{N}'_{pm'} A'_{m's} \right)^2 \quad (4.30)$$

In this form the integrals over the motif activities and σ can be easily performed and we find for the ratio of the probabilities

$$R_{pm} = \left(\frac{\sum_{p',s} \chi_{p's}^2}{\sum_{p',s} \chi_{p's}^2 - \sum_s (\chi_{psm}^2 - \chi_{ps}^2)} \right)^{S(P-M)}, \quad (4.31)$$

where M is the total number of motifs. Since $P \gg M$ we approximate $P - M \approx P$ and we find approximately

$$R_{pm} = \exp \left[\frac{\sum_s (\chi_{psm}^2 - \chi_{ps}^2)}{\langle \chi^2 \rangle} \right], \quad (4.32)$$

where we have defined the average squared-deviation per sample/promoter combination

$$\langle \chi^2 \rangle = \frac{1}{PS} \sum_{p,s} \chi_{ps}^2, \quad (4.33)$$

and made use of the fact that $[1 - x/(SP)]^{-SP} \approx e^x$ for large SP .

In the results shown in the web-server we show, for each predicted target, the logarithm of the likelihood ratio, i.e. the score S_{pm} for motif m targeting promoter p is

$$S_{pm} = \frac{\sum_s \chi_{psm}^2 - \chi_{ps}^2}{\langle \chi^2 \rangle}. \quad (4.34)$$

All targets for which this score is positive, i.e. where removing the motif from the promoter reduces the quality of the fit, are reported.

4.5.9.1 Enriched Gene Ontology categories

To analyze whether there are any Gene Ontology categories whose genes are over-represented among the targets of a motif, we use the “GO::TermFinder” Perl module (145). The ontology files and associations between genes and categories were taken from Gene Ontology (GO) Consortium web-site (48). As a set of target genes for motif m we include all genes associated with promoters that have a target score $S_{pm} > 0$. For microarray chips we create a background set from all the genes which have complementary probes present on the microarray, i.e. have associated probes of the microarray according to our mappings (see Expression data processing). For RNA-seq data we take all genes associated with promoters which have mapped reads. In the web results we display all GO categories with a p -value of 0.01 or less. These p -values are corrected for multiple testing using a simple Bonferroni correction, i.e. multiplied by the number of tests performed.

4.5.10 Principal component analysis of the activities explaining chromatin mark levels

We first performed standard ISMARA analysis on the $n = 10$ data sets measuring expression and 9 different chromatin marks (ChIP-seq), across $S = 8$ cell types (97). For each motif m , and each mark i , we thus obtained estimated activities A_{ms}^i .

We performed principal component analysis (PCA) of the expression and chromatin mark levels across all promoters, separately for each cell type. For a given sample s , let E_{pi} denote the level of mark i at promoter p (suppressing the label s for notational simplicity). We have here already column normalized these levels, i.e.

$$\sum_p E_{pi} = 0, \quad (4.35)$$

for all marks i .

Using singular value decomposition, the matrix $E = U \cdot D \cdot V^T$ can be uniquely decomposed into an orthonormal matrix U (of size $P \times n$), a diagonal positive-semidefinite matrix D (of size $n \times n$), and an orthonormal matrix V (of size $n \times n$)

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

as:

$$E_{pi} = \sum_{k=1}^n U_{pk} D_{kk} V_{ik}, \quad (4.36)$$

where k denotes the index of each component, the column vectors \vec{V}_k with components V_{ik} contain the principal components, and D_{kk}^2 is the fraction of the variance in the E_{pi} values, i.e.

$$\text{var}(E) = \frac{1}{nP} \sum_{p,i} (E_{pi})^2, \quad (4.37)$$

that is explained by component k .

The first principal component \vec{V}_1 , shown in Suppl. Fig. 4.24 top panels, is virtually identical in all cell types and captures approximately 60% of the collective behavior of the expression and 9 chromatin marks (8 histone modification and CTCF binding) across promoters in each sample. As discussed in the main text, this first principal component appears to capture the combination of chromatin mark levels associated with the general ‘activity’ of a promoter. As a consequence, the effect of a given TF on a specific chromatin mark is confounded by its effect on general promoter activity and we therefore decided to subtract it from the activity profiles of all TFs.

For the purpose of removing the first principal component from the motif activities, we will treat each motif m separately and ignore the covariances in the inferred motif activities, i.e. as we assumed previously when calculating the error bars on the motif activities in (4.15). We perform the removal one sample (cell line) at a time. A careful probabilistic analysis must be performed in order to calculate the error bars.

Let’s focus on a given motif m in sample s and denote by A the vector of activities across the marks, i.e. A_i is the activity associated with mark i . In addition, let δA_i denote the standard-deviation (error-bar) of this activity. The posterior distribution $P(A|D)$ of this activity vector given the data is given by a Gaussian, i.e. as in (4.14), of the form

$$P(A|D) \propto \exp \left[-\frac{1}{2} \sum_i \frac{(A_i - A_i^*)^2}{\delta A_i^2} \right], \quad (4.38)$$

where A_i^* is the MAP estimate of the motif activity of mark i . If we introduce a diagonal matrix containing the inverse of the standard-deviation, we can write this expression in matrix-vector form:

$$P(A|D) \propto \exp \left[-\frac{1}{2} (A - A^*)^T \cdot \text{diag} \left(\frac{1}{\delta A^2} \right) \cdot (A - A^*) \right], \quad (4.39)$$

where A^* is a $n \times 1$ vector of the MAP estimates and $\text{diag} \left(\frac{1}{\delta A^2} \right)$ is a 10×10 diagonal precision matrix which elements are set to the inverses of motif activity variances.

4.4.5 Supplementary Methods

Using principal components V of E (4.36) and their orthonormality $V \cdot V^T = \mathbf{1}$ this distribution can be rewritten as

$$P(A|D) \propto \exp \left[-\frac{1}{2} (A - A^*)^T \cdot V \cdot V^T \cdot \text{diag} \left(\frac{1}{\delta A^2} \right) \cdot V \cdot V^T \cdot (A - A^*) \right]. \quad (4.40)$$

We can rewrite the activities in the basis of the principal vectors as $B \equiv V^T \cdot (A - A^*)$ and the precision matrix in the same basis as $M \equiv V^T \cdot \text{diag} \left(\frac{1}{\delta A^2} \right) \cdot V$. In this basis the probability distribution takes the form:

$$P(B|D) \propto \exp \left[-\frac{1}{2} B^T \cdot M \cdot B \right]. \quad (4.41)$$

Note that in this basis, the inverse covariance matrix M contains off-diagonal terms.

We want to integrate out the activities along the first principal component, therefore we separate elements of B and M in the following way

$$B = \left(\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array} \right) \equiv \begin{pmatrix} b_1 \\ B_y \end{pmatrix} \quad (4.42)$$

$$M = \left(\begin{array}{c} m_{11} \\ m_{21} \\ \vdots \\ m_{n1} \end{array} \right) \left(\begin{array}{ccc} m_{12} & \cdots & m_{1n} \\ m_{22} & \cdots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n2} & \cdots & m_{nn} \end{array} \right) \equiv \begin{pmatrix} m_{11} & M_y^T \\ M_y & M_w \end{pmatrix}, \quad (4.43)$$

and the last equivalency holds because the matrix M is symmetric.

Using these definitions, eq. (4.41) can be expanded and rewritten to obtain:

$$\begin{aligned} P(B|D) &\propto \exp \left[-\frac{1}{2} (b_1^2 m_{11} + 2b_1 B_y^T \cdot M_y + B_y^T \cdot M_w \cdot B_y) \right] \\ &= \exp \left[-\frac{1}{2} \left(m_{11} \left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}} \right)^2 + B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}} \right) \right] \end{aligned} \quad (4.44)$$

Where we reordered terms and completed the square to bring out that this posterior is proportional to a Gaussian with respect to b_1 . It is now straightforward to integrate

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

this probability distribution along the first principal direction:

$$\begin{aligned}
 P(B_y|D) &= \int_{b_1=-\infty}^{\infty} P(B|D) \mathbf{d}b_1 \propto \exp \left[-\frac{1}{2} \left(B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}} \right) \right] \cdot \\
 &\quad \cdot \int_{b_1=-\infty}^{\infty} \exp \left[-\frac{1}{2} m_{11} \left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}} \right)^2 \right] \mathbf{d}b_1 \\
 &\propto \exp \left[-\frac{1}{2} B_y^T \cdot \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}} \right) \cdot B_y \right], \quad (4.45)
 \end{aligned}$$

The last proportionality holds because the Gaussian integral yields a constant (with respect to B_y). Since the covariance matrix is the inverse of the precision matrix, the covariance matrix W in the reduced $(n-1)$ -dimensional space (i.e. without the first principal direction) has the form:

$$W = \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}} \right)^{-1} \quad (4.46)$$

Finally, this covariance matrix W needs to be transformed back from the principal component basis to the original basis. To this end we use the principal components contained in columns 2 through n of the V matrix. We obtain for the final covariance matrix K in the original basis

$$K_{ij} = \sum_{k,l=2}^n V_{ik} W_{kl} V_{jl}. \quad (4.47)$$

The standard deviation of activities of the i^{th} mark is given by square root of the corresponding diagonal element of this matrix

$$\delta \tilde{A}_i = \sqrt{K_{ii}}. \quad (4.48)$$

The corrected MAP activities are obtained by first defining

$$B^* = V^T \cdot A^*, \quad (4.49)$$

and then transforming back to the original basis using only the components along principal vectors 2 through n :

$$\tilde{A} = \sum_{k=2}^n V_{ik} B_k^*. \quad (4.50)$$

4.4.6 Fraction of variance explained by the fit

The reported z -value of the i^{th} mark (we introduce back the indices for motif m and sample s omitted previously) is given by

$$z_{ms}^i = \frac{\tilde{A}_i}{\delta \tilde{A}_i} \quad (4.51)$$

After removing the contribution of the first principal component to the motif activities, we re-calculated significance z -values z_m^i for each motif m and each mark i (x-axis in the Suppl. Fig. 4.25)

$$z_m^i = \sqrt{\frac{\sum_{s'} (z_{ms'}^i)^2}{S}}. \quad (4.52)$$

In addition, we calculated a specificity s_m^i which measures the fraction of the overall that is associated with mark i (y-axis in the Suppl. Fig. 4.25)

$$s_m^i = \frac{z_{mk}^2}{\sum_{k'} z_{mk'}^2}. \quad (4.53)$$

That is, a motif m will be highly specific for mark i if it has a high z -value z_m^i , and low z -values for all other marks.

4.6 Fraction of variance explained by the fit

The total variance V in a data set is given by the sum of the squared normalized expression values

$$V = \frac{1}{PS} \sum_{p,s} (E'_{ps})^2. \quad (4.54)$$

After fitting the model, the average squared deviation left unexplained is given by the average of χ_{ps}^2 across all promoters and samples, i.e. as defined by equations (4.28) and (4.33). The fraction of the variance f explained by the fit is thus

$$f = 1 - \frac{\langle \chi^2 \rangle}{V}. \quad (4.55)$$

We find that the fraction of variance explained by the fit typically ranges between 4% and 14%. As an illustration, Suppl. Fig. 4.8 shows a histogram of the fraction of variance explained by the model across all samples in the GNF data set.

The fraction of variance explained in the samples of the second data set (human GNF atlas plus NCI-60 cell lines) is a bit larger than the fraction of variance explained in the samples of data set one (the mouse GNF atlas). It appears that this increase results from the fact that there is a relatively large (and explainable) difference in the expression profiles of the cancer cell lines and the normal cell lines.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

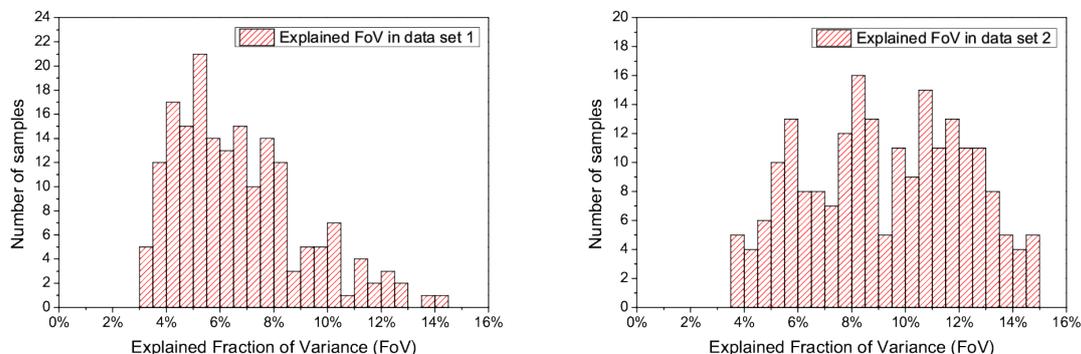


Figure 4.8: Histogram of the fraction of variance explained by the model. **Left panel:** Histogram of the fraction of variance explained for the samples in the mouse GNF atlas (data set 1). **Right panel:** Histogram of the fraction of variance explained for the samples from the human GNF atlas and the NCI-60 cancer cell lines (data set 2).

4.7 Overview of results presented in the web-interface

To illustrate the results that ISMARA provides, we here present a number of figures, that show examples of results on the mouse GNF atlas. Note that almost all of these figures are screen shots from the actual web-interface. All the full results for the mouse GNF data are available at http://ismara.unibas.ch/supp/dataset1/ismara_report.

The main page of results that ISMARA provides for a given data set centers around a list of motifs, sorted by their significance, showing for each motif its significance, the associated TFs, a sequence logo of the motif, and a thumbnail image of its inferred activity across the samples. Supplementary Fig. 4.9 shows an excerpt from this list of motifs.

Each motif name in this list is in fact a link to a separate page with much more extensive results for the motif. Among these more extensive results is, first of all, a figure showing the inferred motif activity (and error bars) across all samples, where the samples are ordered according from left to right, according to the user's input. Supplementary Fig. 4.10 shows the activity profile of the E2F1..5 motif across the mouse GNF samples. Note that such an ordering motif activity across samples is especially helpful when the samples come from a time course, in which case the graph shows the motif activity across time.

However, in many cases, including the GNF atlas analyzed here, there is no preferred natural ordering of the samples. In those cases it is more natural to present the motif activities with samples sorted from those in which the motif is most significantly upregulated, to those where it is most significantly downregulated. ISMARA provides such a list of motif z -values, with sample sorted from largest to smallest z -

4.4.7 Overview of results presented in the web-interface

ISMARA results for GNF SymAtlas, mouse (Lattin, 2008)

ISMARA - Integrated System for Motif Activity Response Analysis is a free online tool that recognizes most important transcription factors that are changing their activity in a set of samples.

All motifs sorted by activity significance

Search:

Showing 1 to 274 of 274 entries

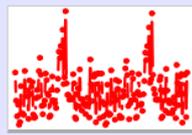
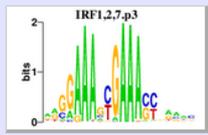
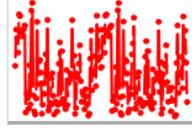
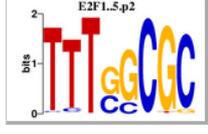
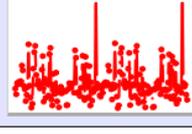
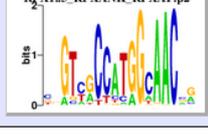
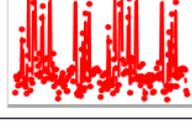
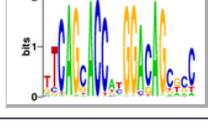
Motif name	Z-value	Associated genes	Profile	Logo
IRF1.2.7.p3	7.114	Irf1 (Irf-1) Irf2 (Irf-2) Irf7		
E2F1.5.p2	6.191	E2f4 E2f5 (E2F-5) E2f2 E2f1 (E2F-1) E2f3 (E2F3b, E2f3a)		
RFX1.5_RFXANK_RFXAP.p2	5.544	Rfx1 Rfxap Rfx5 Rfxank (Tv1) Rfx4 Rfx3 (MRF3) Rfx2		
REST.p3	5.393	Rest (NRSF)		

Figure 4.9: Fragment of the list of regulatory motifs sorted by their significance (z -score). The motifs are sorted from top to bottom. Shown for each motif are, from left to right, the name of the motif (which is a link to a separate page with results for the motif), its z -score, a list of associated TFs (links to NCBI pages for these genes), a thumbnail of the inferred motif activity profile, and the sequence logo of the motif.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

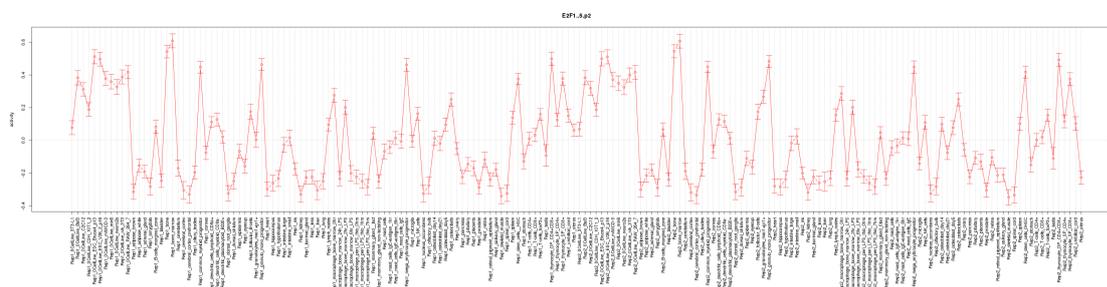


Figure 4.10: Inferred activities of the E2F1..5 motif on the mouse GNF atlas. The samples are ordered, from left to right, in lexicographic order, according to sample names input by the user. The red circles show the estimated activities A_{ms}^* and the error-bars δA_{ms} are shown as red vertical bars. Samples names are indicated on the bottom.

value, as shown in Suppl. Fig. 4.11 for the E2F1..5 motif. In this case, inspection of this sorted list of samples makes clear that E2F is highly upregulated in fast dividing cells, and downregulated in post-mitotic cells. The close association of E2F activity with cell proliferation is something that we have observed across many different data sets (data not shown).

The next important information provided for each motif, is a predicted list of target promoters. ISMARA provides the target promoters p for a motif m sorted by their target score S_{pm} (see section 4.5.9). As an example, the list of targets for the E2F1..5 motif is shown in Suppl. Fig. 4.12. Each row in the table corresponds to one target promoter and information shown includes the promoter ID, its score S_{pm} , associated transcripts and Entrez gene, and a description of the gene. Note that all these pieces of information are links that take the user to additional information on the promoter, the associated transcripts and gene. Note that, to keep the page easily viewable, by default only the top 20 targets are shown. But the user can interactively change the number of targets shown in the list. In addition, a search box allows the user to search whether a particular promoter, transcript, or gene of interest occurs within the full list of targets.

Of particular interest is the additional information provided about each promoter, through the links with the promoter IDs. Following this link takes the user to the genome browser of our SwissRegulon database (146), showing the section containing the proximal promoter regions (500 base pairs up-stream and down-stream of the major TSS of the promoter). In this browser the user is shown all the predicted TFBSs that are used by ISMARA in its modeling of expression or ChIP-seq data. This thus allows the user to determine the precise locations of the TFBSs on the genome, through which a particular TF is predicted to target a given promoter. Supplementary Fig. 4.13 shows, as an example, the promoter of the Rrm2 gene,

4.4.7 Overview of results presented in the web-interface

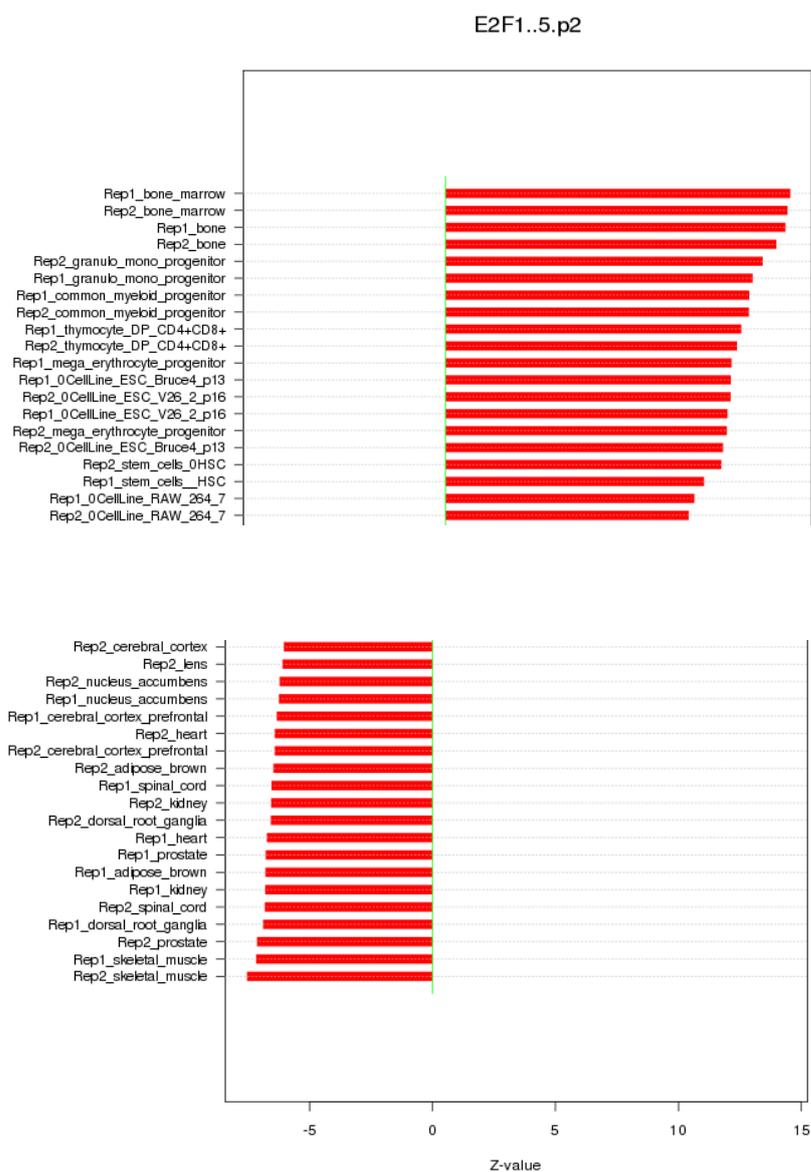


Figure 4.11: Sorted list of z -values for the E2F motif across all samples of the mouse GNF atlas. For readability, only the top 20 and bottom 20 samples are shown. Note that the samples with the highest z -values correspond to fast proliferating cells whereas the samples with the lowest z -values correspond to non-proliferating cells.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Top 20 targets:

Search:

Show **20** entries

Showing 1 to 20 of 200 entries

Promoter	Score	Refseq	Gene	Description
chr12_+_25393083	205.825	NM_009104	Rrm2	ribonucleotide reductase M2
chr10_-_68815606	190.395	NM_007659	Cdk1	cyclin-dependent kinase 1
chr1_-_130256055	178.689	NM_008567	Mcm6	minichromosome maintenance deficient 6 (MIS5 homolog, <i>S. pombe</i>) (<i>S. cerevisiae</i>)
chr15_-_57966551	170.642	NM_027435	Atad2	ATPase family, AAA domain containing 2
chr8_+_77633426	170.047	NM_008566	Mcm5	minichromosome maintenance deficient 5, cell division cycle 46 (<i>S. cerevisiae</i>)
chr10_+_110182506	159.597	NM_178609	E2f7	E2F transcription factor 7
chr1_-_20810238	152.621	NM_008563	Mcm3	minichromosome maintenance deficient 3 (<i>S. cerevisiae</i>)
chr10_-_20880559	148.266	NM_001198914 NM_010848	Myb	myeloblastosis oncogene
chr5_-_138613028	146.050	NM_008568	Mcm7	minichromosome maintenance deficient 7 (<i>S. cerevisiae</i>)
chr8_+_125091889	138.459	NM_026014	Cdt1	chromatin licensing and DNA replication factor 1
chr2_+_72314235	134.555	NM_025866	Cdc6	cell division cycle associated 7
chr11_+_98769122	126.897		Cdc6	cell division cycle 6 homolog (<i>S. cerevisiae</i>)
chr13_-_21925343	122.491	NM_178185 NM_001177544	Hist1h2ah Hist1h2ap Hist1h2al Hist1h2ao	histone cluster 1, H2ah histone cluster 1, H2ap histone cluster 1, H2al histone cluster 1, H2ao
chr6_-_88848650	116.296	NM_008564	Mcm2	minichromosome maintenance deficient 2 mitotin (<i>S. cerevisiae</i>)
chr11_+_98769162	115.939	NM_001025779	Cdc6	cell division cycle 6 homolog (<i>S. cerevisiae</i>)
chr13_-_21879086	111.341	NM_178184	Hist1h2an	histone cluster 1, H2an
chr13_-_22134795	106.518	NM_178186	Hist1h2ag Hist1h2al	histone cluster 1, H2ag histone cluster 1, H2al
chr10_+_127669118	104.620	NM_001136082 NM_001164080 NM_001164081	Timeless	timeless homolog (<i>Drosophila</i>)
chr2_-_157030236	103.984	NM_001139516 NM_011249	Rbl1	retinoblastoma-like 1 (p107)
chr10_+_127669154	103.485		Timeless	timeless homolog (<i>Drosophila</i>)

Figure 4.12: Top target promoters of the E2F1..5 motif for the mouse GNF atlas. Targets are sorted by the log-likelihood score S_{pm} . Shown for each target promoter are the promoter ID (a link to the SwissRegulon web-browser page showing the promoter on the genome), the target score S_{pm} , associated RefSeq transcripts, associated gene symbols (links to NCBI pages), and gene names (which often provide a short description of the gene's function). By default the top 20 targets are shown but this number can be changed using the drop-down menu at the top of the table. A search box allows users to search for genes or transcripts within the entire target list.

4.4.7 Overview of results presented in the web-interface

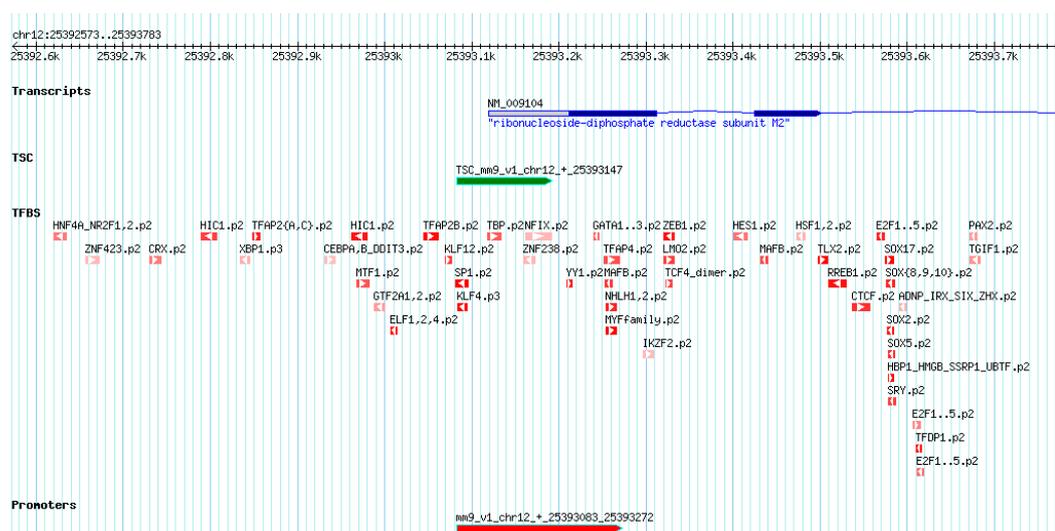


Figure 4.13: Example of a promoter region as display in the SwissRegulon genome browser. The region shown corresponds to the proximal promoter of the *Rrm2* gene (the top target of the E2F1.5 motif) and this is the region that will be displayed when following the link to the promoter as displayed in Suppl. Fig. 4.12. The genome browser shows the RefSeq transcript, the promoter, the associated annotated transcript start cluster (TSC) based on the CAGE data, and all the predicted TFBSs. Here the intensity of the color indicates the posterior probability assigned to each site, and the name of the cognate motif is written above each side. The arrows inside the TFBSs indicate on which strand the motif occurs.

which is the top predicted target of the E2F1.5 motif.

Beyond a list of individual targets, a user would typically like to gain some intuition of the pathways and particular biological processes that are targeted by a particular motif. One way of visualizing the functional structure of the predicted targets of a motif, is to represent these as a network, with links between pairs of genes that are known to be functionally related. The STRING database (47) maintains a curated collection of functional links between proteins, where ‘functional link’ can range from direct physical interaction, to over-representation of the protein pair within abstracts of scientific articles. For any set of proteins, STRING provides visualizations of the network of known functional interactions between these proteins, which intuitively brings out groups of proteins known to be functionally related. ISMARA provides, for each motif, such STRING network pictures of the set of predicted targets of the motif (for visibility at most the top 200 targets are shown). Supplementary Fig. 4.14 shows the STRING network for the predicted targets of E2F1.5. Note that the picture is itself a link to the STRING database, where the figure is interactive and allows the user more detailed information on each of the proteins in the network and each

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

functional link between the proteins.

Apart from the STRING network, ISMARA also provides list of Gene Ontology categories that are enriched among the predicted targets of a motif. Lists are provided for the ‘biological process’, ‘cellular component’, and ‘molecular function’ hierarchies. A p -value for enrichment is calculated using a simple hypergeometric test and only categories with a p -value below 0.05 are shown. The categories are sorted by the fold-enrichment of targets relative to what would be expected by chance. As an example, Suppl. Fig. 4.15 shows the top categories of the biological process hierarchy for the E2F1..5 motif.

For many of the motifs incorporated into the ISMARA analysis, there is more than one TF that can potentially bind to sites for the motif. As a consequence, it is not always clear which individual TFs are responsible for the observed motif activity in a particular system. To help determine which TFs are most likely involved in the activity of a given motif, ISMARA provides some simple correlation analysis. In particular, a table is provided showing the Pearson correlation between the motif’s activity profile and the mRNA expression profiles of each of the TFs that can bind to the sites of the motif. The TFs in the list are sorted by their p -value. Supplementary Fig. 4.16 shows the list of correlations for the E2F TFs.

For each of the correlations a link is also provided to a simple scatter plot showing the mRNA expression levels and motif activities across the samples. Supplementary Fig. 4.17 shows example scatter plots for the TFs E2F1 and E2F2, which are both significantly correlated with motif activity. The fact that both motifs correlate *positively* with motif activity strongly suggests that these TFs act as activators, i.e. as their mRNA levels go up, the expression of target genes is affected positively. To show an example of opposite behavior, Suppl. Fig. 4.18 shows the mRNA expression levels of the TF REST against its inferred motif activity, across the mouse GNF samples. The clear negative correlation strongly suggests that REST acts as a *repressor* of its targets, and this is indeed well-known to be the case.

Finally, one of our our aims is to understand the causal structure of the transcription regulatory network, and a first step in that direction are predictions of direct regulatory interactions between the motifs. For each motif, we check its list of predicted targets for promoters of TFs that are associated with other motifs. Using this we build a regulatory network where nodes correspond to motifs and a directed edge from motif m to motif m' occurs whenever a promoter of at least one of the TFs associated with motif m' is a predicted target of motif m . On the page with results of a given motif, a part of this regulatory network centered around the motif in question is shown, i.e. all edges from or to the motif in question as well as edges between the direct neighbors of the motif. Supplementary Fig. 4.19 shows this network for the E2F1..5 motif. Note that a slider on the left-hand side of the network allows the user to vary a cut-off on the target score S_{pm} , i.e. showing only nodes and edges over the cut-off. In addition, placing the mouse pointer over a node brings up a pop-up with

4.4.7 Overview of results presented in the web-interface

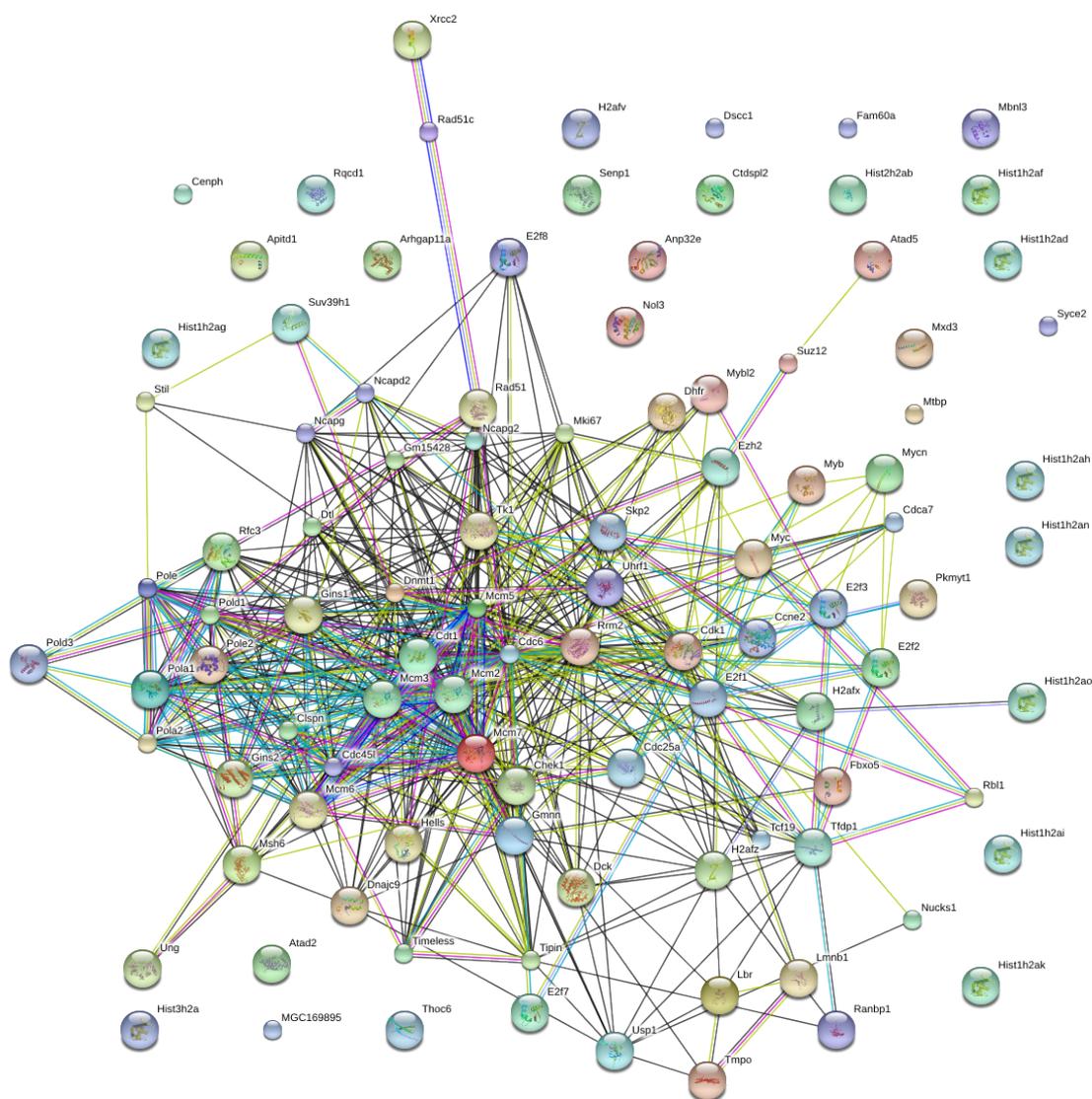


Figure 4.14: Network of target genes of the E2F motif as displayed by the STRING database (47). Each node corresponds to a predicted target gene of the E2F1.5 motif (in the mouse GNF atlas, i.e. data set 1). Links are drawn by STRING whenever there is any evidence that the two genes may interact or be functionally linked, where evidence may range from measured direct protein-protein interaction to significant co-occurrence of the gene names within abstracts of articles.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Gene overrepresentation in process category:

Search:

Show **10** entries

Showing 1 to 20 of 160 entries

enrichment	p-value	GO term	description
18.42	1.97e-02	GO:0019985	translesion synthesis
16.57	8.61e-08	GO:0006270	DNA-dependent DNA replication initiation
15.07	4.51e-07	GO:0071897	DNA biosynthetic process
13.15	2.04e-02	GO:0000731	DNA synthesis involved in DNA repair
12.28	6.83e-05	GO:2000104	negative regulation of DNA-dependent DNA replication
11.84	1.41e-05	GO:0090329	regulation of DNA-dependent DNA replication
11.05	1.00e-02	GO:0000084	S phase of mitotic cell cycle
11.05	1.00e-02	GO:0006268	DNA unwinding involved in replication
10.05	2.10e-02	GO:0051320	S phase
9.69	2.87e-05	GO:0006297	nucleotide-excision repair, DNA gap filling
8.29	8.81e-04	GO:0007062	sister chromatid cohesion
7.37	1.18e-02	GO:0032392	DNA geometric change
7.21	2.20e-08	GO:0006261	DNA-dependent DNA replication
6.82	2.13e-29	GO:0006260	DNA replication
6.82	1.76e-03	GO:0033261	regulation of S phase
6.31	3.67e-04	GO:0008156	negative regulation of DNA replication
5.84	3.20e-04	GO:0043966	histone H3 acetylation
4.93	4.31e-04	GO:0006289	nucleotide-excision repair
4.87	1.33e-03	GO:0051053	negative regulation of DNA metabolic process
4.72	9.25e-06	GO:0006473	protein acetylation

Figure 4.15: Top over-represented categories from the Gene Ontology hierarchy of biological processes among the predicted targets of the E2F1..5 motif. The categories are sorted by their enrichment, i.e. how much more frequent targets from this category are than expected by chance (first column) and only categories that are significantly enriched at a p -value of 0.05 (second column) are shown. The third and fourth columns in the table show the GO identifier and a description of the categories and these are again links to pages with more extensive information on the GO category. Finally, the user can interactively change the number of top categories shown using the drop-down menu or search for keywords.

4.4.7 Overview of results presented in the web-interface

Activity-expression correlation:

Gene	Promoter	Pearson	P-value	Plot
E2f1	chr2_-154394864	0.69	1.3e-26	Click!
E2f2	chr4_+135728600	0.68	1.7e-26	Click!
E2f3	chr13_-30077931	0.63	7.5e-22	Click!
E2f4	chr8_+107828491	0.53	1.8e-14	Click!
E2f5	chr3_+14578783	-0.02	7.5e-01	Click!

Figure 4.16: Correlations between the E2F1..5 motif activity and mRNA expression profiles of TFs that kind bind to sites of the motif. The table shows the names of the associated TF genes, the IDs of the associated promoters of these genes, the Pearson correlation coefficient, the p -value for the correlation, and a link to a figure showing a scatter of the motif activity and mRNA expression levels across the samples (Suppl. Fig. 4.17) below.

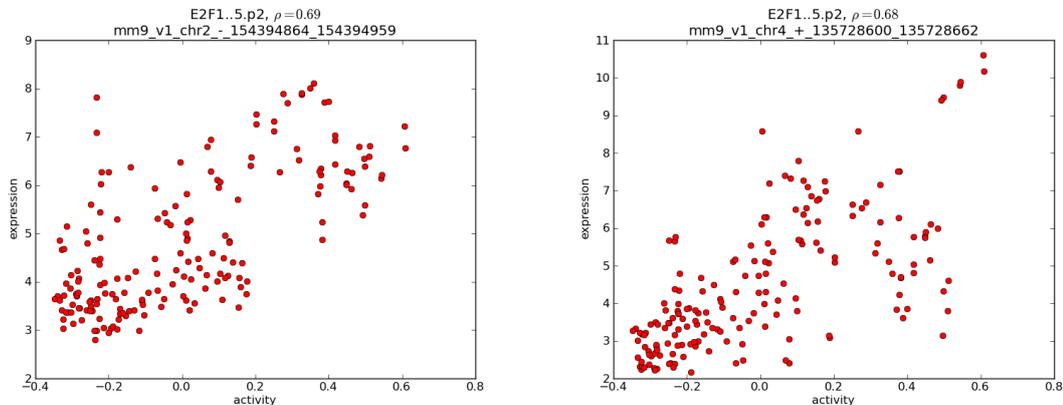


Figure 4.17: Example scatter plots showing the correlations between E2F1..5 motif activity and the mRNA expression of the E2F1 (left panel) and E2F2 (right panel) TFs, across the samples of the mouse GNF atlas. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

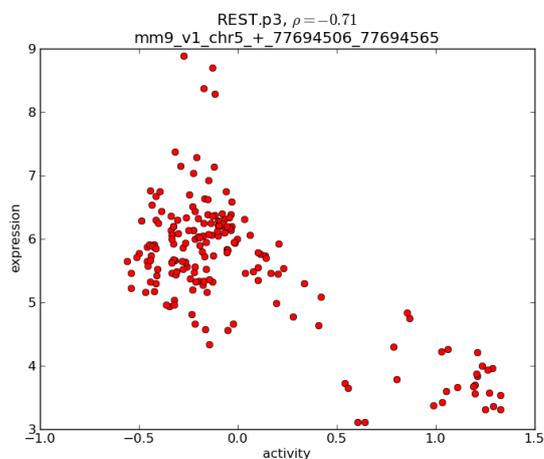


Figure 4.18: Scatter plots showing the correlation between REST motif activity and the mRNA expression of the REST TF, across the samples of the mouse GNF atlas. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

the z -value of the motif, and placing the mouse pointer on an edge will bring up a pop-up with the target score of the link.

4.4.7 Overview of results presented in the web-interface

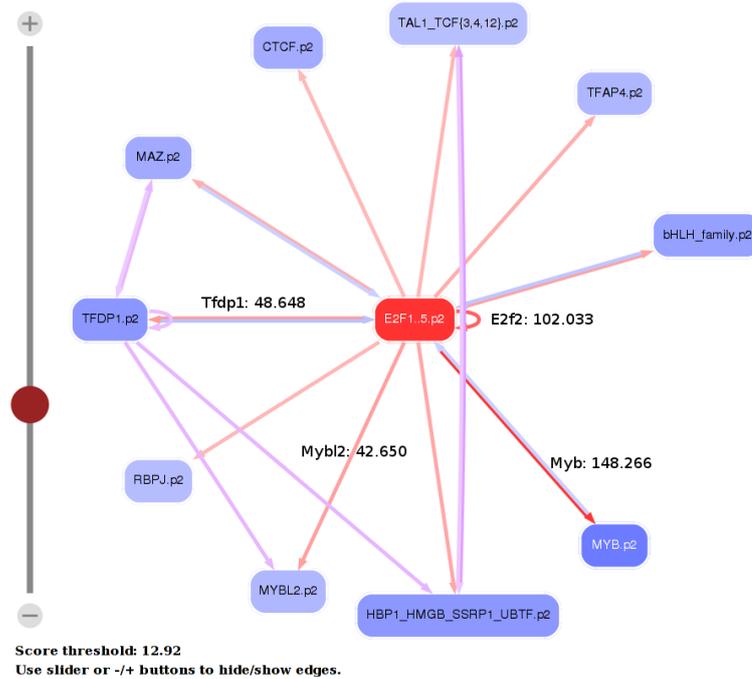


Figure 4.19: Predicted direct regulatory interactions between E2F and other motifs. Edges are drawn from motif m to m' whenever a promoter p , associated with motif m' , is a predicted target of motif m , with a target score S_{pm} larger than a given cut-off c . In the web browser, the user can interactively change the cut-off c using the slider on the left of the figure. In this example the cut-off was set at 12.92. When the cursor is placed on an edge the target score S_{pm} is shown, and in this figure the target scores of the 4 most significant targets are shown. The intensity of the color of each motif corresponds to its z -score. Finally, for each motif (in this case E2F1.5) only the direct neighborhood in the network is shown, i.e. edges that are directly linked to E2F1.5, or that link between motifs that directly link to E2F1.5.

4.8 HNF1a activity in pancreas

Besides its well-known role in liver and kidney, ISMARA also predicts that HNF1a is one of the most active motifs in pancreas. Supplementary Fig. 4.20 shows the motifs with the most positive and most negative z -values in the two pancreas samples of the mouse GNF atlas.

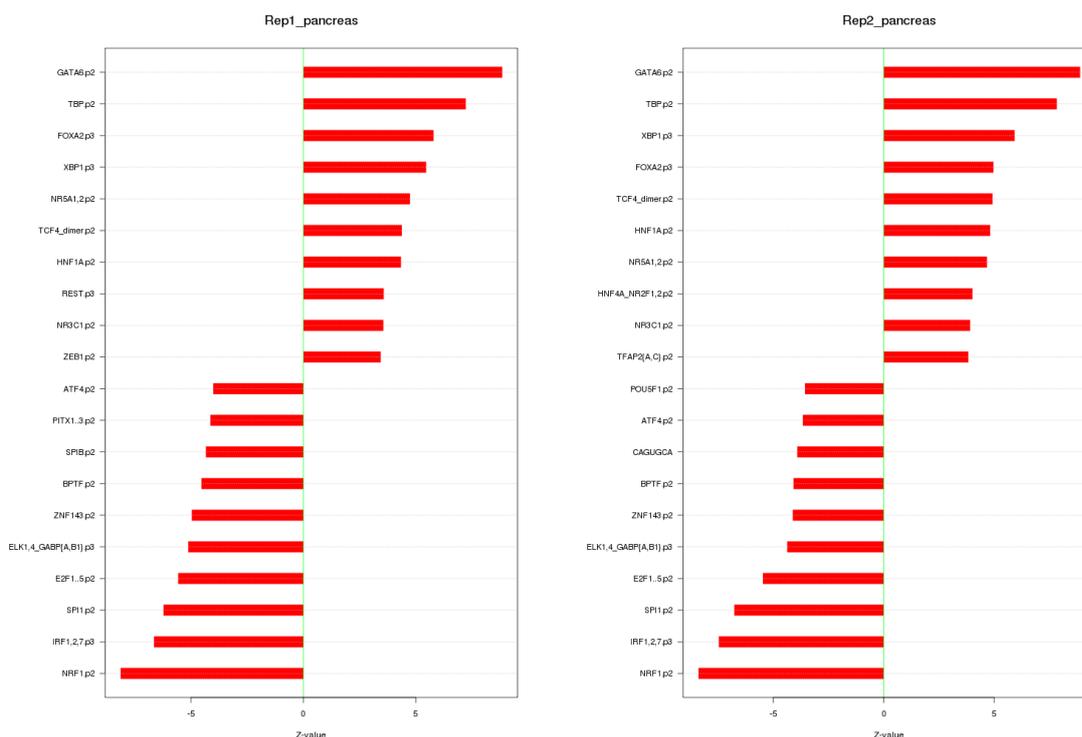


Figure 4.20: Motifs with the most positive and most negative z -values in two replicate pancreas samples from the mouse GNF atlas. Note that the HNF1a is the 7th and 6th most upregulated motif, respectively, in these samples.

4.9 Reproducibility of motif activities

The inferred motif activities depend both on our binding site predictions, and on the assumed simple linear relationship between predicted numbers of sites and mRNA expression. As explained in the main text, there are many reasons why such a ‘cartoon’ model is very unlikely to produce an accurate quantitative model of genome-wide expression profiles. As a consequence, one may wonder how robust the inferred motif activities are. However, as shown in Suppl. Figure 4.21, the motif activities inferred

4.4.9 Reproducibility of motif activities

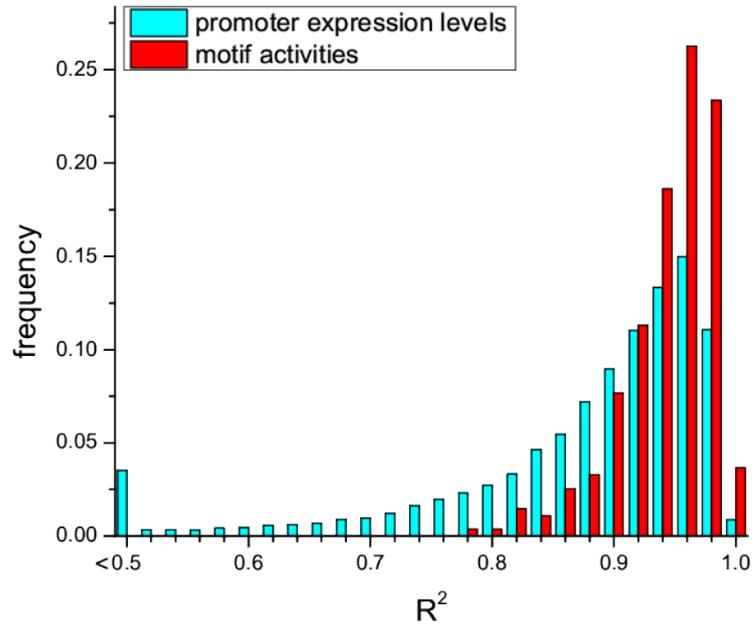


Figure 4.21: Reproducibility of the inferred motif activities and the expression profiles of promoters. For each motif, and each promoter, we calculated the Pearson correlation coefficient of the activity/expression profiles for the two replicates of the samples in the mouse GNF atlas. The figure shows the distribution of observed correlation coefficients for the motif activities (red) and promoter expression profiles (blue). The motif activities are generally considerably more reproducible than the expression profiles of the promoters from which they are inferred.

from the two replicates of the mouse GNF atlas are typically more reproducible across these replicates than the expression levels of individual promoters which are used to infer the motif activities. The reason for this is that the motif activity is inferred from the behavior of the hundreds to thousands of predicted targets of the motif. Thus, although at each individual promoter the expression is likely a complex function of the regulatory sites and the linear model is likely a poor approximation, these complications are effectively averaged out when inferring motif activities from the joint behavior of all targets.

4.10 Motifs dis-regulated in tumor cells

To identify motifs whose motif activities are consistently dis-regulated in tumors, we first separate all samples s from the GNF and NCI-60 data sets into the set of tumor samples T and non-tumor samples N . Next, we use the replicate averaging described in section 4.5.8 to calculate, for each motif, an average activity $\langle \bar{A}^T \rangle$ in tumor samples, an associated error-bar $\delta \bar{A}^T$, an average activity in non-tumor samples $\langle \bar{A}^N \rangle$, and an error-bar $\delta \bar{A}^N$ associated with the average activity in non-tumor samples. From these, we calculate a z -value z_m for each motif m that quantifies the significance of the difference in the average activities in tumor and non-tumor samples. Tables 4.2 and 4.3 show the motifs with highest and lowest z -values, respectively. That is, these are the motifs most significantly dis-regulated in tumor cells.

Motif	z -values
bHLH_family.p2	2.398858
HIF1A.p2	2.230493
E2F1..5.p2	2.140652
ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	2.071274
BPTF.p2	1.977484
NFY{A,B,C}.p2	1.920594
FOXD3.p2	1.915846
TFDP1.p2	1.901083
ELF1,2,4.p2	1.874818
ZNF143.p2	1.802732
ATF4.p2	1.786143
YY1.p2	1.735238
EHF.p2	1.718308
NRF1.p2	1.674024
ELK1,4_GABP{A,B1}.p3	1.667680
CCUUCAU (hsa-miR-205)	1.525379
PAX5.p2	1.500615
UCAAGUA (hsa-miR-26a, hsa-miR-26b, hsa-miR-1297, hsa-miR-4465)	1.404557
BACH2.p2	1.371868
GUAACAG (hsa-miR-194)	1.349047
HES1.p2	1.317505

Table 4.2: Motifs that are most consistently upregulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their z -value (shown in the second column).

4.4.10 Motifs dis-regulated in tumor cells

Motif	<i>z</i> -values
SMAD1..7,9.p2	-2.194113
HAND1,2.p2	-2.185943
TGIF1.p2	-2.117814
MAZ.p2	-2.076224
TFCP2.p2	-2.071225
KLF12.p2	-1.958392
GGCUCAG (hsa-miR-24)	-1.918863
FOX{D1,D2}.p2	-1.839199
TBX4,5.p2	-1.805228
FOXP3.p2	-1.740035
EVI1.p2	-1.701934
HBP1_HMGB_SSRP1_UBTF.p2	-1.688854
AAAGUGC (hsa-miR-17, hsa-miR-20a, hsa-miR-20b, hsa-miR-93, hsa-miR-106a, hsa-miR-106b, hsa-miR-519d)	-1.628037
GAGAUGA (hsa-miR-143, hsa-miR-4770)	-1.619611
HIC1.p2	-1.607936
NANOG{mouse}.p2	-1.576193
FEV.p2	-1.574951
MYOD1.p2	-1.565920
NR1H4.p2	-1.562673
POU1F1.p2	-1.556216
TCF4_dimer.p2	-1.536692
MYFfamily.p2	-1.514719
TAL1_TCF{3,4,12}.p2	-1.499900
POU5F1.p2	-1.480033
NR3C1.p2	-1.473553
HOX{A5,B5}.p2	-1.440485
STAT1,3.p3	-1.417964
GTF2A1,2.p2	-1.416557
RORA.p2	-1.391819
CAGCAGG (hsa-miR-214, hsa-miR-761, hsa-miR-3619-5p)	-1.356781
ETS1,2.p2	-1.337667
EN1,2.p2	-1.337051
AR.p2	-1.330996
RREB1.p2	-1.330444
CUCCCAA (hsa-miR-150)	-1.318296
CACAGUG (hsa-miR-128)	-1.318135
JUN.p2	-1.313498

Table 4.3: Motifs that are most consistently down-regulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their *z*-value (shown in the second column).

4.11 XBP1 motif activity and mRNA expression

The XBP1 motif is the third most significant motif in the innate immune response time course in which HUVEC cells were treated with $\text{TNF}\alpha$. The motif is upregulated during the time course. However, as shown in Suppl. Fig. 4.22, the mRNA expression of the XBP1 gene is almost constant across the time course, and not significantly correlated with the motif's activity. In fact, it has been established that XBP1's activity is regulated post-transcriptionally, i.e. through alternative splicing (82; 83).

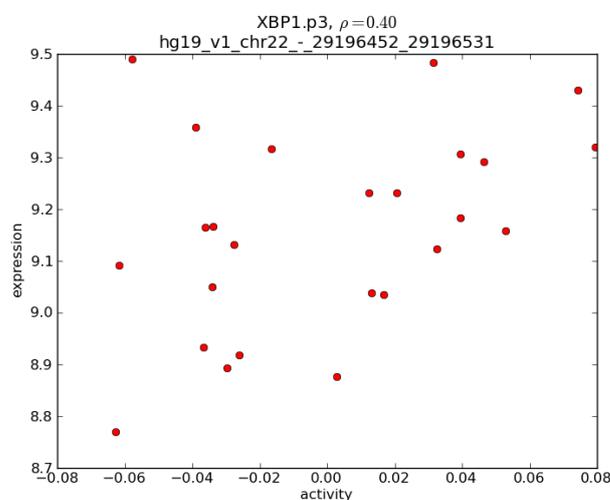


Figure 4.22: Scatter plot showing the correlation between the inferred activity of the XBP1 motif and the mRNA expression of the XBP1 gene for the innate immune response time course. The mRNA expression is shown on a logarithmic scale (base 2) along the vertical axis. Note the small range in expression variation.

4.12 Analysis of the ENCODE ChIP-seq data

To illustrate ISMARA's performance on ChIP-seq data we used data from the ENCODE project in which expression and 9 different chromatin modifications were measured across 8 different cell types (97). Supplementary table 4.4 shows the list of cell types used together with their description and Suppl. table 4.5 shows a list of all the signals that were measured. For simplicity, we will refer to all 10 signals (which include expression and the binding of the CTCF transcription factor) as 'marks' in our description below.

We first ran ISMARA separately on the data sets for each of the 10 signals. For all the ChIP-seq data we thus modeled the occurrence of each of the marks at promoters

4.4.12 Analysis of the ENCODE ChIP-seq data

Cell	Description
GM12878	B-lymphocyte, lymphoblastoid
HepG2	hepatocellular carcinoma
HMEC	mammary epithelial cells
HSMM	skeletal muscle myoblasts
Huvec	umbilical vein endothelial cells
K562	chronic myelogenous leukemia
NHEK	epidermal keratinocytes
NHLF	lung fibroblasts

Table 4.4: Human tissues and cell lines used as the source of experimental material in the ENCODE data sets for which we analyzed ChIP-seq data of chromatin marks. We used all available samples for which a consistent measurement platform was used.

Profiling	Platform
expression	Affymetrix HT Human Genome U133A Array
H3K4me3	Illumina Genome Analyzer II
H3K27me3	Illumina Genome Analyzer II
H3K27ac	Illumina Genome Analyzer II
H3K9ac	Illumina Genome Analyzer II
H3K36me3	Illumina Genome Analyzer II
H3K4me1	Illumina Genome Analyzer II
CTCF	Illumina Genome Analyzer II
H3K4me2	Illumina Genome Analyzer II
H4K20me1	Illumina Genome Analyzer II

Table 4.5: List of the signals (i.e. expression, histone modifications, and the binding of one TF) and corresponding measurement platforms from ENCODE data sets, that we used to demonstrate ISMARA's performance on ChIP-seq data sets. We used available BED and CEL files from the GSE26386 and GSE26312 GEO series.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

in terms of the predicted TFBSs at the promoters. Supplementary table 4.6 lists all the data sets that were analyzed in this paper and shows, including references to the original publications, and lists for each data set the URL at which ISMARA's results for the corresponding data set can be found. Note that, for data sets 1, 2, and 5, there are also replicate averaged results available. These can be found by replacing 'ismara_report' at the end of the URL with 'averaged_report'.

4.4.12 Analysis of the ENCODE ChIP-seq data

Data Set	ISMARA URL http://ismara.unibas.ch/supp/
<p>GNF SymAtlas, mouse (44)</p> <p>GNF SymAtlas + NCI-60 cancer cell lines, human (54; 55)</p> <p>Inflammatory response time course, HUVEC (69)</p> <p>Mucociliary differentiation, bronchial epithelial cells, human (84)</p> <p>Epithelial-Mesenchymal Transition, human (88)</p> <p>ENCODE cell lines, expression (97)</p> <p>ENCODE cell lines, H3K4me3 (97)</p> <p>ENCODE cell lines, H3K27me3 (97)</p> <p>ENCODE cell lines, H3K27ac (97)</p> <p>ENCODE cell lines, H3K9ac (97)</p> <p>ENCODE cell lines, H3K36me3 (97)</p> <p>ENCODE cell lines, H3K4me1 (97)</p> <p>ENCODE cell lines, CTCF (97)</p> <p>ENCODE cell lines, H3K4me2 (97)</p> <p>ENCODE cell lines, H4K20me1 (97)</p>	<p>dataset1/ismara_report</p> <p>dataset2/ismara_report</p> <p>dataset3/ismara_report</p> <p>dataset4/ismara_report</p> <p>dataset5/ismara_report</p> <p>dataset6.1_ENCODE_expression/ismara_report</p> <p>dataset6.2_ENCODE_H3K4me3/ismara_report</p> <p>dataset6.3_ENCODE_H3K27me3/ismara_report</p> <p>dataset6.4_ENCODE_H3K27ac/ismara_report</p> <p>dataset6.5_ENCODE_H3K9ac/ismara_report</p> <p>dataset6.6_ENCODE_H3K36me3/ismara_report</p> <p>dataset6.7_ENCODE_H3K4me1/ismara_report</p> <p>dataset6.8_ENCODE_CTCF/ismara_report</p> <p>dataset6.9_ENCODE_H3K4me2/ismara_report</p> <p>dataset6.10_ENCODE_H4K20me1/ismara_report</p>

Table 4.6: URLs with the results of ISMARA’s analyses of the data sets discussed in this paper.

4.12.1 PCA analysis

We first performed principal component analysis of the 10 marks across all promoters genome-wide, separately for each of the 8 cell types, as described in section 4.5.10. As shown in Suppl. Fig. 4.24, we find that the first principal component explains approximately 60% of the variation in each of the 8 cell types. In addition, the first principal component is almost identical in each of the cell types. This strongly suggests that this first principal component is a general feature of the distribution of chromatin marks. Moreover, the fact that this component aligns positively with expression and activity-associated chromatin marks, suggests that this first component reflects general promoter activity. We then pooled the data from all samples and performed principal component analysis on this complete data set, i.e. treating each promoter sample combination (p, s) as if it were a separate promoter. The resulting first principal component is shown in Fig. 6B of the main article.

Next, as described in section 4.5.10, we took the inferred motif activities for all marks and removed the component along the first principal component. That is, we removed the contribution to the motif activities that comes from the general ‘promoter activity’. As an illustration, Suppl. Fig. 4.23 shows the inferred motif activities for 5 motifs (SNAI, IRF, HNF4a_NR2F1, TEAD1, and GATA6) both before (left panels) and after (right panels) the contribution from general promoter activity has been removed, for expression and the activation associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3. As the figure shows, before removal of the first PCA component, the activities for all marks are highly correlated, but this correlation disappears when the first PCA component is removed. This confirms that the highly correlated motif activities and the activation-associated chromatin marks is accounted for by the first PCA component that captures the relative chromatin mark levels associated with the general activity of a promoter. The remaining activities (right panels) thus provide a clearer insight in the specific role of a motif for specific marks across the cell-types. For example, for the SNAI motif the two acetylation marks are highly positive in HepG2 cells, whereas expression and H3K36me3 are clearly negative. Thus, promoters carrying SNAI sites tend to have higher histone acetylation levels than expected based on their general activity, and lower gene expression and H3K36me3 levels than expected based on the general activity.

As described in section 4.5.10, after removing the contribution of the first principal component to the motif activities, we re-calculated significance z -values z_m^i for each motif m and each mark i . In addition, we calculated a specificity s_m^i which measures the fraction of the overall that is associated with mark i . That is, a motif m will be highly specific for mark i if it has a high z -value z_m^i , and low z -values for all other marks. To identify motifs that are either most significant or highly specific for particular marks, we plotted scatter plots showing the significance and specificity for each motif (Suppl. Fig. 4.25). In each of the scatters we have indicated in red

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Figure 4.23: Inferred motif activities for 5 example motifs on the ENCODE ChIP-seq data sets measuring chromatin (97). Each row (labeled A through E) shows the activities for explaining expression (black), H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown) levels, for one motif. The left panels show the motif activities as inferred from the original data the right panel the motif activities after the contribution along the first principal component has been subtracted. The names of the motifs are indicated above each panel and sequence logos are shown as insets. Note that the motif activities for the different marks go from highly correlated to essentially uncorrelated as the first principal component is removed.

those motifs that had either very high significance or high specificity for the motif. Interestingly, we often find that the motifs with highest significance for a particular mark also have high specificity. For example, HNF1a is both most significant and most specific for H3K4me2 levels in promoters. Not surprisingly, the occurrence of CTCF motifs is the most significant determinant of the observed levels of bound CTCF.

4.4.12 Analysis of the ENCODE ChIP-seq data

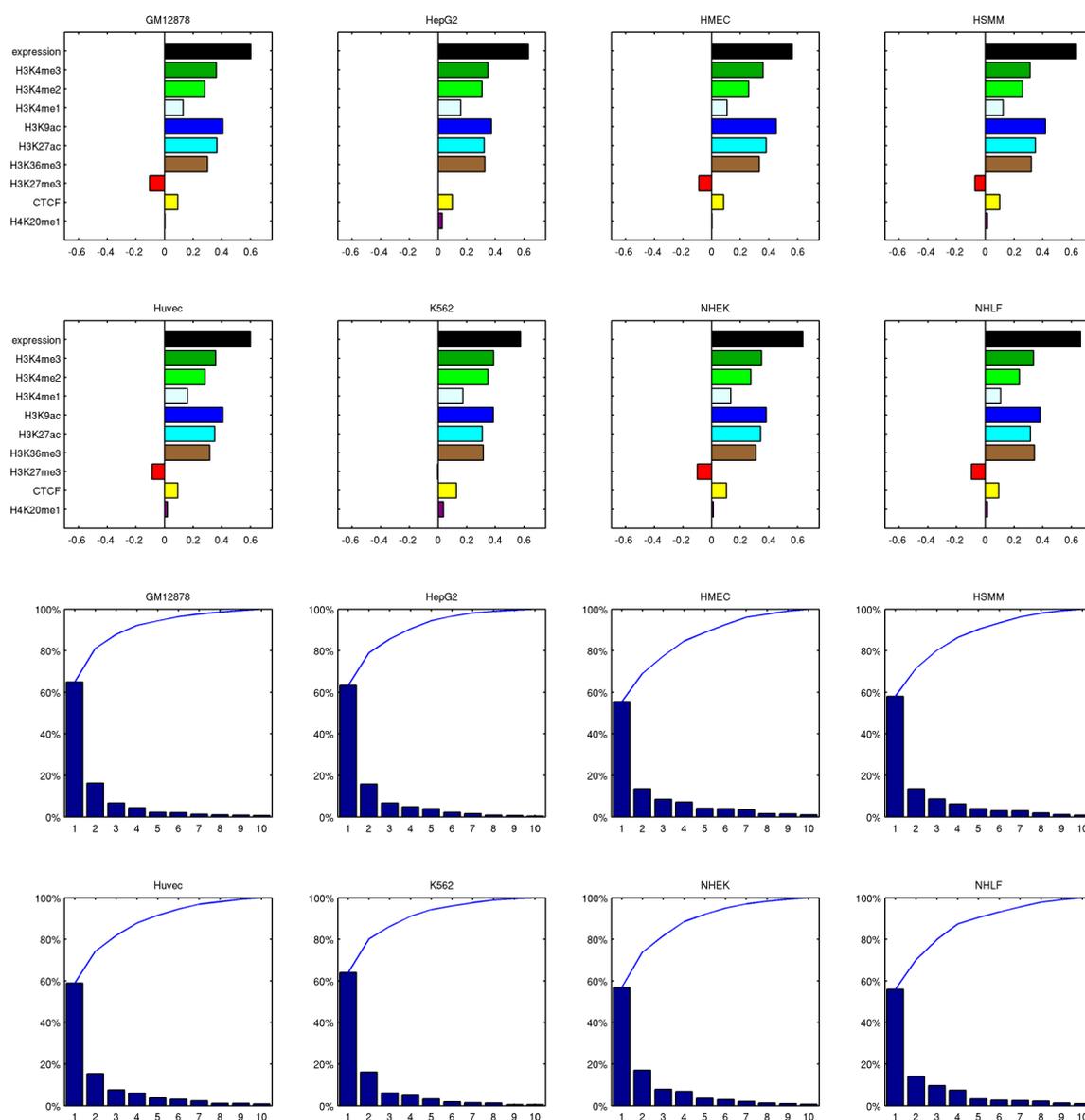


Figure 4.24: First principal component explaining the largest amount of chromatin mark and expression levels associated with each promoter, separately for each of the 8 cell types (top 8 panels). The bars indicate the relative contributions of expression and each of the chromatin marks to the first principal component. Note that the first principal component is virtually identical in each cell type. The bottom 8 panels show the fraction of the total variance explained by each subsequent principal component (bars) and the cumulative fraction of variance explained by consecutive components. Note that, for each cell type, close to 60% of the variance in expression and the 9 chromatin marks is explained by the first component.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

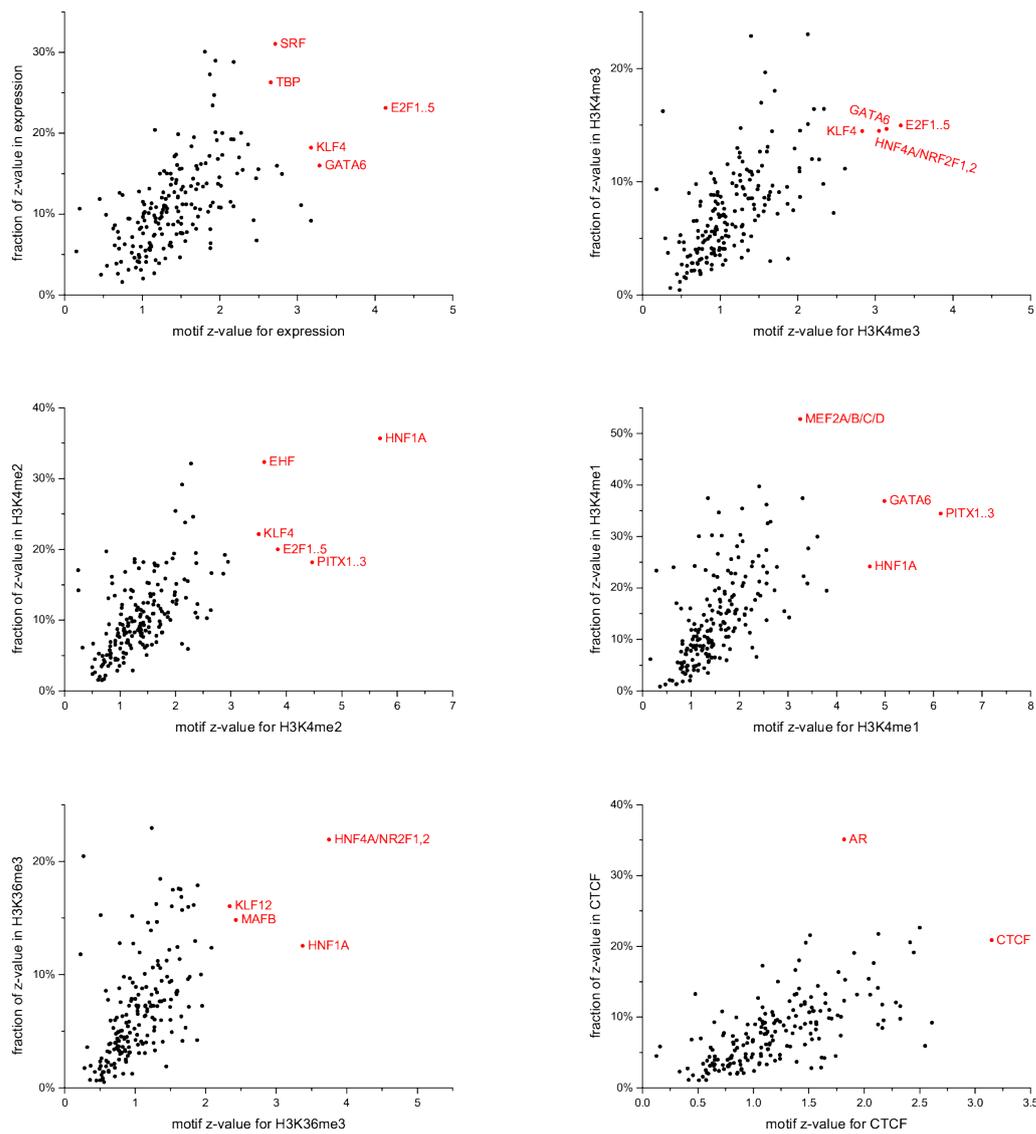


Figure 4.25: (continued on the next page)

4.4.12 Analysis of the ENCODE ChIP-seq data

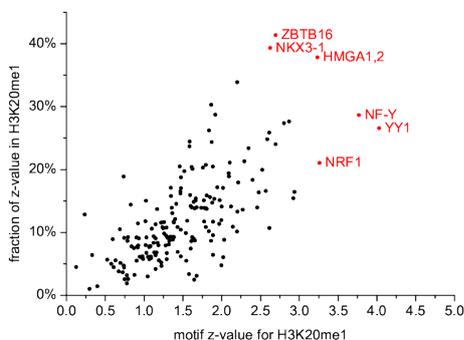


Figure 4.25: Significances and specificities of the motifs for explaining variations in different chromatin marks. Each panel corresponds to one mark (as indicated on the axes) and each dot corresponds to one motif. The significance of each motif is quantified by a z -value of the motif's activity for a given mark, after motif activities along the first principal component have been removed (see section 4.5.10). The specificity of a motif for a given mark is the fraction of all significance associated with a given mark (its z -value squared relative to the sum of all z -values squared, see section 4.5.10). the most significant and/or specific motifs for each mark are indicated in red.

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Chapter 5

The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.

Harukazu Suzuki, Alistair R R Forrest, Erik van Nimwegen, Carsten O Daub, Piotr J Balwierz, Katharine M Irvine, Timo Lassmann, Timothy Ravasi, Yuki Hasegawa
et alli

Nature Genetics, 41(5):553-62 2009, PMID:19377474
doi:10.1038/ng.375

Using deep sequencing (deepCAGE), the FANTOM4 study measured the genome-wide dynamics of transcription-start-site usage in the human monocytic cell line THP-1 throughout a time course of growth arrest and differentiation. Modeling the expression dynamics in terms of predicted cis-regulatory sites, we identified the key transcription regulators, their time-dependent activities and target genes. Systematic siRNA knock-down of 52 transcription factors confirmed the roles of individual factors in the regulatory network. Our results indicate that cellular states are constrained by complex networks involving both positive and negative regulatory interactions among substantial numbers of transcription factors and that no single transcription factor is both necessary and sufficient to drive the differentiation process.

5.1 Introduction

Development, organogenesis and homeostasis in multicellular systems involve the proliferation of precursor cells, followed by growth arrest and the acquisition of a differentiated cellular phenotype. Upon stimulation with phorbol myristate acetate (PMA), human THP-1 myelomonocytic leukemia cells cease proliferation, become adherent and differentiate into a mature monocyte- and macrophage-like phenotype(147; 148). This study aimed to understand the transcriptional network underlying growth arrest and differentiation in mammalian cells using THP-1 cells as a model system.

Most existing methods for regulatory network reconstruction collect genes into coexpressed clusters and associate these clusters with regulatory motifs or pathways (for example, see refs. (2; 149; 150)). Alternatively, one can model the expression patterns of all genes explicitly in terms of predicted regulatory sites in promoters and the post-translational activities of their cognate transcription factors (TFs)(4; 43; 151). Although this approach is challenging in complex eukaryotic genomes owing to large noncoding regions, ChIP-chip data(152) indicates that the highest density of regulatory sites is found near transcription start sites (TSSs) and regulatory regions originally thought to be distal may often be alternative promoters(12; 13). Precise identification of TSS locations is thus likely to be a crucial factor for accurate modeling of transcription regulatory dynamics in mammals.

In this study, we extend our previous observations of genome-wide TSS usage by Cap Analysis of Gene Expression (CAGE)(1) and using deep sequencing to identify promoters active during a time course of differentiation and quantify their expression dynamics. DeepCAGE data are used in combination with cDNA microarrays, other genome-scale approaches, novel computational methods and large-scale siRNA validation to provide a comprehensive analysis of growth arrest and differentiation in the THP-1 cell model.

5.2 Results

5.2.1 Outline of the analysis strategy

In most cell line models, only a subset of cells undergoes growth arrest and differentiation. To maximize the sensitivity in this study, we identified a subclone of THP-1 cells in which the large majority of cells became adherent in response to PMA. Our strategy began with deepCAGE, which identified active TSSs at single-base-pair resolution, and simultaneously measured their time-dependent expression (using normalized tag frequency) as cells differentiated in response to PMA. The same RNA was subjected to cDNA microarray analysis on an Illumina platform. The differentiation of the cells was evident from the large increase in expression of macrophage-specific genes such

as CD14 and CSF1R detected by both deepCAGE and microarray in all replicates.

Figure 5.1 summarizes our Motif Activity Response Analysis (MARA) strategy. Promoters were defined as local clusters of coexpressed TSSs and promoter regions as their immediate flanking sequences (Fig. 5.1 a,b). To reconstruct transcription regulatory dynamics we refined earlier computational methods(4; 43; 151) by incorporating comparative genomic information and each TF's positional preferences relative to the TSS in the prediction of regulatory sites. Binding sites for a comprehensive and unbiased collection of mammalian regulatory motifs were predicted in all proximal promoter regions (Fig 5.1c) and the observed promoter expression profiles (Fig. 5.1d) were combined with the predicted site-counts (Fig. 5.1e) to infer time-dependent activity profiles of regulatory motifs (Fig. 5.1f). We inferred individual regulatory interactions (edges) between motifs and promoters by comparing the promoter expression and motif activity profiles (Fig. 5.1g). Rigorous Bayesian probabilistic methods were developed for all steps of the computational analysis. Finally, a core network was constructed by selecting the motifs that explained the greatest proportion of the expression variance, obtaining all predicted regulatory edges between TFs corresponding to these motifs and selecting those regulatory edges that had independent experimental support. Using this approach, we reconstructed the transcriptional regulatory dynamics associated with cellular differentiation in human THP-1 cells, and validated a subset of predicted regulatory interactions.

The transcriptional network...

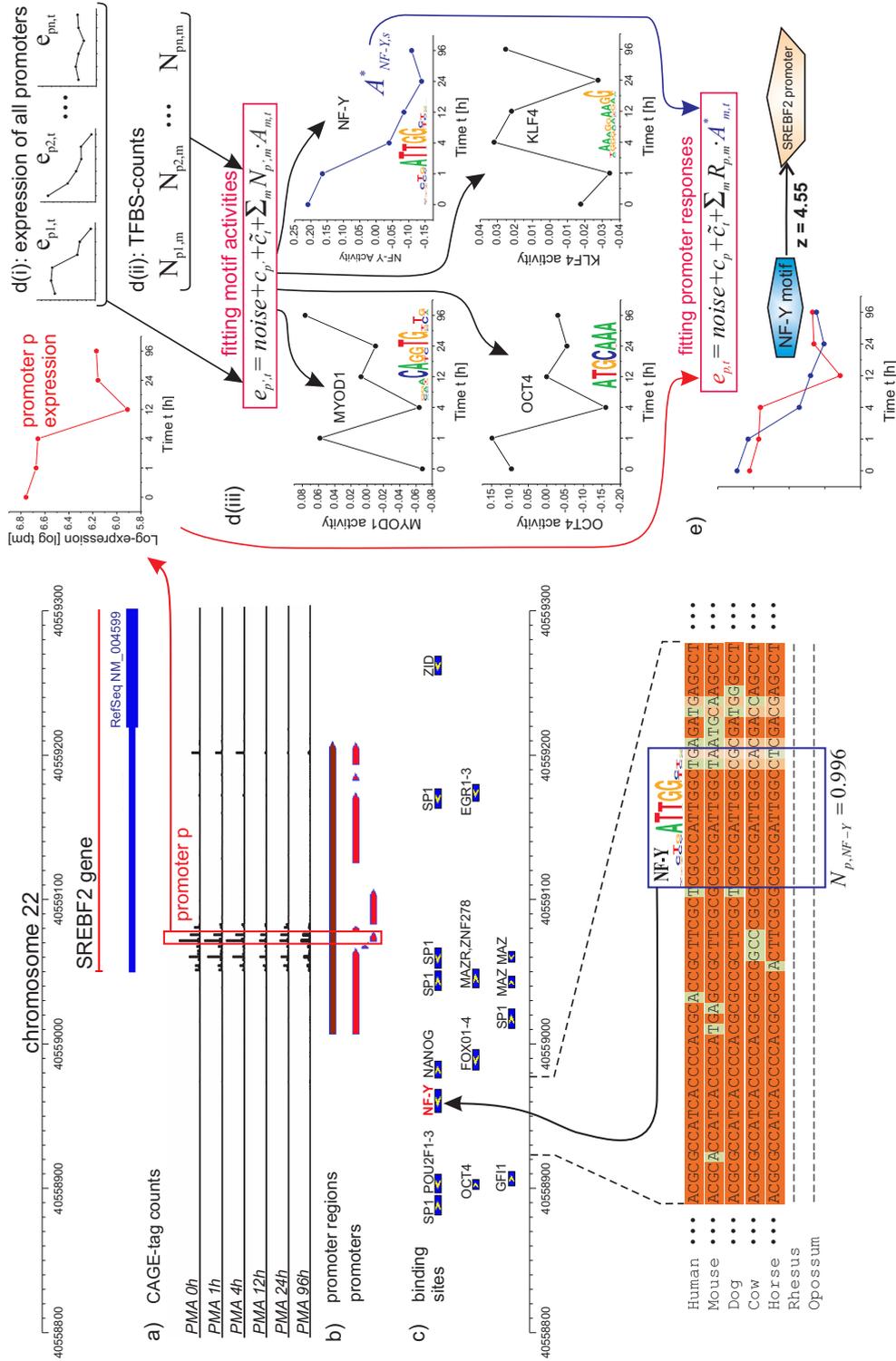


Figure 5.1: Motif Activity Response Analysis (MARA) (continued on the next page)

Figure 5.1: (a) CAGE tags are mapped to the human genome and their expression is normalized; vertical lines represent TSS positions, and their height is proportional to the normalized expression. (b) Mapped tags are clustered into promoters on the basis of their relative expression, and neighboring promoters are joined into promoter regions. (c) A window of -300 to +100 flanking each promoter region is extracted, multiply aligned and the MotEvo algorithm is used to predict binding sites for known motifs. (d–f) Observed expression of all promoters (d) and predicted site-counts (e) are used to infer motif activities (f). (g) The statistical significance of the regulatory edge from motif to promoter is calculated based on correlation of the promoter expression and motif activity profiles.

5.2.2 DeepCAGE quantification of dynamic TSS usage

CAGE tags generated from mRNA harvested at each time point were mapped to the human genome. Promoters were defined as clusters of nearby TSSs that showed identical expression profiles (within measurement noise) and were substantially expressed in at least one time point (Fig. 5.1a,b). Using these criteria we identified 29,857 promoters expressed in THP-1 cells containing 381,145 unique TSS positions (which is a subset of the nearly 2 million TSSs detected at least once in THP-1). These promoters were contained within 14,607 promoter regions (separated by at least 400 bp; Methods). The deepCAGE data was validated using genome tiling-array ChIP for markers of active transcription. Of the promoters identified, 79% and 78% were associated with H3K9Ac and RNA polymerase II, respectively (both markers of active transcription(153; 154)), compared to 18% and 27% for inactive promoters.

Among the identified promoters 84% (24,984) were within 1 kb of the starts of known transcripts and 81% (24,327) could be associated with 9,452 Entrez genes. Approximately half of the remaining promoters were more than 1 kb away from the loci of known genes (Supplementary Fig. 5.7). These newly identified promoters are conserved across mammals, suggesting that they are true transcription starts of currently unknown transcripts. The association of 24,327 promoters with 9,452 Entrez genes extends previous evidence of alternative promoter usage(12) – in this case even within a single cell type (Supplementary Table 5.1) – and demonstrates that promoter regions frequently contain multiple promoters with distinguishable expression profiles. In addition, for genes with known multiple promoters deepCAGE frequently identified only one promoter to be active in the THP-1 samples (Supplementary Fig. 5.8). Hence, deepCAGE samples a distinct aspect of transcriptional activity that can and does vary independently of mRNA abundances as measured by hybridization to representative microarray probes.

5.2.3 Promoter expression

Using the normalized tags per million (tpm) counts assigned to the promoters, we tested reproducibility among the three biological replicates and compared the outcome to the Illumina array from the same samples (Supplementary Fig. 5.9). DeepCAGE expression measurements were comparatively noisy (Supplementary Fig. 5.9a). Nevertheless, the median Pearson correlation between the replicate-averaged expression profiles of CAGE and microarray was around 0.72 (Supplementary Fig. 5.9b), which is comparable to that observed with other deep transcriptome sequencing datasets(155). As predicted, the correlation is lower for genes with multiple promoter regions (Supplementary Fig. 5.9b).

5.2.4 Comprehensive regulatory site prediction

Known binding sites from the JASPAR and TRANSFAC databases(156; 157) were used to construct a set of 201 regulatory motifs (position-specific weight matrices, WMs), which represent the DNA binding specificities of 342 human TFs. We predicted transcription factor binding sites (TFBSs) for all motifs within the proximal promoter regions (-300 to +100 bps) of all CAGE-defined promoters. Extending the proximal promoter regions beyond the -300 to +100 window decreased the quality of the fitted model described below (data not shown). In contrast to previous approaches that used simple WM scanning(151), we incorporated information from orthologous sequences in six other mammals and used a Bayesian regulatory-site prediction algorithm that uses explicit models for the evolution of regulatory sites(39; 158) (Fig. 5.1c and Methods). Notably, different motifs had distinct and highly specific positional preferences with respect to TSS (Supplementary Fig. 5.10), extending a previous genome-scale analysis²⁰. Positional preferences were incorporated in the TFBS prediction by assigning each site a probability that it is under selection and correctly positioned. This analysis generated approximately 245,000 predicted TFBSs for the 201 motifs genome-wide. For each promoter–motif combination, the TFBS prediction was summarized by a count N_{pm} , which represents the estimated total number of functional TFBSs for motif m in promoter p . The TFBS predictions were compared with published high-throughput protein–DNA interaction datasets (ChIP-chip) and predicted target genes were significantly (P values ranged from 0.02 for ETS1 to $6.60E-263$ for GABPA) enriched among genes for which binding was observed.

5.2.5 Inferring key TFs and their time-dependent activities

The details of our Motif Activity Response Analysis (MARA) are described in Methods. Briefly, for each motif m and each time point t , there is an (unknown) motif activity A_{mt} , which represents the time-dependent nuclear activity of positive and

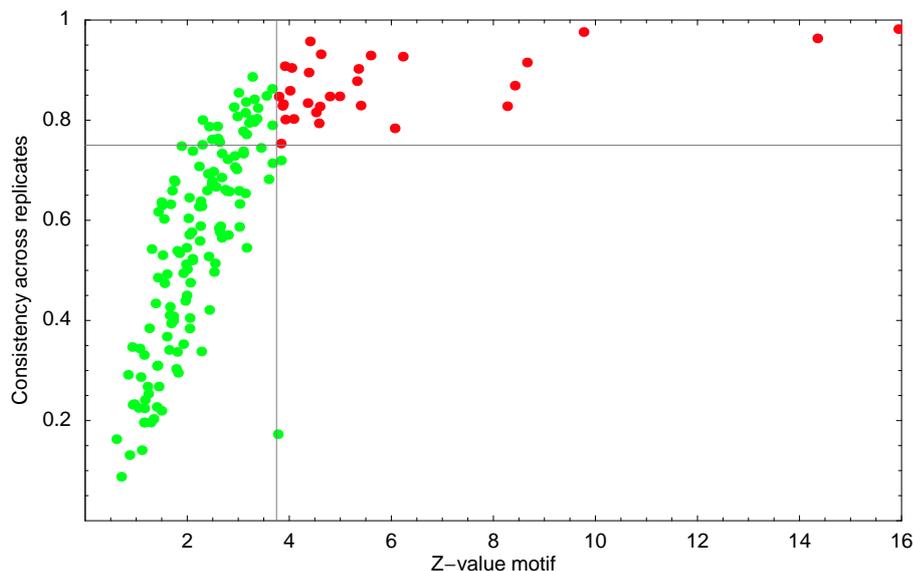


Figure 5.2: Statistical significance and consistency across replicates of the inferred motif activity profiles. Each dot corresponds to a motif. The significance of each motif in explaining the observed expression variation is quantified by the z value of its activity profile (horizontal axis, see Methods). The consistency of the inferred activity profile of each motif is quantified by the fraction of the variance (FOV) in the activity profile across all six replicates (three biological replicates for both CAGE and Illumina), which is reproduced in each replicate (vertical axis, see Methods)

negative regulatory factors that bind to the sites of the motif (for example, the E2F activity will depend on nuclear E2F1-8, and DP1-2 levels, as well as RB1 phosphorylation status). As in previous work(4; 5; 43; 151), motif activities were inferred by assuming that the expression e_{pt} of promoter p at time t is a linear function of the activities A_{mt} of those motifs that have predicted sites in p . Additionally, the effect of motif m on the expression of promoter p is assumed to be proportional to the predicted number of functional sites N_{pm} . Assuming that the deviations of the predicted expression levels $e_{pt}^{theo} = constant + \sum_m N_{pm}A_{mt}$ from the observed levels e_{pt} are Gaussian distributed, and using a Gaussian prior on the activities, we determine fitted activities A_{mt}^* that have maximal posterior probability (Methods).

The inferred motif activities were validated using a number of internal tests. First, our Bayesian procedure quantifies both the significance of each motif in explaining the observed expression variation as well as the reproducibility of its activity across replicates (Fig. 5.2 and Supplementary Table 4 online). The activity profiles of the top motifs are extremely reproducible across replicates and different measurement

The transcriptional network...

technologies (Figs. 5.25.3a and Supplementary Fig. 5.11). It should be stressed that, although motif activities are inferred by fitting the expression profiles of all promoters, the model cannot be expected to predict expression profiles of individual genes from the predicted TFBS in proximal promoters alone. The effects of chromatin structure, distal regulatory sites, nonlinear interactions between regulatory sites, and the contribution of the large numbers of human TFs for which no motif is known, are not considered. Furthermore, especially for genes that are dynamically regulated, mature mRNA abundance can be dynamically regulated independently of transcription initiation and promoter activity through selective mRNA elongation, processing and degradation. Our aim is not to predict expression profiles of individual genes but rather to predict the key regulators and their time-dependent activities, which can be inferred from integration of global expression information in a system undergoing dynamic change. We validated the significance of the inferred activity profiles by comparing the fraction of the 'expression signal' (expression variance minus replicate noise) that is explained by the model, compared to randomized versions, and under a tenfold cross-validation test (Supplementary Fig. 5.12). The explained expression signal is highly significant and this significance is maintained under tenfold cross-validation (Methods). In addition, the highly peaked positional profiles of TFBSs (Supplementary Fig. 5.10) suggest that knowing the exact TSS is important for accurate TFBS prediction. Indeed, the predicted TFBSs from CAGE promoters explain substantially more of the expression signal in microarrays than predicted TFBSs of the associated RefSeq promoters (Supplementary Fig. 5.12). We observe that the model better predicts the expression profiles of those promoters that are more strongly expressed, more reproducible across replicates, and have higher expression variance (Supplementary Fig. 5.13). Similarly, samples at the start and end of the differentiation time course are better predicted than those at intermediate time points (Supplementary Fig. 5.14), possibly because individual cells differentiate at different rates and leave the cell populations less homogeneous at intermediate time points.

Motif activities that were independently inferred from all 11,995 expressed microarray probes were combined with the inferred motif activities from all CAGE and microarray replicates into a final set of time-dependent motif activities (Methods). From these, we selected 30 'core' motifs that contribute most to explaining the expression variation (red dots in 5.2) and segregated their activity profiles using a Bayesian procedure into nine clusters (Fig. 5.3b and Methods), including three clusters of up-regulated motifs, three clusters of downregulated motifs and three clusters containing single motifs with profiles involving different transient dynamics. The genome-wide set of target promoters for each of the motifs was determined as described in Methods. The significance of each regulatory 'edge' from a motif to a putative target promoter (containing a predicted TFBS) was quantified by the z value of the correlation between the motif's activity profile and the promoter's expression profile (Fig. 5.1e).

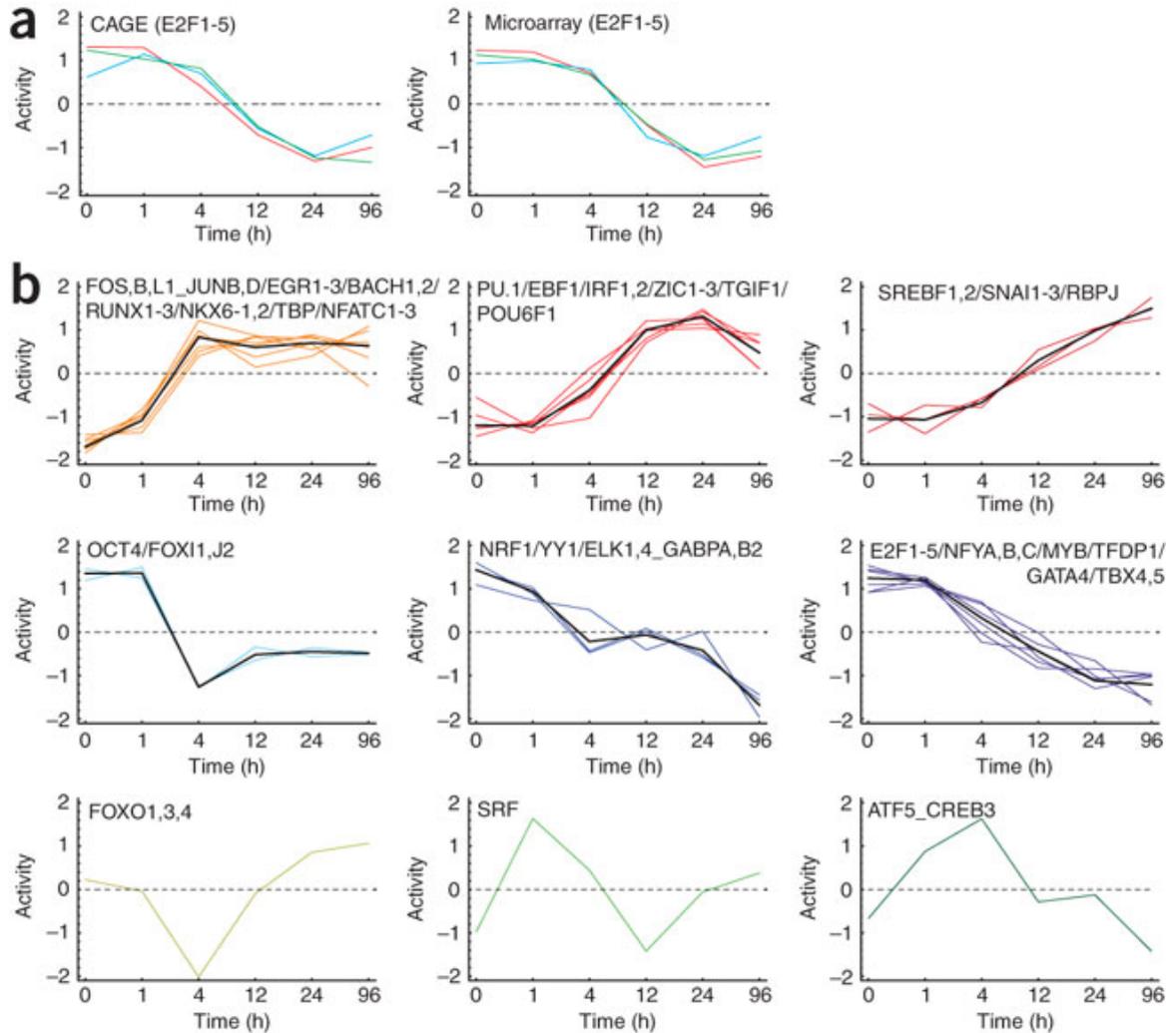


Figure 5.3: Inferred time-dependent activities of the key regulatory motifs. (a) The time-dependent activity profile of the E2F1-5 regulatory motif as inferred from CAGE (left) and microarray (right) data. The three biological replicates are shown in red, blue and green. (b) The 30 most significant motifs with consistent activity profiles across all replicates (CAGE and microarray) were clustered into nine sets of motifs with similar dynamics. Each panel shows the activity of the members of the cluster (colored curves), the names of motifs contributing and the cluster average activity profile (black).

5.2.6 Core transcriptional regulatory network

The final aim in reconstructing transcriptional regulatory networks is to infer not only the key regulators and their target gene sets, but also the way in which the actions of these key regulators are coordinated. For this purpose, we collected all 199 predicted regulatory edges (z value greater than or equal to 1.5) between the 30 core motifs. Recognizing that the prediction of individual regulatory edges is still prone to error, we constructed a core regulatory network (Fig. 5.4) of 55 highly trusted edges by filtering the predicted edges according to experimental validation, either within our data or in existing literature. In addition, for each core motif we extracted the set of predicted target genes (z value ≥ 1.5) and checked for enrichment of gene ontology terms. A selection of significantly enriched terms is shown as oval nodes in Figure 5.4.

Whereas our method infers the key regulators *ab initio*, the majority of factors within this core network are known to be important in the monocyte-macrophage lineage, thereby validating the method. In addition the predicted targets of these motifs are enriched for biological processes known to be involved in differentiation of the monocytic lineage.

The gene ontology enrichments can broadly be divided into four groups. Down-regulated motifs E2F1-5, NFYA,B,C and MYB are associated with cell cycle-related terms, consistent with the growth arrest observed during PMA-induced differentiation and the specific downregulation of numerous genes required for DNA synthesis and cell cycle progression within 24 h of PMA addition. Notably, MYB targets are also enriched specifically for microtubule-cytoskeleton-associated genes. Conversely, targets of upregulated motifs are associated with the terms immune response, cell adhesion, plasma membrane, vacuole and lysosome, all of which are consistent with differentiation into an adherent monocyte-like cell. The targeting of lysosomal genes by cholesterol-regulated SREBFs (sterol regulatory element-binding transcription factors) is of note, as lipid homeostasis is important in the macrophage in atherosclerosis and lysosomal storage diseases(159). We also saw enrichment of signal transduction genes among targets of the early induced motifs EGR1-3 and TBP. Finally, there is a set of motifs whose targets are enriched in TFs. These motifs correspond to the transiently induced/repressed motifs, ATF5_CREB3, FOXO1,3,4 and SRF, and the repressed pair of OCT4 and FOXI1,J2 motifs.

The transcriptional network...

Figure 5.4: (*continued from the previous page*) An edge $X \rightarrow Y$ is drawn whenever the promoter of at least one of the TFs associated with motif Y has a predicted regulatory edge for motif X (z value greater ≥ 1.5) and the edge has independent experimental support. The color of each node reflects its cluster membership and the size of the node reflects the significance of the motif. Edges confirmed in the literature, by ChIP or by siRNA are shown in red, blue and green, respectively. In cases where there are multiple lines of support only one evidence type is shown. GO terms significantly enriched among target genes are shown as white nodes with black edges. FOS/JUN (FOS,B,L1_JUNB,D), CREB (ATF5_CREB3), GABPA (ELK1,4_GABPA,B2).

5.2.7 Validation of edge predictions

THP-1 cells, even in an 'undifferentiated' state, are clearly a myeloid cell line. In seeking to validate the transcriptional network, we noted that there was a large set of TF genes expressed constitutively in the cells that were rapidly downregulated in response to PMA, of which MYB is an example, and another set that was expressed but further upregulated during differentiation. It is technically difficult to apply siRNA knockdown to genes that are only expressed later in the differentiation. To validate predicted edges empirically, we therefore chose to carry out siRNA knockdowns in undifferentiated THP-1 cells for genes encoding 28 TFs that are expressed in the undifferentiated state and for which we have associated motifs. To assess whether siRNA knockdown carried out in the undifferentiated state is appropriate to address factors that increase expression during the time course, we carried out the technically more difficult experiment of siRNA knockdown combined with PMA treatment for SPI1 (more commonly known in the literature as PU.1). All knockdowns were carried out in biological triplicate and qRT-PCR was used to confirm RNA-level knockdown, which in most cases was greater than 80%. Changes in gene expression caused by TF knockdown were measured by Illumina microarrays. For each knocked-down TF gene, we obtained the list of predicted regulatory targets for the associated motif and divided the microarray probes into predicted targets and nontargets for a range of z -value thresholds. Higher-confidence targets in general show greater expression changes upon knockdown (Fig. 5.5a shows the example TF genes MYB, SNAI3, EGR1 and RUNX1; additional examples are shown in Supplementary Fig. 14 online). For SPI1, even in the absence of PMA treatment siRNA knockdown caused significant downregulation of predicted SPI1 targets, but the effects were much stronger when knockdown was combined with 1 h or 24 h of PMA treatment (Fig. 5.5b), confirming that PMA causes upregulation of SPI1 activity. A good correlation between target confidence (z -value cut-off) and average log expression ratio was observed for the large majority of experiments (Fig. 5.5c). For an intermediate cut-off of $z = 1.5$ we quantified the difference in log expression ratio of predicted targets and nontargets (Fig.

5.5d) and found significant changes (z-value larger than 2) for 23 of 33 cases with SPI1 knockdown combined with 24 h of PMA treatment and MYB knockdown being the most significant. Notably, for the TF genes LMO2, MXI1 and SP1, the knockdown led to a significant upregulation of their targets, suggesting that the three encoded TFs act primarily as repressors in undifferentiated THP-1 cells (Fig. 5.5d, also see Supplementary Fig. 5.15a). Together these results provide compelling experimental validation of our predicted regulatory edges.

5.2.8 Single TF knockdowns affect multiple motif activities

Besides validating predicted targets, the siRNA knockdowns can also be used to assess the effects of the knockdown of one TF gene on the motif activities of other TFs. In addition to the 28 TFs perturbed above, we included a further 24 TFs that lacked motifs but were naturally repressed during PMA differentiation, or had been reported to have a role in myeloid differentiation or leukemia.

The motif activity inference method was used to determine the changes in activities of all motifs upon knockdown of each TF gene. To assess the role of each TF in differentiation, we defined the differentiative overlap between a TF gene knockdown and the PMA time course as the fraction of all motifs that significantly changed their activity in the same direction upon TF gene knockdown as in the PMA differentiation (Methods). By far the largest differentiative overlap (69%) was observed for the MYB knockdown, which not only affected MYB motif activity, but also the activity of most motifs in the core network, with the most significant activity changes all in the same direction as in the PMA time course (Fig. 5.6a). Knockdown of 13 other TF genes generated an overlap greater than the negative control (Supplementary Table 9 online), and Figure 5.6 shows three further examples (E2F1, HOXA9 and CEBPG).

As for MYB, E2F1 knockdown reproduced some of the downregulation of MYB and E2F activity observed upon PMA stimulation, but it failed to reproduce the upregulation of SREBF1,2, PU.1, NFATC1-3 and FOS,B,L1_JUNB,D activity (Fig. 5.6b). Similarly, the activity changes that HOXA9 knockdown induced were mostly in the same direction as in the PMA differentiation; however, the SNAI1-3 and IRF1,2 motif activities failed to be induced and the GATA4 and TBX4,5 motif activities failed to be downregulated (Fig. 5.6c). Notably, knockdown of CEBPG, encoding one of the PMA-downregulated factors, for which we do not have a motif, also generated activity changes that significantly overlapped those observed in response to PMA (Fig. 5.6d). Finally, instead of comparing the motif activity changes that different knockdowns induced, we can also directly compare the expression changes of all genes with the expression changes observed in the PMA time course. We found that MYB, HOXA9, CEBPG, GF11, CEBPA, FLI1 and MLLT3 knockdowns all generated changes in gene expression that reiterated some of those observed with PMA treatment. MYB knockdown was exceptional, as it induced 35% (340/967) and repressed

The transcriptional network...

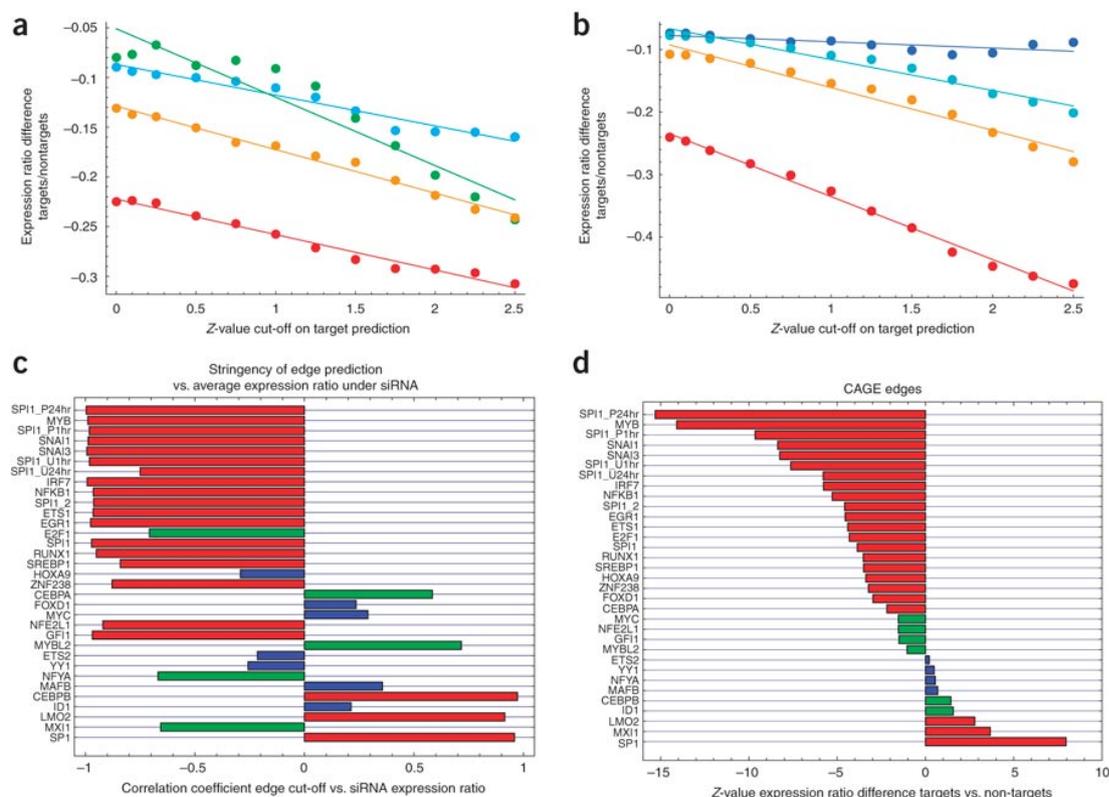


Figure 5.5: Validation of predicted target promoter sets using siRNA knockdowns. (a) Difference in the average log expression ratio upon knockdown between predicted target promoters and predicted nontargets (vertical axis) as a function of the z-value cut-off on target prediction (horizontal axis, more stringent cut-offs are on the right) for knockdown of the TF genes MYB (red), SNAI3 (orange), RUNX1 (green) and EGR1 (light blue). (b) As in (a) but now for knockdown of SPI1 followed by 1 h without treatment (light blue), 24 h without treatment (dark blue), 1 h of PMA treatment (orange) and 24 h of PMA treatment (red). All straight lines are linear regression fits. (c) Pearson correlation coefficients between the average log expression ratio difference of targets and nontargets and the cut-off on target predictions (horizontal axis). Red bars indicate correlation coefficients larger than 0.75 in absolute value; green bars, absolute values between 0.5 and 0.75; and blue bars, less than 0.5. (d) Significance (z value) of the difference in log expression ratio between predicted targets and nontargets (cut-off $z = 1.5$) for all 28 TFs associated with a motif, measured as a z value (number of standard errors). Red bars correspond to significant changes, that is, greater than two standard errors; green bars, changes between 1 and 2 standard errors; and blue bars, changes less than 1 standard error. siRNA knockdowns were carried out in biological triplicate and knockdown was assessed by qRT-PCR.

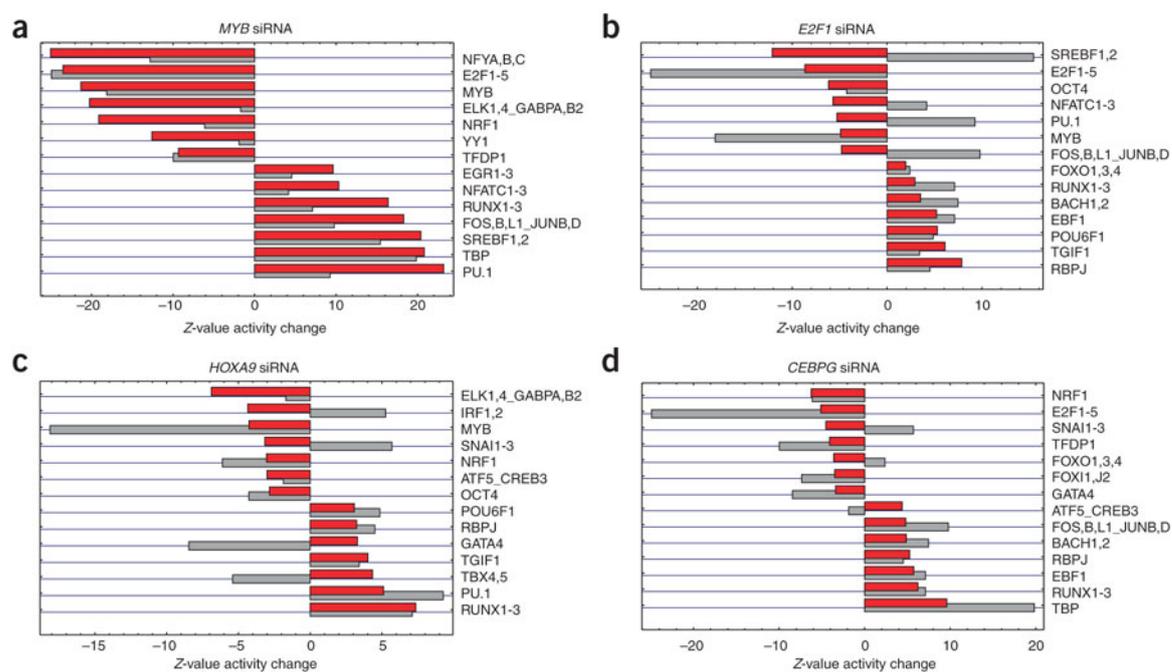


Figure 5.6: Most significant motif activity changes (as measured by z value, red bars) for four TF gene knockdowns that induce motif activity changes that have a differentiative overlap with the PMA time course of more than 50%. The corresponding motif activity changes observed in the PMA time course are shown as gray bars.

The transcriptional network...

19% (172/916) of the genes upregulated and downregulated with PMA, respectively. In addition the cells became adherent and began to express the monocytic markers CD11B (ITGAM), CD54 (ICAM1), CD14, APOE and CSF1R, three of which we confirmed by flow cytometry. This development of adherence could be linked to the GO enrichment for cytoskeleton-associated genes among MYB targets noted above. Given these observations one might wonder whether MYB is a master regulator of the differentiation process and whether stronger and longer knockdown would have reproduced the complete differentiation observed under PMA treatment. Several observations argue strongly against this. First, the gene sets perturbed by MYB and by the other pro-differentiative TFs overlap only partially. Second, of the six other pro-differentiative TF genes only two (CEBPG and GFI1) are affected by MYB knockdown. Both these facts indicate that the other pro-differentiative TF genes are not simply downstream of MYB. Third, MYB downregulation does not occur until after the second hour of the PMA time course (Fig. 5.3b), which is at odds with the idea of MYB sitting at the top of the regulatory hierarchy. It is also worth noting that THP-1 cells harbor a leukemogenic fusion(160) between MLL (mixed-lineage leukemia) and MLLT3 (MLL translocation partner 3) and that the MLLT3 siRNA targets this leukemogenic fusion (note that full-length MLLT3 does not seem to be expressed in THP-1 as there is no CAGE 5' signal for this gene). Our data indicate that this fusion interferes with differentiation and that neither PMA treatment nor MYB knockdown affects MLL-MLLT3 levels, suggesting these stimuli can bypass the differentiative block. Conversely, MLLT3 knockdown had no effect on MYB levels. These results agree with previous RNAi studies that conclude that downregulation of MLL leukemogenic fusion proteins can promote growth arrest but is not required for terminal differentiation(161; 162). Thus, individual TF gene knockdowns affect the activities of multiple motifs and elicit different, but overlapping, subsets of the regulatory changes observed in the PMA time course. Taken together, the data indicate that the independent perturbation of expression of multiple TFs in response to PMA is both necessary and sufficient to initiate partial differentiation.

5.2.9 Many TFs are involved in the differentiation process

The network predictions and the siRNA results above suggest that upregulation and downregulation of the activities of multiple cooperating TFs is required for differentiation. Of a curated list(163) of 1,322 human TFs, 610 were detected by both CAGE and microarray in at least one time point; however, only 155 of these are covered by weight matrices, suggesting that other factors may well be important in these cells. Of the 610 expressed TFs 64 were most highly expressed in the undifferentiated and 34 in the differentiated state. In addition, 101 TFs were transiently induced or repressed during differentiation. To elucidate the connection of these TFs to the inferred network, we compared the predicted regulatory inputs of co-regulated subsets

of TFs with the predicted regulatory inputs of the set of all 610 expressed TFs.

Whereas no motifs are overrepresented among inputs of statically expressed TFs, inputs of dynamically expressed TFs showed enrichment for a subset of motifs. TFs downregulated from 0 to 96 h PMA were most enriched for three downregulated motifs of the core network: OCT4 (3.4 times), GATA4 (3.3 times) and NFYA,B,C (2.2 times). Similarly, TFs upregulated from 0 to 96 h were most enriched for core network motifs that increase activity during differentiation: SNAI1-3 (4.6 times) and TBP (5.2 times). Finally, transiently regulated TFs were enriched for the SRF (3.5 times) and NHLH1,2 (3 times) motifs.

Notably, TFs that are predicted targets of SRF are mostly induced in the first hour of PMA-induced differentiation. During this first hour 55 of the 57 genes whose expression was perturbed are induced and 30% encode TFs (Supplementary Fig. 5.16a). The regulatory inputs of these early-induced TFs are enriched for the motifs SRF, TBP and FOSL2, which all correspond to known PMA-responsive TFs (164; 165; 166; 167). Among the early-induced TFs, five correspond to upregulated core network motifs themselves (FOSB, EGR1-3 and SNAI1) and two (MAFB and EGR1) are known to induce pro-differentiative changes(168; 169). It is also worth noting that significant downregulation did not occur until the second hour, and this may require both early induction of transcriptional repressors and the RNA degradation proteins BTG2 and ZFP36 (tristetraprolin)(170; 171) (Supplementary Fig. 5.16b). Together, these results suggest that induction of SRF target genes in the first hour is critical to establishing the differentiative program and is required before factors maintaining the undifferentiated state are downregulated (Supplementary Fig. 5.16b,c).

5.2.10 Web interface to data and analysis results

To facilitate the use of the data and analysis of results amassed here, we provide an online tool, EdgeExpressDB, as part of the FANTOM4 web resource, which allows users to explore our annotations of the structure, expression and regulation of promoters genome-wide. It also integrates published TF-promoter interactions, the siRNA perturbations and genome-wide chromatin immunoprecipitation experiments. Our complete set of regulatory-interaction predictions provides a large collection of hypotheses that can be targeted for validation, for example, through chromatin immunoprecipitation, gel shift assays or reporter assays. The value of this resource is illustrated by detailed examination of individual loci. For example, the osteopontin gene (SPP1) is massively induced from 12 h of differentiation. Our predictions confirm RUNX and PU.1 as regulators and support a previous analysis in mouse implicating the TGIF1 factor. In addition our analysis identifies NFAT, STAT, NKX6.2 and LIM domain and homeobox proteins as candidates for further testing.

Finally, our set of human promoters, TF motifs, genome-wide annotation of TF-

binding sites and their predicted effects on the expression of the target promoters are available through the SwissRegulon website. A web interface, allowing researchers to automatically perform Motif Activity Response Analysis (MARA) of their own expression data in terms of our genome-wide predictions of TFBSs, is also available at SwissRegulon.

5.3 Discussion

We have devised a new integrated approach that combines genome-wide identification of TSSs and their time-dependent expression with computational modeling to reconstruct the transcriptional regulatory dynamics of a differentiating human cell line. The CAGE tag sequencing used here is tenfold deeper than in previous studies(12), and this is the first study to our knowledge to quantitatively monitor dynamic expression changes of individual TSSs genome-wide. Using this data we developed a new computational method in which promoter expression profiles were modeled directly in terms of the TFBSs occurring in their proximal promoter regions. This method allowed us to infer which regulatory motifs are most predictive of expression changes and the time-dependent activities of the corresponding TFs *ab initio*. We identified more than two dozen different regulatory motifs that significantly change their activity during PMA-induced differentiation and a complex network of regulatory interactions between them that have independent experimental support. Notably, although the modeling considers only TFBSs in proximal promoter sequences, the core network in Figure 5.4 contains most of the known regulators of macrophage differentiation. Furthermore, siRNA perturbation of these TFs confirmed many of their predicted targets, and by analyzing changes in motif activity we found that each knockdown led to a distinct transcriptional state that was associated with changes in the activities of multiple motifs.

The changes in motif activity that we observed during THP-1 macrophage differentiation do not necessarily imply that the factor(s) that act upon a motif are themselves transcriptionally regulated. For example, PU.1 (SPI1) activity increases significantly in response to PMA and we have confirmed that, besides a moderate increase in mRNA expression, the SPI1 protein is also activated by phosphorylation(172) and nuclear translocation(173) (data not shown). For other motifs such as E2F, multiple redundant factors can bind to the same sites(174). Motif activity analysis is conducted without any assumptions about the TFs that act through these regulatory elements. That is, because motif activity is inferred directly from expression changes of predicted targets, the most active motifs can be identified before ascertaining the responsible TF(s) and their mode of regulation. Thus, motif activity analysis is a powerful approach compared to analysis of TF mRNA expression alone.

What do our results teach us about the general structure of regulatory networks

in cellular differentiation? An often evoked picture is that differentiation pathways consist of well-defined cascades of regulatory events which are initiated by master regulators that sit at the top of fixed regulatory hierarchies. A prime candidate for such a master regulator in our system would be MYB, as its siRNA-mediated knock-down reconstituted a significant fraction of the expression and phenotypic changes observed under PMA-induced differentiation. Indeed, this observation is consistent with earlier reports that MYB antisense treatment of myeloid leukemia lines causes differentiative growth arrest(175) and that MYB is a repressor of expression of mature macrophage-expressed genes such as CSF1R(176). Our data indicate that MYB probably acts on such genes indirectly, by a transcriptional program that represses upregulation of SPI1 activity and downregulation of proliferation.

However, several observations argue against MYB as a master regulator: MYB downregulation is not among the first events in the PMA time course, MYB knock-down far from completely mimics the PMA-induced differentiation and there are several other TFs, which are not downstream of MYB, whose knockdown reconstituted different subsets of the PMA-induced expression changes. Moreover, it is known that additional factors can also drive differentiation, for example, enforced expression of SPI1 and CEBPA in mouse fibroblasts is sufficient to drive acquisition of a macrophage-like phenotype(177), and overexpression of EGR1 and MAFB also drives differentiation, as we noted above. Yet, evidence from mouse knockouts indicates that the whole EGR family is dispensable for macrophage proliferation, differentiation and function(178).

Rather than a fixed hierarchy with one or very few master regulators at the top, the picture that emerges is that of a recurrent network in which multiple TFs mutually coordinate their activity changes to implement the differentiation. In addition, whereas different partial differentiation pathways can be initiated by multiple independent perturbations, it appears that complete differentiation requires the coordinated downregulation of multiple factors that maintain the undifferentiated state. This observation draws some similarities to the TF network that both maintains proliferation and prevents differentiation in embryonic stem cells(179). Enforced expression of four stem cell transcription factors (MYC, OCT4, KLF4, SOX2) is sufficient to dedifferentiate committed adult cells into a stem cell-like state(112). Maintenance of an undifferentiated proliferative state is important in cancer, and it is worth noting that 10 of the 64 downregulated TFs (16%) have Entrez gene annotations containing the term 'myeloid leukemia' (compared to 50 of the remaining 1,258 TFs (4%)). In addition we have demonstrated that knockdown of the MLL-MLLT3 leukemogenic fusion found in THP-1 also partially promotes differentiation.

From our time-course analysis, we see distinct phases of early, middle and late, induction and repression. Our modeling predicts, and the literature supports, SRF as the major effector of transcriptional activation of immediate early genes (IEG)(180). However, SRF activation and IEG responses are not restricted to the PMA stimulus,

The transcriptional network...

the monocytic lineage or differentiation(165; 181; 182; 183), suggesting that this response has a more general function. We speculate that a generalized immediate early response may be used to put the cell into a transient receptive state, which permits downregulation of the multiple TFs that maintain the undifferentiated state. This fits with the concept of stable cellular states as attractors of the regulatory network dynamics. The associated attractor basins(184; 185) of cellular states are analogous to local minima in energy landscapes surrounded by slopes, and homeostatic interactions between the TFs can be considered as providing a kind of inertia to maintain this state. We suggest that the immediate early response may help overcome this inertia, that is, by moving the system out of its attractor basin.

5.4 Supplementary Figures

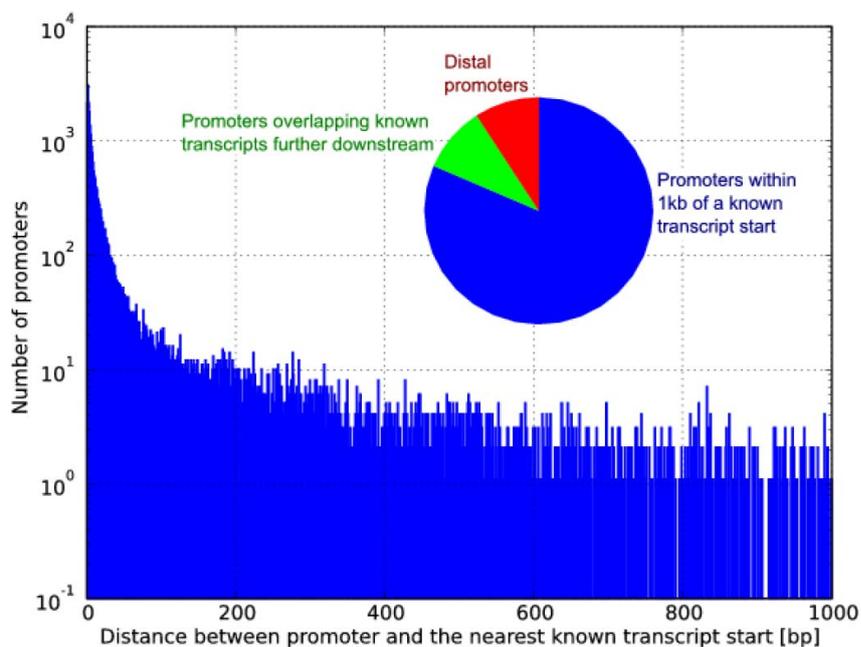


Figure 5.7: Distribution of distances between promoters and the start of the nearest known transcript. Note the vertical axis is shown on a logarithmic scale. The inset shows the fractions of promoters within 1Kb of a known start, those further than 1Kb from a known start but within 1Kb of a gene locus, and those distal to gene loci.

The transcriptional network...

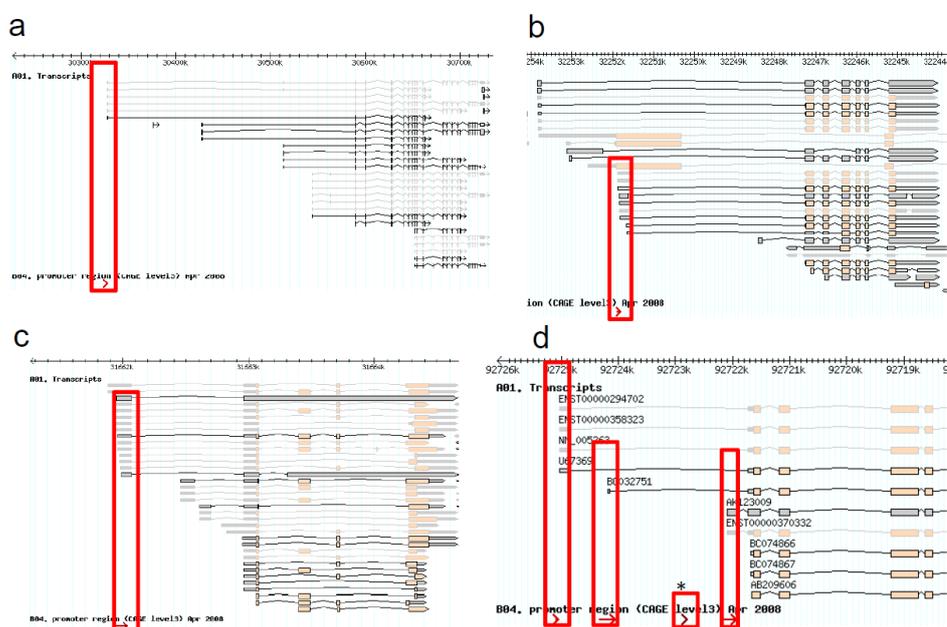


Figure 5.8: Active promoters in THP-1 differentiation. Red boxes show the promoter regions detected by deepCAGE for the example genes (a) DTNA, (b) AGPAT1, (c) LST1 and (d) GFI1. Note, the third promoter in GFI1 does not map to a full length transcript however there is EST support (BM149905).

5.5.4 Supplementary Figures

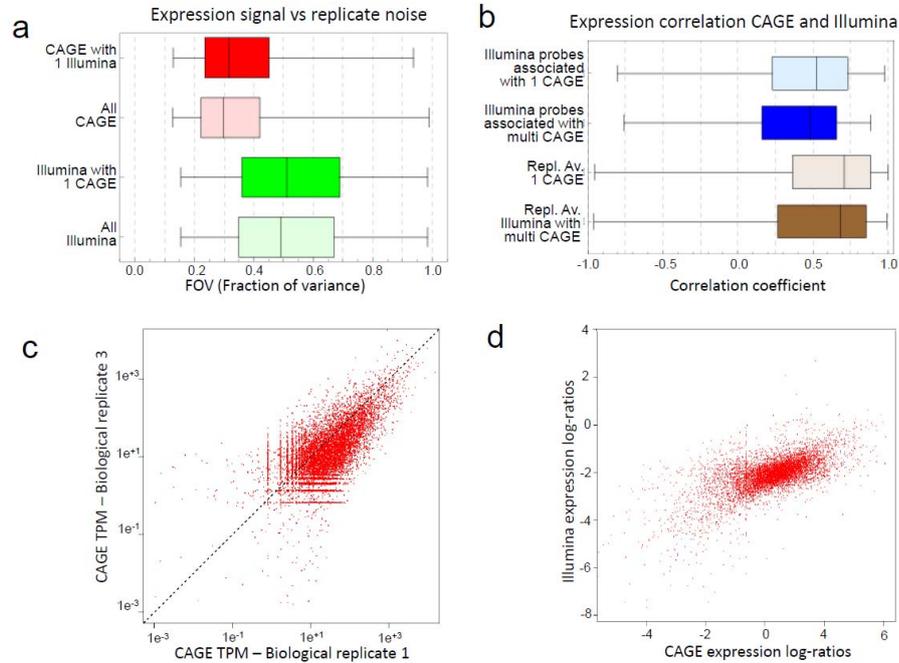


Figure 5.9: Reproducibility of the expression profiles across the three biological replicate time series, and correlation between the expression profiles based on CAGE and microarray measurements. (a) Distributions of the “expression signal” of the promoters/probes defined as the fraction of expression variance (FOV) that is reproduced across the three replicates. The whiskers denote 5 and 95 percentiles, the bar the 25 and 75 percentiles and the vertical line denotes the median fraction of variance for CAGE promoters that are associated with 1 microarray probe (red), all CAGE promoters (light red), microarray probes associated with 1 CAGE promoter (green) and all microarray probes (light green). (b) Distribution of Pearson correlation coefficients of the expression profiles of microarray probes and associated CAGE promoters. Whiskers denote 5 and 95 percentiles, boxes 25 and 75 percentiles and the vertical line the median correlation coefficient for probes associated with 1 CAGE promoter (light blue), probes associated with multiple CAGE promoters (blue), correlations of the replicate-averages for microarray probes associated with 1 CAGE promoter (light brown) and probes associated with multiple CAGE promoters (brown). (c) Representative scatterplot of deepCAGE biological replicates for undifferentiated THP-1 cells. (d) Representative scatterplot of median normalized log expression ratios for Illumina and CAGE for undifferentiated THP-1 cells.

The transcriptional network...

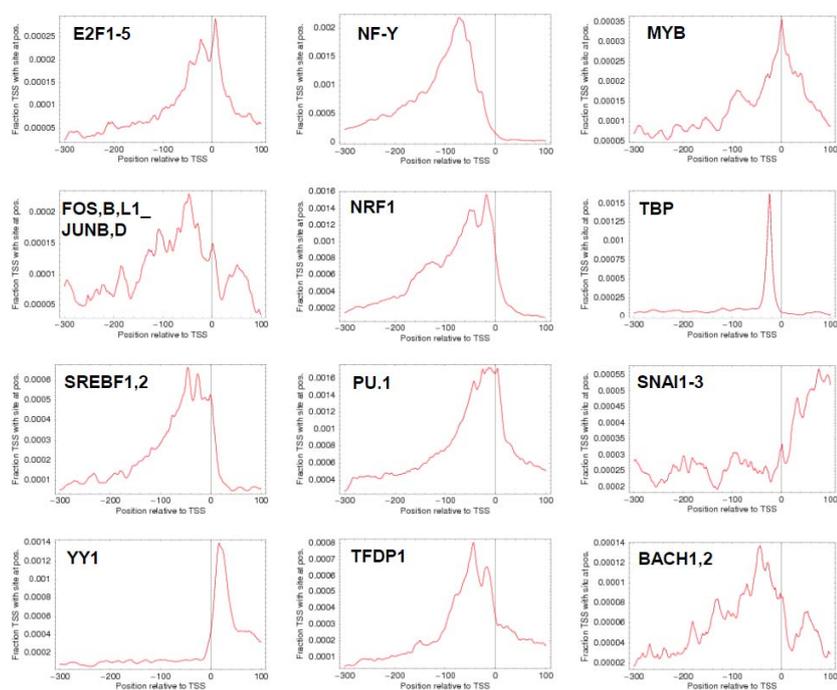


Figure 5.10: Positional distribution relative to TSS of predicted TFBSs for the 15 most significant motifs. The horizontal axis shows the position relative to TSS and the vertical axis shows the fraction of all promoters that have a site for the motif centered precisely at the corresponding position.

5.5.4 Supplementary Figures

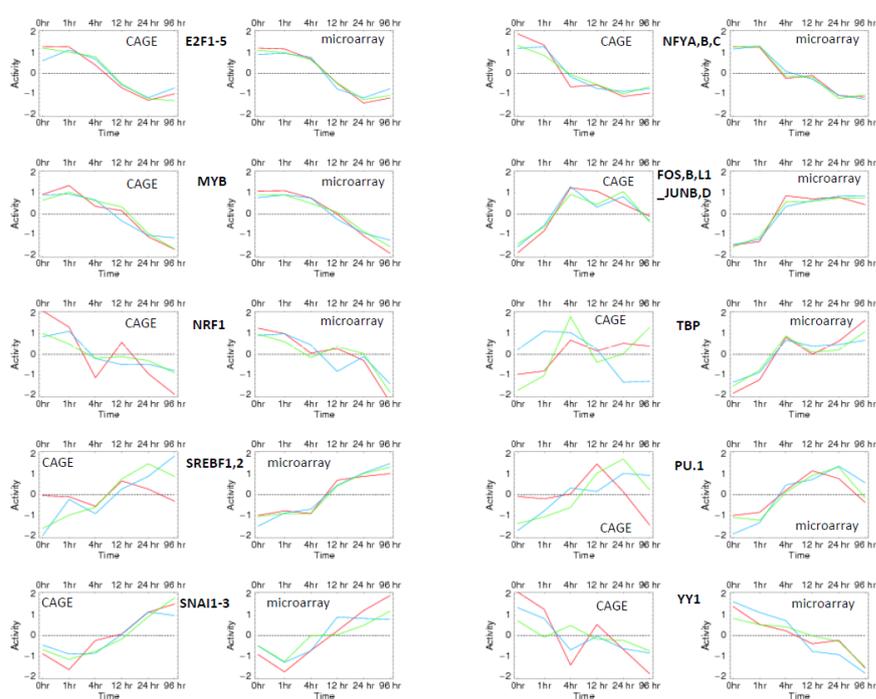


Figure 5.11: Inferred motif activities across replicates (CAGE and microarray) for the top 10 most significant motifs. Motifs are ordered by significance from top left to bottom right. Each pair of panels corresponds to the activities inferred from CAGE (left) and microarray data (right). The activities inferred for the three biological replicates are shown in red, green, and blue.

The transcriptional network...

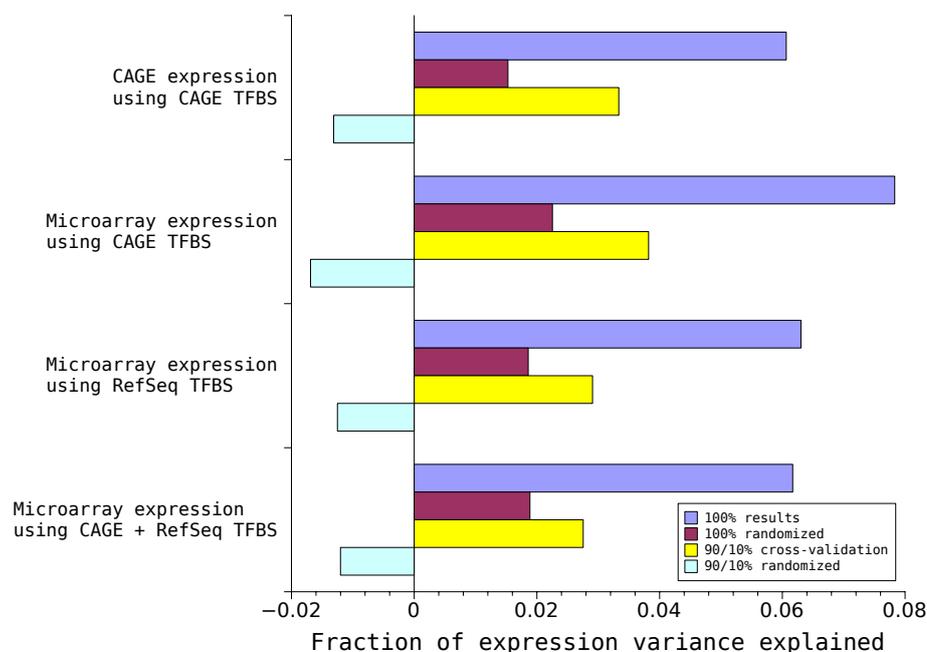


Figure 5.12: Fraction of expression signal explained by the motif activities for different data sets under permutation and 10-fold cross validation tests. Different combinations of expression data and TFBS predictions tested were (a) expression variance of 29,857 CAGE promoters modeled using TFBS predictions from CAGE defined promoters, (b) expression variance of the 8,416 expressed array probes that are associated with both a RefSeq and a CAGE promoter, using TFBSs from CAGE defined promoters, (c) expression variance of the same 8,416 array probes using TFBSs from Refseq defined promoters, and (d) expression variance of all 11,995 expressed array probes using CAGE TFBS predictions whenever available, and Refseq TFBS prediction when no CAGE promoter was associated with the transcript. For each we determined the fraction of expression signal (expression variance minus variance in replicate noise) that is explained by the model (dark blue), when the association between promoters and expression profiles is randomly permuted (purple/brown), under 10-fold cross-validation (yellow), and under 10-fold cross-validation of the randomly permuted data (light blue). (*continued on the next page*)

Figure 5.12: (*continuation from the previous page*) The model explains 6% of the expression signal of all 29,857 promoters, comparable with statistics obtained in recent work(151) for the comparatively simpler task of explaining expression differences between pairs of samples for a selected set of highly varying genes. Comparison of the amount of expression signal explained by the model compared to a data-set in which the assignment between promoters and expression profiles is randomly permuted (1.5% of expression signal explained) demonstrates the extreme significance of the inferred activity profiles (estimated p-value $2.85 * 10^{-1554}$). A 10-fold cross-validation test (on average 3.4% explained versus -1.2% ‘explained’ for permuted promoters in a 1000 iterations, which corresponds to a difference of 170 standard deviations) further demonstrates the validity of the fitting. The fact that the 10-fold cross-validation of the randomized data resulted in negative values indicates that the residual variance after prediction is larger than the original variance. Comparison of the explained expression signal in (b) and (c), where we considered the 8,416 expressed microarray probes that are associated with both CAGE and RefSeq promoters, demonstrates that the predicted TFBSs in CAGE promoters provide significantly better fits than the TFBSs in the corresponding RefSeq promoters, i.e. 7.8% versus 6.3% of explained expression signal. Note that, because the set of promoters/probes fitted in (a), (b,c), and (d) are different, the fractions of expression signal explained cannot be compared across these different data-sets. Only the values in (b) and (c) can be directly compared.

The transcriptional network...

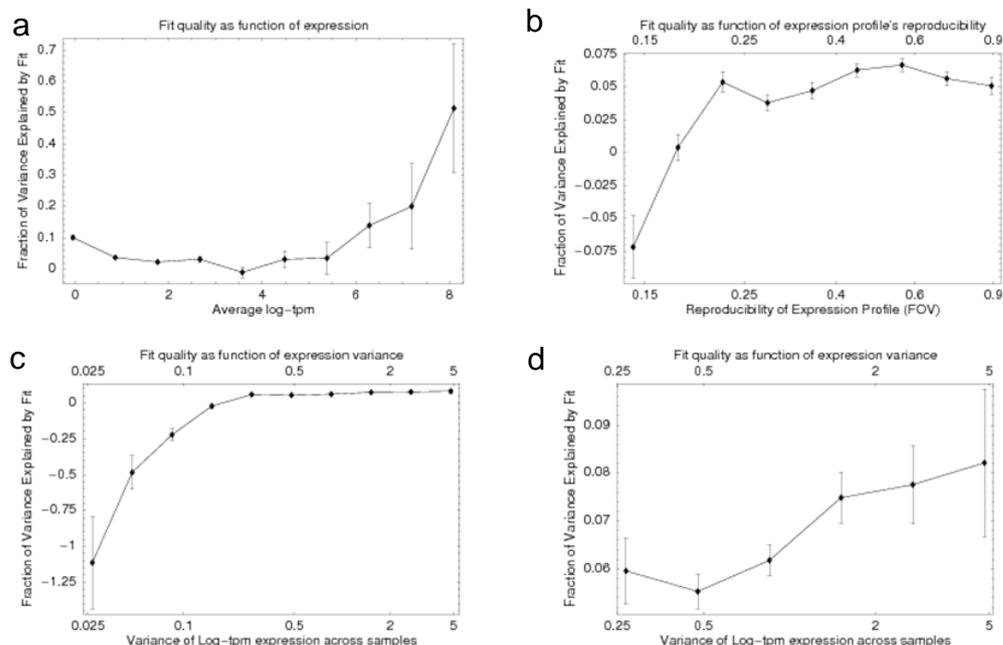


Figure 5.13: Quality of the fits as a function of various CAGE promoter statistics. (a) Mean fraction of expression variance (FOV) explained by the fits as a function of the absolute expression (average log-tpm) of the promoter. (b) Mean fraction of expression variance (FOV) explained by the fits as a function of the reproducibility of the promoter's expression profile, as estimated by the fraction of the variance in the promoter's expression profile that is reproduced across the 3 replicates (FOV). (c) Mean fraction of expression variance (FOV) explained by the fits as a function of the variance of the promoter's expression profile. (d) Blow up of the right half of panel (c). For each statistic all CAGE promoters were divided into 10 bins and for each bin the average FOV and its standard-error (shown as error bars) were determined. Note that all FOVs are as determined from a single fit of activities based on the expression of all promoters, i.e. we do not re-estimate motif activities based on different promoter subsets.

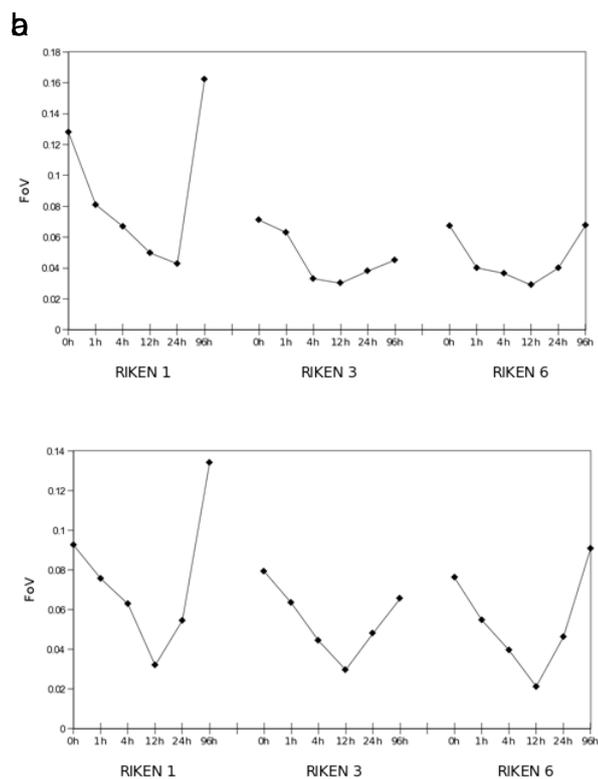


Figure 5.14: Quality of the fits at each time point for all replicates. (a): Quality of the fits as measured by FOV (Fraction of Variance in the expression across all promoters explained by the fit) for each time point in each of the CAGE replicates. (b): Quality of the fits as measured by FOV (Fraction of Variance in the expression across all probes explained by the fit) for each time point in each of the Illumina micro-array replicates.

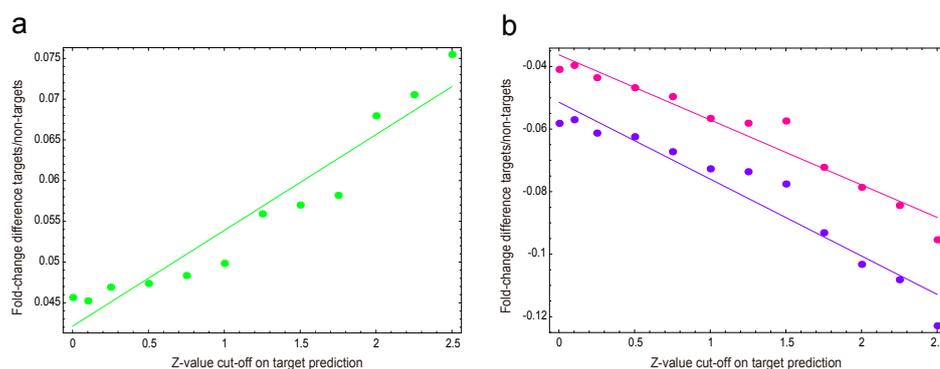


Figure 5.15: Log expression ratio (fold-change) differences of predicted targets and non-targets for several different siRNAs. (a) Difference in average log expression ratio upon siRNA knockdown between predicted targets and non-targets as a function of the z-value cut-off on the target prediction for knockdown of SP1. (b) Difference in average log expression ratio upon siRNA knockdown between predicted targets and non-targets as a function of the z-value cut-off on the target prediction for knockdowns of PU.1 (SPI1) using two different siRNAs (PU.1 in pink and PU.1_2 in purple). All lines are linear regression fits to the data.

5.5.4 Supplementary Figures

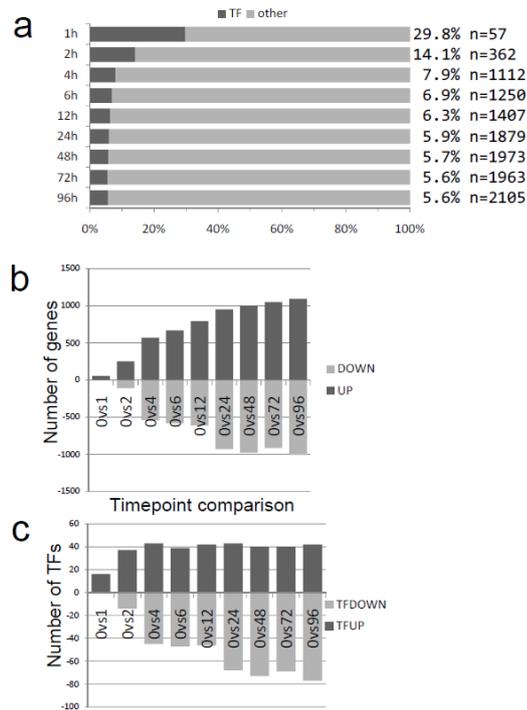


Figure 5.16: Early differentiation involves proportionally more TFs than non-TFs. (a) Using microarrays we count the number of genes with significant differences in expression levels compared to the undifferentiated state. Note: Early differentiation is enriched for changes in TFs. (b) Induction and repression of all genes during PMA differentiation (relative to 0h). (c) Induction and repression of TFs during PMA differentiation relative to 0h.

5.5 Supplementary Tables

Number of promoters	Number of genes
1	3885
2	2176
3	1305
4	780
5	446
6	279
7	178
8	119
9	96
10	56
11	32
12	28
13	13
14	14
15	10
16	11
17	6
18	8
19	2
20	2
21	1
22	2
24	2
29	1
Total	9452

Table 5.1: Distribution of the number of promoters per gene (zero counts not shown)

We identified 9452 genes with at least one CAGE-defined promoter. The promoters shown in this table account for 24,327 out of the 29,857 promoters identified in total. 300 promoters are associated with two genes and 8 promoters with three genes. The remaining 5530 promoters were not assigned to any gene.

5.6 Supplementary Methods

5.6.1 DeepCAGE

The preparation of the CAGE library from total RNA was a modification of methods described by Shiraki et al(1) and Kodzius et al(30), adapted to work with the 454 Life Sciences sequencer.

5.6.2 CAGE tag mapping

Deep sequencing of CAGE tags was done in triplicate at 0, 1, 4, 12, 24 and 96 hours of PMA treatment for a total of 18 samples. A novel alignment method, nexalign (16), was used to align all CAGE tags to the human genome reference sequence (hg18) using a layered, iterative approach. Firstly, tags were matched exactly to the genome and their positions recorded. Secondly, tags that did not match in the first pass were subjected to single base pair substitutions at every position and realigned. Finally, those tags that still did not map were subjected to mapping with indels and aligned to the genome. After this, the match that contained the fewest errors for a given tag was designated the “best” match. For the majority of tags the “best” match was unique on the genome. However, if a tag matched multiple locations at a best match level, a multi-mapping CAGE tag rescue strategy, previously described by Faulkner et al.(17) was used to assign tags to their most probable location. Finally, a filter was applied to remove rRNA-derived tags.

5.6.3 CAGE expression normalization, noise analysis, and promoter construction

The detailed procedures and mathematical derivations involved in our normalization, noise-analysis, and promoter construction are described in Chapter 3. To normalize the deepCAGE expression data we chose a reference power-law distribution with exponent -1.25. The multiplicative noise size σ was estimated to be $\sigma = 0.245$. The promoterome of the THP cell line was constructed based on 18 samples (i.e. the different cell lines and tissues were not included as opposed to the results described in Chapter 3).

5.6.4 Gene assignment for CAGE promoters

We obtained the genomic mappings of all human mRNAs from the UCSC BLAT alignments, discarded mRNAs whose 5' ends don't map, and then associated each promoter with all mRNAs whose mapped TSS is within 1000 base pairs of the CAGE promoter. Using the mapping from mRNAs to Entrez genes provided by NCBI we

The transcriptional network...

associated promoters with Entrez genes and constructed the gene locus (union of all mRNA mappings) of each gene.

5.6.5 deepCAGE expression Analysis

We define the normalized expression e_{ps} of promoter p in sample s as

$$e_{ps} = \log \left(t_{ps} + \frac{1}{2} \right) - \left\langle \log \left(t_p + \frac{1}{2} \right) \right\rangle$$

where t_{ps} is the normalized number of tags per million from promoter p in sample s , and the second term is the average of the first term over the 6 time points in the replicate. For the microarray probes the expression e_{ps} is similarly given by the log-intensity of the probe in sample s minus the average log-intensity of the probe across the 6 time points in the replicate. Probes with detection probability less than 0.99 in all samples were discarded.

5.6.6 Expression signal versus replicate noise

For each CAGE promoter and each microarray probe we compared the total variance in the expression profile, i.e. across all time points and replicates, with the variance across replicates for each time point to estimate the fraction f_p of the total variance (FOV) that is reproducible across replicates.

For each promoter we estimated the fraction of the variance in its expression values that could be explained theoretically, i.e. the fraction that is not due to noise. To do this we compared the variance of expression at the same time point across replicates with the total variance, i.e. across all replicates and time points. For each promoter p we started from the log-expression values

$$x_s^i = \log \left(t_s^i + \frac{1}{2} \right) - \left\langle \log \left(t^i + \frac{1}{2} \right) \right\rangle,$$

where t_s^i is the normalized tag-per-million count of the promoter in replicate i and time point s , and the average in the second term is over the 6 time points in the replicate. That is, $\sum_s x_s^i = 0$ for each replicate i when summed over the time points s . We assume that x_s^i is the sum of a “true” expression value δ_s (which is of course the same for all replicates) and replicate noise. We denote by σ^2 the size of the replicate noise, and by τ^2 the size of the variance in true expression. Using this, the prior probability of the true expression values is given by a Gaussian:

$$P(\delta_s | \alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp \left(-\frac{\alpha}{2} (\delta_s)^2 \right),$$

5.5.6 Supplementary Methods

where $\alpha = \frac{1}{\tau^2}$. Similarly the probability of the observed expression values given the true expression values and size of the noise is

$$P(x_s^i | \delta_s, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x_s^i - \delta_s)^2\right),$$

where $\alpha = \frac{1}{\tau^2}$, where $\beta = \frac{1}{\sigma^2}$. Using these two expressions we obtain for the probability of the data given α and β

$$\begin{aligned} P(x_s | \alpha, \beta) &= \int_{-\infty}^{\infty} P(\delta_s | \alpha) \prod_{i=1}^r P(x_s^i | \delta_s, \beta) d\delta_s \propto \\ &\propto \sqrt{\frac{\alpha}{\alpha + \beta r}} \beta^{r/2} \exp\left(-\frac{1}{2} \left[\frac{\beta^2 r^2}{\alpha + \beta r} \text{var}(x_s) + \frac{\alpha \beta r}{\alpha + \beta r} \langle (x_s)^r \rangle \right]\right), \end{aligned}$$

where r is the number of replicate (3 in our case),

$$\text{var}(x_s) = \frac{1}{r} \sum_{i=1}^r (x_s^i - \langle x_s \rangle)^2$$

is the variance in expression across the replicates for time point s , and

$$\langle (x_s)^2 \rangle = \frac{1}{r} \sum_{i=1}^r (x_s^i)^2$$

is the average squared log-expression at time point s . To get the probability over all time points we simply take the product of the above expression over all time points, i.e.

$$P(x | \alpha, \beta) = \prod_s P(x_s | \alpha, \beta).$$

We are interested in calculating the fraction f of the total expression variance (FOV) that is reproducible across the replicate. This fraction f is given by

$$f = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{\beta}{\alpha + \beta}.$$

We write $P(x | \alpha, \beta)$ in terms of f and β and we integrate over β to obtain the probability of the data as a function of f only, i.e.

$$P(x | f) = \int P(x | f, \beta) \frac{d\beta}{\beta}.$$

We then finally find:

$$P(x | f) \propto \left(\frac{1-f}{1+(r-1)f} \right)^{n/2} \left(\frac{rf \text{var}(x) + (1-f)\langle x^2 \rangle}{1+(r-1)f} \right)^{-nr/2},$$

The transcriptional network...

where n is the number of time points (6 for our case),

$$var(x) = \frac{1}{n} \sum_{s=1}^n \langle (x_s)^2 \rangle$$

is the average squared log-expression across all replicates and time points. Finally, we use the expression $P(x|f)$ to calculate the expected value of f , i.e.

$$\langle f \rangle = \frac{\int f P(x|f) df}{\int P(x|f) df}.$$

Since this integral can generally not be performed analytically we approximate it numerically (for each promoter and probe) by a sum over 100 equal-sized bins of size 0.01 (given the relatively small number of samples per promoter this bin-size is always small compared to the width of the distribution over f).

The distribution of FOV across all promoters and across all probes was summarized by their 5, 25, 50, 75, and 95 percentiles (Fig. 5.9). As shown in Supplementary figure 5.9a the FOV we observe for CAGE promoters are clearly lower than the FOVs observed for Illumina probes. That is, the expression profiles of CAGE promoters typically vary more across replicates than the expression profiles of micro-array probes. One contributing factor is the limited depth of the CAGE sequencing. That is, CAGE measures a much larger number of independent expression profiles than the micro-array, and many of the CAGE promoters have low overall expression. Because of the Poisson sampling noise in CAGE sequencing, promoters with low expression will generally show noisier expression profiles. Since deepCAGE is a relatively new technology, we currently have only limited insight into other factors that may contribute to noise in the expression profiles. One possible contributing factor is the addition of barcodes to the CAGE tags, as we have observed that replicate samples using different barcodes show larger variations than replicates using the same barcodes (data not shown).

To compare deepCAGE and microarray expression measurements we associated microarray probes with CAGE promoter regions whenever the probe intersected a known mRNA whose mapped 5' end was within 1000 bps of the promoter region. We selected all probe/promoter region pairs that are one-to-one associated with each other and calculated the Pearson correlation coefficients of their expression profiles across all samples and replicates. For each probe and promoter region we calculated an average expression profile by averaging the 3 replicate measurements at each time point, and also obtained the Pearson correlation coefficients of the average expression profiles of all probe/promoter region pairs. We collected all microarray probes that were associated with multiple CAGE promoter regions and calculated Pearson correlation coefficients between the probe expression profiles and the total expression from the associated CAGE promoter regions (summing the tags from the different promoter regions) (Fig. 5.9b)

5.6.7 Construction of position specific weight matrices

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for human transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory. Our main objectives were:

1. To remove obvious redundancy, we aim to have no more than 1 WM representing any given TF, and where multiple TFs have WMs that are indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.
2. Associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.
3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consists of:

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM.
2. The collection of TRANSFAC vertebrate WMs (version 9.4)
3. The amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs.
4. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).
5. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles (186).

We start by removing the most basic redundancy from TRANSFAC. TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and should therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically

The transcriptional network...

highly similar and appear redundant. Therefore, for each TF with multiple WMs in TRANSFAC we choose only a single ‘best’ WM based on TRANSFAC’s own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score.

Next we ran Hmmer with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all transcription factors (E-value cut-off 10^{-9}) associated with either JASPAR or TRANSFAC matrices. We then replaced each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR or TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF. From this we create a list of ‘necessary WMs’. For each human TF we obtain the JASPAR WM with the highest percent identity. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations:

1. Different WMs for TFs with identical or near identical DBDs.
2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtain three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.
2. The overlap of the binding sites predicted by each WM. We use MotEvo as described in the methods to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the sites predicted by the two WMs.
3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the WMs and calculate the likelihood-ratio of the sets of sites assuming they derive from a single underlying WM and assuming the set of sites for each WM derives from an independent WM.

For each of these three criteria we set a cut-off: 95% identity of the DBDs, 60% overlap of predicted TFBSs, and a likelihood-ratio of $exp(40)$. Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster and calculating the sum of the base-counts in each column. For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2) and we used these to replace the corresponding WM in the list. For PU.1 we inferred a new WM from the top 50 target regions according to our ChIP-chip data.

Finally, we used MotEvo to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed *new* WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates the position-specific prior probabilities) and using only sites with a posterior probability of at least 0.5. Our final list contains 201 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the human TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 342 human TFs are associated with our 201 WMs. The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (<http://www.swissregulon.unibas.ch>).

5.6.8 Binding Site Predictions

We extracted all position specific weight matrices (WMs) from the JASPAR and TRANSFAC[®] databases that are associated with TFs of multi-cellular eukaryotes. For a few TFs (SP1(187), OCT4, NANOG(188)) we extracted WMs from the literature, and for PU.1 we inferred a new WM using the PhyloGibbs algorithm(135) (see below). WMs were associated with human TFs by matching their DNA binding domain sequences. Whenever both TRANSFAC and JASPAR WMs were available for a given TF only the JASPAR WM was used. Redundancy was removed by clustering WMs that are either highly similar themselves, are associated with equal or highly similar TFs, or predict highly overlapping sets of sites. All clusters were checked manually. For each cluster a fused WM was obtained by aligning matrices within the cluster. After a first round of prediction using these curated WMs, new matrices were constructed from the predicted sites, weighing each predicted site by its posterior probability.

For each promoter region, orthologous regions in Rhesus Macaque, Dog, Cow, Horse, Mouse and Opossum were extracted using the pairwise genome alignments

The transcriptional network...

provided by UCSC. The sets of orthologous sequences from 300 base pairs upstream to 100 base pairs downstream of each promoter region were aligned using T-Coffee(129). In a completely analogous manner multiple alignments were created for the proximal promoter regions of all RefSeq starts.

TFBSs were predicted for all 201 motifs in all multiple alignments of proximal promoters using the MotEvo algorithm(39). Like the Monkey algorithm(158) MotEvo incorporates comparative genomic information by using a specific evolutionary model for the evolution of regulatory sites for the motif as well as for the neutral background evolution. In contrast to Monkey, MotEvo incorporates the possibility that sites are under selection in only a subset of the species in the alignment. In addition, MotEvo uses a more advanced background model that distinguishes neutrally evolving background sequences from background sequences that are under purifying selection.

To incorporate the positional preferences of different motifs we adapted MotEvo to employ position-dependent prior probabilities. For each motif m the prior $\pi_m(x)$ denotes the probability that, in a randomly chosen promoter, a site for m occurs at position x relative to the TSS of the promoter. For each motif the prior $\pi_m(x)$ was fitted using expectation maximization starting from a uniform prior. Using the fitted priors, posterior probabilities were assigned to all predicted binding sites. Finally, all binding sites with posterior less than 0.25 were discarded and for each promoter/motif combination the score N_{pm} is given by the sum of the posterior probabilities of the remaining sites for m in promoter p . Motifs that had less than 150 predicted sites across all promoters were removed from further analysis, leaving 167 motifs.

5.6.9 Motif Activity Inference

With e_{ps} the expression level of promoter p in sample s , N_{pm} the predicted number of functional sites for motif m in promoter p , and A_{ms} the activity of motif m in sample s , we fit a model of the following form

$$e_{ps} = c_p + \tilde{c}_s + \sum_m N_{pm} A_{ms} + noise,$$

where c_p is a promoter-dependent constant (i.e. the basal expression of the promoter) and \tilde{c}_s is a sample dependent-constant. We first fit these constants. Using the fact that $\sum_s e_{ps} = 0$ for each promoter, and defining the site-count averages

$$\langle N_m \rangle = \frac{1}{P} \sum_p N_{pm},$$

where P is the total number of promoters, we can rewrite the model as

$$e_{ps} = \sum_m (N_{pm} - \langle N_m \rangle) A_{ms}.$$

5.5.6 Supplementary Methods

The noise is assumed to be Gaussian of unknown variance with the noise variance σ^2 the same at each promoter (but possibly varying from sample to sample). Under this assumption the likelihood for sample s is given by

$$L_s \propto \sigma^{-\frac{P}{2}} \exp \left[-\frac{\sum_p (e_{ps} - \sum_m (N_{pm} - \langle N_m \rangle) A_{ms})^2}{2\sigma^2} \right].$$

To minimise over-fitting we use a prior probability over activities that is centered around zero

$$P(A_{ms}) \propto \exp \left[-\frac{1}{2} \left(\frac{A_{ms}}{\tau} \right)^2 \right]$$

and we set $\tau = 0.1$. The posterior distribution for the activities in sample s takes then the general form

$$P(A_s | e) \propto \exp \left[-\frac{P}{\chi_s^2} \sum_{m, \tilde{m}} (A_{ms} - A_{ms}^*) C_{m, \tilde{m}}^{-1} (A_{\tilde{m}s} - A_{\tilde{m}s}^*) \right],$$

where the A_{ms}^* are the activities with maximal posterior probability which are determined by singular value decomposition, the activity covariance matrix $C_{m, \tilde{m}}$ is a function of the site-counts N_{pm} , and χ_s^2 is the residual variance after fitting, i.e.

$$\chi_s^2 = \frac{1}{P} \sum_p \left(e_{ps} - \sum_m (N_{pm} - \langle N_m \rangle) A_{ms}^* \right)^2.$$

From this we can rigorously calculate a standard-error σ_{ms} for the activity of each motif in each sample, and calculate a z-value

$$z_{ms} = \frac{A_{ms}^*}{\sigma_{ms}}.$$

Note that, given the Gaussian form of the posterior for the activity of each motif, the p-value for the significance of the motif's activity can be directly determined from the z-value. Finally, we calculate an overall significance of the motif by averaging its z-value over the samples

$$z_m = \sqrt{\frac{1}{S} \sum_s (z_{ms})^2},$$

where S is the number of samples. Analogously, the posterior distribution of motif activities is inferred from the expression profiles of microarray probes and the site-counts of associated promoters. Final motif activities A_{mt} as a function of time are inferred by combining the posterior distributions from the 3 replicates for both CAGE

The transcriptional network...

and the microarrays assuming one underlying activity for each motif at each time point.

To quantify the quality of the fits we first calculate the “expression signal”, i.e. the total variance that could possibly be explained by the fit. The expression variance of a promoter is given by $v_p = \frac{1}{S} \sum_s (e_{px})^2$ and with f_p the FOV for this promoter, the total expression signal is $E = \sum_p f_p v_p$. The fraction ρ of expression variance explained by the fit is then

$$\rho = \frac{\sum_s (\sum_p (e_{ps})^2 - \chi_s^2)}{E}.$$

To select core motifs we combined the posterior distributions over motif activities from the posterior distributions of the 3 replicates for both CAGE and the microarrays (5.6.10). The result is a final average motif activity A_{mt}^f for each motif at each time point, plus a standard-error σ_{mt}^f . Using this we calculate a final significance z_m^f for each motif. In addition we calculate the fraction of variance in motif activity that is reproduced across the replicates of both CAGE and microarray (5.6.10). The 30 selected core motifs are all motifs with z-values at least 3.75 and FOV at least 0.75 (Fig. 5.2). We clustered the activity profiles of the core motifs using a Bayesian hierarchical clustering method (5.6.11). Briefly, starting from the posterior distributions of motif activities for all motifs we can calculate, for any pair of motifs, the probability that their activity profiles are the same (i.e. within noise). We iteratively clustered the two motifs with highest probability of being the same and determined the new posterior probability of motif activities for the cluster. We stopped when the probability for the highest scoring pair fell below a cut-off that we determined by hand.

5.6.10 Combining motif activities from replicates and motif FOV

For each motif m and each sample s our inference provides a fitted activity A_{ms}^* and its associated standard-error σ_{ms} . Therefore, if we ignore covariances between the inferred activities of different motifs, the posterior distribution for the activity of motif m in sample s is given by

$$P(A_{ms}) = \frac{1}{\sqrt{2\pi}\sigma_{ms}} \exp\left(-\frac{1}{2} \left(\frac{A_{ms} - A_{ms}^*}{\sigma_{ms}}\right)^2\right).$$

For each of the 6 time points we have 6 independent posterior distributions of motif activity, namely 3 replicates for both CAGE and microarray data. We now infer an overall motif activity by combining the 6 posterior distributions. Let's focus on a single motif and let α_t denote the final inferred activity of the motif at time t ,

let C_t^i be the inferred activity from CAGE replicate i , σ_t^i its standard-error, M_t^i the inferred activity from microarray replicate i , and τ_t^i its standard-error. The posterior distribution for α_t is now given by:

$$\begin{aligned} P(\alpha_t | C, M, \sigma, \tau) &\propto \prod_i \exp\left(-\frac{1}{2} \left(\frac{\alpha_t - C_t^i}{\sigma_t^i}\right)^2 - \frac{1}{2} \left(\frac{\alpha_t - M_t^i}{\tau_t^i}\right)^2\right) \propto \\ &\propto \exp\left(-\frac{1}{2} \left(\frac{\alpha_t - \alpha_t^*}{\sigma_t^*}\right)^2\right), \end{aligned}$$

with

$$\alpha_t^* = \frac{\sum_{i=1}^r C_t^i (\sigma_t^i)^{-2} + M_t^i (\tau_t^i)^{-2}}{\sum_{i=1}^r (\sigma_t^i)^{-2} + (\tau_t^i)^{-2}}$$

and

$$\sigma_t^* = \frac{1}{\sqrt{\sum_{i=1}^r (\sigma_t^i)^{-2} + (\tau_t^i)^{-2}}}.$$

That is, the posterior distribution is again Gaussian but with updated mean and standard-error. Finally, we calculate a z-value for the combined activity profile of the motif

$$z_m = \sqrt{\frac{1}{6} \sum_{t=1}^6 \left(\frac{\alpha_t^*}{\sigma_t^*}\right)^2}.$$

For each motif we also quantify the extent to which the different replicates and measurement technologies (CAGE and microarray) lead to the same inferred activity profiles. For this we calculate a FOV exactly as described in 5.6.12.

5.6.11 Clustering motifs on activity profiles

We noticed the inferred activity profiles of several motifs are highly similar suggesting there are clusters of motifs with essentially the same activity profiles. We thus devised a clustering procedure that joins together motifs whose inferred activity profiles are statistically indistinguishable. To this end we needed to calculate, for any set C of motifs, the probability of the data under the assumption that their inferred activity profiles all derive from the a common underlying activity profile. Let α_{mt}^* denote the inferred combined activity of motif m at time t , let σ_{mt}^* denote the standard-error associated with this activity, let C denote a cluster of motifs, and let γ_t be the (unknown) common activity profile of the motifs in the cluster. The probability of the inferred activities given γ and the standard-errors is then given by

$$P(\alpha | \gamma, \sigma) = \prod_{m \in C} \left[\prod_t \frac{1}{\sqrt{2\pi\sigma_{mt}^*}} \exp\left(-\frac{1}{2} \left(\frac{\alpha_{mt}^* - \gamma_t}{\sigma_{mt}^*}\right)^2\right) \right].$$

The transcriptional network...

We now use a prior over the underlying activity profile γ that is the same as we used for inferring the activity profiles from the independent data-sets, i.e.

$$P(\gamma_t | \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{1}{2} \left(\frac{\gamma_t}{\tau}\right)^2\right),$$

where we again use $\tau = 0.1$. By integrating over the unknown activity profile γ we then obtain the probability of the inferred activity profiles in the cluster under the assumption that they are all the same up to noise, i.e.

$$P(\alpha | \sigma) = \prod_{m \in C} \left[\int \prod_t P(\alpha_t^* | \gamma_t, \sigma_t^*) P(\gamma_t | \tau) d\gamma_t \right].$$

These integrals are all Gaussian integrals and can be performed analytically.

We use the result to hierarchically cluster the 30 core motifs base on their activity profiles. We start with letting each motif be a cluster by itself and calculate, for each pair, the likelihood-ratio of the probability of the data before and after clustering. We then iteratively cluster the pair of motifs with highest likelihood-ratio. Note that when two motifs are clustered we recalculate their average activity profile and associated standard-error of the average exactly in the same way as we do when we combine the data from the replicates (i.e. we treat the inferred activities of the different motifs in the cluster just like we treat the inferred activities from different replicates for the same motif). At each iteration we also keep track of the total probability of the data in the current clustering state. The cut-off for termination of the hierarchical clustering was chosen by hand (essentially where the first large drop in likelihood of the clustering state is observed).

5.6.12 Permutation and Cross-validation tests

We tested the significance of the fits using the following permutation test: We randomly permuted the association between the site-counts N_{pm} and the expression profiles e_{ps} so that each promoter is now assigned the site-counts from a randomly chose other promoter. The model was then fitted on this randomized data set and the fraction of expression signal explained by the fit was calculated exactly as for the original data. This procedure was repeated 1'000 times. For the CAGE promoters, the average fraction of expression signal explained was 0.015 with a standard-deviation of 0.00054, corresponding to a difference of 84.5 standard-deviations with the fitted fraction on the real data (0.061). Assuming the fitted fractions for the permuted data-sets are Gaussian distributed this would correspond to a p-value of $2.85 \cdot 10^{-1554}$.

For the cross-validation test we randomly divided the promoters in 10 subsets of equal size. For each subset we use the remaining 90% of the promoters to fit motif activities and used these to predict the expression values of the promoters in

the set. Combining the results from all 10 subsets we again calculated the fraction of expression signal explained by the fit. Cross-validation was also applied to the data-set with permuted promoters.

For comparison of the fits based on CAGE versus RefSeq promoters we selected all microarray probes that intersect a RefSeq transcript and that are one-to-one associated with a CAGE promoter region. We fitted the expression data of all these probes once using the site-counts N_{pm} from the associated CAGE promoters and once using N_{pm} from the RefSeq promoters.

5.6.13 Motif target predictions

We predict a regulatory edge between a motif and a promoter when the promoter has predicted binding sites for the motif ($N_{pm} \geq 0.25$) and the expression profile of the promoter correlates significantly with the inferred final activity profile A_{mt}^f of the motif. In particular, the correlation between the expression profile and activity profile is given by $c_{pm} = \frac{1}{6} \sum_t e_{pt} A_{mt}^f$, where e_{pt} is time-dependent expression profile of the promoter averaged over the replicates. Using only a single motif to explain the expression profile, the residual variance is

$$\chi_{pm}^2 = \frac{1}{6} \sum_t \left(e_{pt} - c_{pm} A_{mt}^f \right)^2.$$

Finally, the z-value that quantifies the significance of the regulatory interaction between motif and promoter is

$$z_{pm} = \sqrt{\frac{6}{\chi_{pm}^2} c_{pm}}.$$

Note that, although c_{pm} can be negative, we only consider regulatory interactions with non-negative correlation. For the Gene Ontology analysis, target gene sets of core motifs (z-value ≥ 1.5 for the association of a motif to promoters of target genes, z-values were averaged if there was more than one promoter associated with a gene) were tested for functional enrichment (189). All genes with CAGE defined promoters were chosen as the background.

5.6.14 siRNA edge validation and core network construction

Predicted regulatory interactions were tested using siRNA knockdowns of 28 TFs that are associated with motifs. For each TF knockdown we collected all microarray probes that are associated with promoters and calculated, for each probe, the average z-value of the predicted regulatory interaction from the TF's motif to the promoters associated with the probe. At different cut-offs in z-value we then divided the probes into "targets" of the motif, i.e. those with a z-value above the cut-off, and "non-targets"

The transcriptional network...

of the motif, i.e. all probes with z-value below the cut-off (this includes probes for which there are no predicted TFBSs in the associated promoters), and calculated the difference in average expression ratio (knockdown minus mock) of targets and non-targets. For each knockdown we calculated the Pearson correlation coefficient between the z-value cut-off on target prediction and the observed difference in average expression ratio of targets and non-targets. To assess the significance of the differences in average expression ratio we set an intermediate cut-off of $z = 1.5$, calculated the distribution of expression ratio for targets and non-targets, determined their means (μ_t and μ_{nt}) and variances (v_t and v_{nt}), and determined a z-value for the expression ratio difference as

$$z = \frac{\mu_t - \mu_{nt}}{\sqrt{v_t/N_t + v_{nt}/N_{nt}}},$$

where N_t and N_{nt} are the number of target and non-target probes, respectively.

The core network was constructed by first selecting all predicted regulatory interactions (z-value at least 1.5) between core motifs and promoters that are associated with a gene which is a TF that in turn is associated with a core motif. This set of predicted regulatory interactions was then filtered by choosing only interactions that have independent experimental support of at least one of the following types. 1) The regulatory interaction has been reported in the literature 2) There is a ChIP-chip experiment in which binding of one of the TFs associated with the motif to the promoter of the target gene has been reported. 3) In our siRNA experiments the target promoter is observed to be perturbed in expression (B-statistic larger than zero) after knockdown of a TF associated with the motif.

5.6.15 Motif Activity Analysis of TF knockdowns

We applied the motif activity analysis to the microarray expression profiles of all siRNA samples including negative controls. As a result we obtained fitted motif activities A_{ms}^* and standard-errors σ_{ms} for each motif m in each of the siRNA samples s . We combined the inferred activities from replicates and control experiments, and calculated a z-value for the activity change between siRNA and negative control for each TF that was knocked down:

$$z_m^{TF} = \frac{\langle A_m^{TF} \rangle - \langle A_m^{NC} \rangle}{\sqrt{(\sigma_m^{TF})^2 + (\sigma_m^{NC})^2}},$$

where $\langle A_m^{TF} \rangle$ is the average activity of motif m across the replicates in which the TF was knocked down, σ_m^{TF} the standard-error of this average activity, $\langle A_m^{NC} \rangle$ the average activity of motif m in the negative controls, and σ_m^{NC} its the standard-error. The z-values z_m^{TF} characterize the expression changes observed upon siRNA knockdown of the TF in terms of observed changes in motif activities. That is, if z_m^{TF} is highly

positive it indicates that predicted targets of motif m are up-regulated in response to knockdown of the TF. We similarly calculated z-values for motif activity changes across the PMA time course:

$$z_m^{PMA} = \frac{\langle A_m^{96} \rangle - \langle A_m^0 \rangle}{\sqrt{(\sigma_m^{96})^2 + (\sigma_m^0)^2}},$$

where $\langle A_m^{96} \rangle$ is the average activity of motif m after 96 hours of PMA treatment, σ_m^{96} its standard-error, $\langle A_m^0 \rangle$ is the average activity before PMA treatment, and σ_m^0 its standard-error. Given the z-value for the change in motif activity the probability that the motif is up-regulated is given by

$$p_{up}(z) = \frac{1}{2} \text{Erfc} \left(\frac{z}{\sqrt{2}} \right)$$

and the probability that the motif is down-regulated is given by

$$p_{down}(z) = \frac{1}{2} \text{Erf} \left(\frac{z}{\sqrt{2}} \right).$$

Using this we calculated, for each motif m , the probability p_m that the motif is changing in the same direction in both the PMA time course and the TF knockdown:

$$p_m = p_{up}(z_m^{TF})p_{up}(z_m^{PMA}) + p_{down}(z_m^{TF})p_{down}(z_m^{PMA}).$$

Finally, the overlap o^{TF} between TF and PMA time course is defined as the sum of p_m over all motifs divided by the total number of motifs, i.e. the estimated fraction of motifs that change activity in the same direction in knockdown and PMA time course. We calculated the significance of the differentiative overlaps by a permutation test; we randomly permuted the order of the motifs 1000 times and calculated the differentiative overlap for each.

5.6.16 The data and analysis results available from the FANTOM4 web resource

In addition to the data and analysis results amassed here the full set of several Supplementary Tables are available from the FANTOM4 web resource. The data is also available from “GNP Platform” (http://genomenetwork.nig.ac.jp/index_e.html).

The transcriptional network...

Chapter 6

Conclusions

The classical promoter architecture model has been challenged in the FANTOM3 project, which showed widespread transcription initiation events across chromosomes and thus that alternative promoter usage substantially contributes to the complexity of mammalian proteome (12). Due to the introduction of 454 Life Sciences sequencing and the depth achieved in FANTOM4 project, it has become possible to robustly estimate expression levels of individual TSSs. We have shown in Chapter 3 that nearby TSSs are co-regulated rather than independently expressed. Based on this principle, we defined Transcription Start Clusters as genomic loci, which show coherent expression patterns across different experimental conditions. On average, such co-regulation exists on the distance of about 15 nucleotides; however this average is not representative, as we have shown that high- and low-CpG promoters are inherently different. The low-CpG promoters are usually short (only 22% are longer than 10 bps), whereas high-CpG are much larger (80% longer than 10 bps and 40% longer than 100 bps). This latter fact is especially surprising as this class drives expression of many housekeeping, highly-expressed and well-studied genes.

To our knowledge, until then, there was no rigorous model of noise in sequencing data. We have developed such a model from the need of quantitative analysis of deepCAGE data, and used it extensively in the promoterome construction. Similarly, it was assumed that the sequencing data do not need elaborate normalization techniques - simple “*reads per kilobase per million reads*” normalization was assumed to be sufficient. Our work, among others published during the same period, showed the need for more sophisticated normalization schemes.

We have applied the MARA algorithm for the inference of regulatory interactions, which is able to handle large quantities of (possibly noisy) expression data to reliably predict changes in motif activities. It benefits from high-quality promoter annotations and state-of-the-art binding site predictions. The results are straightforward to interpret: the activities indicate how much a motif of interest contributes to promoter expression across conditions, and a list of important motifs is provided. Although

Conclusions

the model is phenomenological, it captures the strengths and directions of activity changes of multiple motifs - quantities that are of great interest when working with a new system.

Currently all the samples are treated independently and in the same way. However, it is more and more often possible to track changes of a perturbation/developmental time course with a fine temporal resolution. In such case the activities of the neighboring time points are expected to differ only slightly. A key extension of the model would be to model the dynamics of these small changes as a property of the regulatory network itself: the activities of the previous time point(s) are directly causing them.

Another future improvement of the MARA strategy is to allow motifs to work as activators for one group of promoters and as repressors for another group. A recent work by Bauer *et al.* (124) has shown that the inclusion of such a possibility in a model of *Drosophila* embryo segmentation allows for a large improvement in predictions of cis-regulatory module expression which cannot be explained by a simple increase in number of free parameters.

Another extension could be the inclusion of distal regulatory elements. It is not clear, however, if assigning enhancers to the closest TSS and the treating them the same ways as proximal promoters would improve the fit. It might not be productive due to the fact that enhancers/silencers are often tissue specific, turned off and on by chromatin modifications, yet they might regulate the same TSS. A model which includes a total amount of TFBSs summed over all the enhancers might become less efficient than a promoter-based model.

We aimed to robustly infer transcription factor activity from expression data. As an available estimate of transcription initiation data, we used microarray and RNA-seq data. The mRNA levels, however, are a result of a more complex process involving (among other factors) transcription elongation, termination and post-transcriptional control. It is hard to reliably judge the importance of the first two processes on a genome-scale level. We do, however, know that the mRNA post-transcriptional control plays a crucial role in gene expression, and that it can happen at virtually any step of RNA metabolism (some of the steps affecting RNA levels, others not). There are multiple types of RNA binding domains which specifically recognize oligonucleotide motifs, and there is often more than one RNA binding domain per protein. The post-transcriptional control is thus a combinatorial process, whereby different motifs regulate gene expression in various ways (190). An important class of RNA binding proteins are the Argonaute family members, endonucleases recognizing specific RNA motifs with the help of small RNAs. Initial results, not shown in this work, of extending MARA by including microRNA site predictions show that the explanatory potential is very close that one of transcription factor motifs in proximal promoters.

The regression model is general enough that it can take different types of data as the input. The applications are beyond the scope of this work. MARA was recently used for explaining histone modification data (H3K27 trimethylation, H3K4

dimethylation and DNA methylation) to explain histone- and DNA methyltransferases attraction by transcription factors in a developing neural cell.

To validate the predicted edges in the transcription regulatory network of a myeloid leukemia cell line, we performed a broad range of validations including literature search, siRNA assays against transcription factors and chromatin immunoprecipitation. A typical usage of the MARA web pipeline would not include such an extensive validation scheme. Sadly, we are still far from a scheme for robust, automated TF \rightarrow TF regulatory edge predictions without the need of validation.

The study of the myeloid leukemia cell line is one of the most complete studies of differentiation and transcription regulatory network reconstruction in a mammalian genome. The results obtained from the different types of expression data (deepCAGE and microarray) and three replicates largely overlap, showing the robustness of the MARA approach. The most striking result of the analysis is the complexity of the resulting regulatory network. Not only does it contain 30 motifs changing activity under the stimulus, but the interconnections between them are also far from a classical cascade structure. There is no master regulator; no knockdown reproduced differentiation completely and there are many loops and feed-forward connections. It seems that in order to guide the network along the differentiation path, downregulation of multiple factors is needed. This is coherent with the view of cellular states as attractor basins in a wider “landscape” (see Fig. 6.1). In order to move out from an undifferentiated state, the transcriptional network requires a perturbation of multiple factors which homeostatically maintain the undifferentiated state. Such a destabilization in response to PMA is accomplished by immediate early genes.

Conclusions

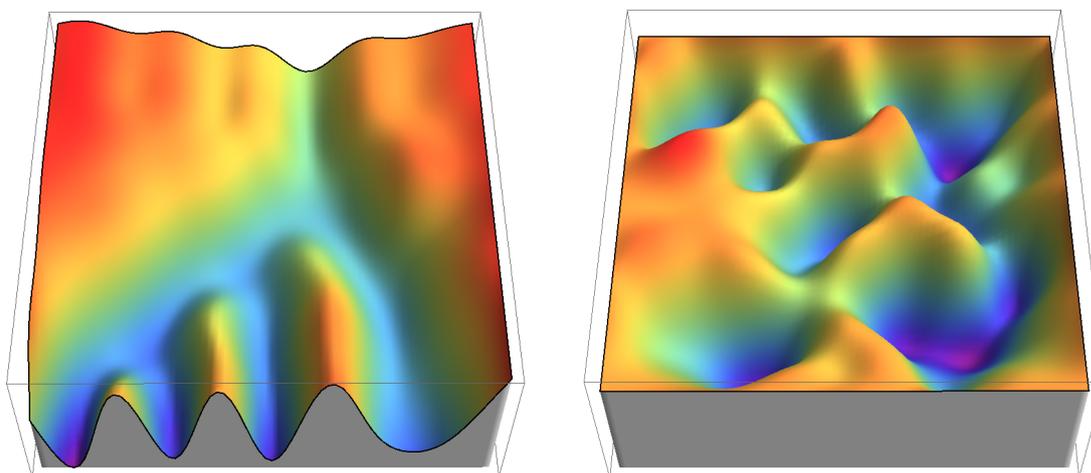


Figure 6.1: A cartoon view of a cellular state landscape: canalization and attractor basins

The phenotypical state corresponds to a point in multidimensional space (here: two-dimensional) of regulatory protein activities. The wiring of the network, especially edges between regulators, determines a direction of the movement, here depicted as slopes of the landscape.

Left panel: the classical view of a cellular differentiation process. The canals represent possible phenotypic states through which a cell travels top to bottom. In the beginning it starts as a stem cell and while traveling, dependent on stimuli and morphogenes, it makes irreversible decisions as to which differentiation path to choose. Intermediate points represent semi-differentiated cell progenitor states, and the bottom states represents a phenotypic space of the terminally developed cells.

Right panel: the attractor view of cellular states. Different cellular states correspond to the attractors (basins). The homeostatic interactions between regulators are the principle which cause such attractors to emerge. A perturbation is needed to move a cell out of its current basin. Subsequently, the cell migrates to a new cellular state depending on the strength and direction of the perturbation. Strong perturbations might lead to nonviable states, as drawn in the middle of the left side of the figure. This view is consistent with recent reports of reprogramming differentiated cells into stem cells (112) or into another differentiated state (113) by perturbing a handful of TFs.

Part II

Function and processing of box C/D snoRNAs

Chapter 7

Introduction

The common function of a diverse class of non-coding RNAs (ncRNAs) is the recognition of a specific locus in a target nucleic acid molecule for enzymatic catalysis by a partner protein. Typically, ncRNA action is driven by base pairing and involves several partner proteins. Such a complex is called non-coding ribonucleoprotein (ncRNP). ncRNPs regulate different levels of gene expression including mRNA splicing, histone pre-mRNA formation (U7), tRNA and rRNA point modifications, pre-tRNA cleavage, mRNA editing, transcription-elongation control, translation, protein trafficking, gene silencing (at mRNA level and chromatin modification level) and telomere synthesis.

snoRNAs are a well-studied classes of small ncRNAs. Evolutionarily, they stem from a point earlier than the split between the Archaea and the Eukarya domains of life. Their name comes from the nucleolar localization of the first members of this family. The family is diverse and at the top level is subdivided into two branches: the box C/D and the box H/ACA snoRNAs. Both of these act primarily as guide molecules for modification of other ncRNAs. The box C/D snoRNAs direct 3'-O-ribose methylation, whereas the box H/ACA snoRNAs guide pseudouridylation. The guided modifications of rRNA are necessary for the proper function of a ribosome. Other targets include snRNAs (eukaryotes), tRNAs (archaea), spliced leader RNAs (trypanosomes).

There exists a number of snoRNAs that have no known target snRNAs or rRNAs. Recently it was shown that one of the member of these “orphans”, HBII-52, exhibits long, perfect complementarity to a pre-mRNA region of serotonin receptor (191). It was shown that it has an important role in choosing a “correct” exon-intron junction in the splicing process. Importantly, deletion of a locus containing HBII-52 leads to a quite common (1 in 12,000 newborns) genetic disease, Prader–Willi syndrome. Encouraged by this initial finding, we performed a computational screen for other potential targets of MBII-52 (which is a mouse ortholog of HBII-52), accompanied with experimental verification. The results and methodology can be found in Chapter 8.

Introduction

A distinctive feature which gives its name to the box C/D snoRNAs is the presence of highly conserved motifs called C and D boxes. There are one or two pairs of these, each pair accompanied with an antisense element. The function of the antisense elements is the recognition of target sites by formation of double stranded RNA hybrid structures. In mature RNP 15.5K/NHPX protein (or L7Ae in archea) binds to the C and D boxes stabilizing the structure; then, close paralogue proteins NOP56 and NOP58 join the complex and recruit enzyme fibrillarin. Fibrillarin is a 2'-O-methyltransferase which methylates precisely that ribonucleotide which is paired with the 5th base of the antisense element. See Figure 7.1.

Constraints on the RNA:RNA duplex formation and stability are key for precise recognition of targets by an antisense element. Fortunately, it was a subject of study by Cavallé and Bachellerie (193). Through extensive experiments these authors have shown that a duplex structure can tolerate many types of irregularities. A bulged nucleotide can be introduced at various positions without a dramatic decrease in the extent of reaction. On the substrate strand duplex bulges are tolerated at almost any position. On the snoRNA strand, the bulge can be positioned anywhere beyond base pair 7. Similarly, the particularly destabilizing G·A apposition and multiple G·U wobbles are tolerated. Shortening of a duplex length from 16 to 12 bp can have a severe impact on methylation status, but a change of binding free energy by increased GC content can fully compensate for the length reduction. Interestingly, there exist naturally occurring methylation sites which require a duplex of only 10 bp in length and which is not particularly GC-rich, suggesting a presence of co-factors. In the following chapters, we use the constraints for an *in silico* screen of putative target sites of MBII-52 snoRNA in mRNAs (Chapter 8) and of v-snoRNA1 in the rRNA (Chapter 9).

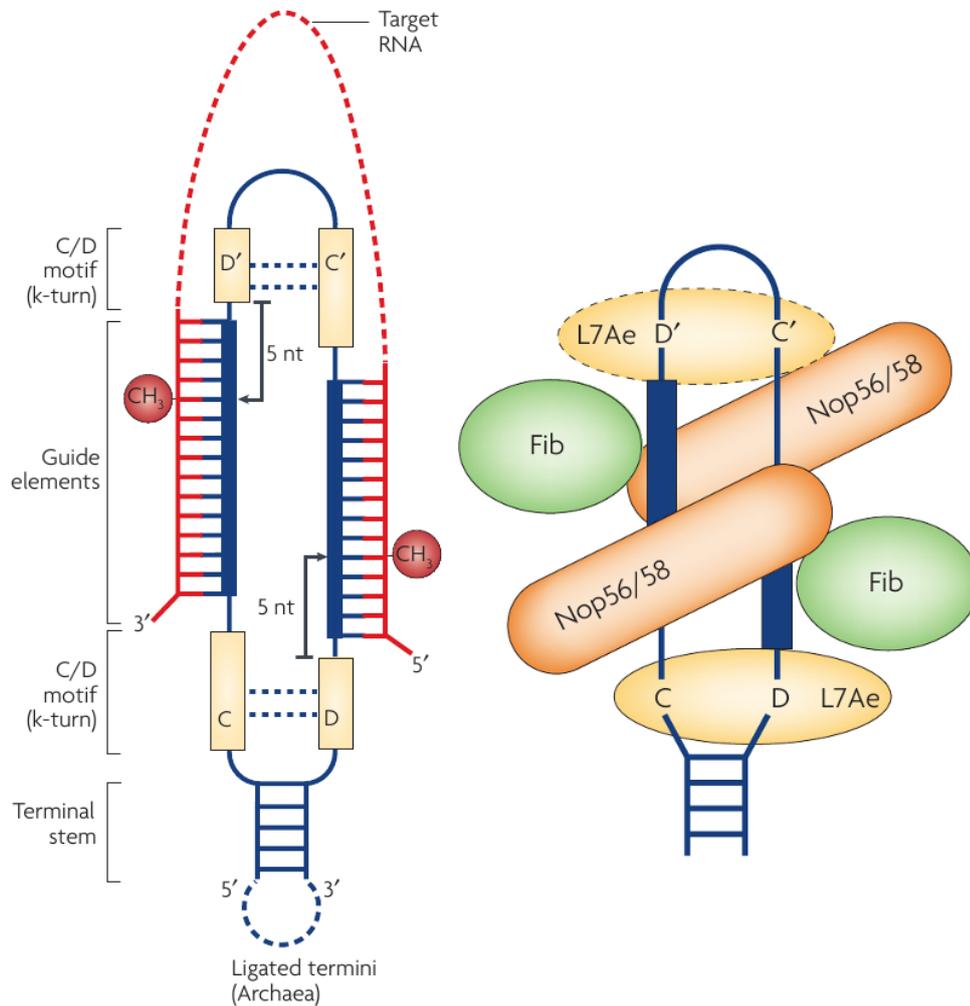


Figure 7.1: Structure of box C/D snoRNAs and snoRNPs

Left panel: Secondary structure of box C/D snoRNAs. The conserved box C (PuUGAUGA) and D (CUGA) sequence elements are tethered by the terminal stem-loop and apical loop and form kink-turns (k-turns). A C/D pair is associated with an antisense element (blue) located upstream of box D which forms base pairs with the target RNA (red). Target RNA is methylated on the ribose of the nucleotide which is base paired with the guide RNA that is 5 nucleotides upstream of box D.

Right panel: Core archeal C/D RNPs. Eukaryotic homologue of L7Ae is 15.5 kD protein also known as NHPX and Snu13p, Nop56/58 is replaced by a pair of NOP56 and NOP58 paralogues.

Reprinted by permission from Macmillan Publishers Ltd: Matera et al. Nature Reviews Molecular Cell Biology 8, 209–220 (March 2007) (192) | doi:10.1038/nrm2124

Introduction

Chapter 8

The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing

Shivendra Kishore, Amit Khanna, Zhaiyi Zhang, Jingyi Hui, Piotr J Balwierz, Mihaela Stefan, Carol Beach, Robert D. Nicholls, Mihaela Zavolan and Stefan Stamm

Human Molecular Genetics, 19(7):1153-64 2010

The loss of HBII-52 and related C/D box small nucleolar RNA (snoRNA) expression units have been implicated as a cause for the Prader-Willi syndrome (PWS). We recently found that the C/D box snoRNA HBII-52 changes the alternative splicing of the serotonin receptor 2C pre-mRNA, which is different from the traditional C/D box snoRNA function in non-mRNA methylation. Using bioinformatic predictions and experimental verification, we identified five pre-mRNAs (DPM2, TAF1, RALGPS1, PBRM1 and CRHR1) containing alternative exons that are regulated by MBII-52, the mouse homolog of HBII-52. Analysis of a single member of the MBII-52 cluster of snoRNAs by RNase protection and northern blot analysis shows that the MBII-52 expressing unit generates shorter RNAs that originate from the full-length MBII-52 snoRNA through additional processing steps. These novel RNAs associate with hnRNPs and not with proteins associated with canonical C/D box snoRNAs. Our data indicate that not a traditional C/D box snoRNA MBII-52, but a processed version lacking the snoRNA stem is the predominant MBII-52 RNA missing in PWS. This processed snoRNA functions in alternative splice-site selection. Its substitution could be a therapeutic principle for PWS.

8.1 Introduction

It has been estimated that 95% of human multi-exon genes undergo alternative splicing(194; 195), indicating that this pre-mRNA processing step is central for human gene expression. Unlike promoter activity that is predominantly reflected in the abundance of transcripts, alternative splicing influences the structure of the mRNAs and their encoded proteins. As a result, it influences binding properties, intracellular localization, enzymatic activity, protein stability and post-translational modification of numerous gene products (reviewed in (196)).

We recently found that usage of the alternative exon Vb of the serotonin receptor 2C (HTR2C) is regulated by expressing a C/D box snoRNA, HBII-52. SnoRNAs are small nuclear RNAs that can be detected in the nucleolus. They reside in introns from which they are released through nuclease action during the processing of the host pre-mRNA. On the basis of their sequence, snoRNAs can be subdivided into C/D and H/ACA snoRNAs. C/D box snoRNAs have C and D boxes as characteristic sequence elements at the ends of the RNA. The 5' and 3' ends of the snoRNA form a short stem that precedes the C and D boxes, which together form a kink-turn (K-turn) structure(197).

A well-understood function attributed to C/D box snoRNAs is the guiding of 2'-O-methylation in ribosomal, transfer and small nuclear RNAs. This activity is achieved through the formation of a specific RNA:RNA duplex between the snoRNA and the target. Most snoRNAs contain two regions to interact with other RNAs, termed the antisense boxes. Each antisense box exhibits sequence complementarity to its target and forms a short, transient double strand with it. On the target RNA, the nucleotide that base pairs with the fifth snoRNA nucleotide upstream of the snoRNA D-box is methylated on the 2'-O-hydroxyl group (reviewed in (192)). Several snoRNAs are complementary to pre-rRNA, but the rRNA is not 2'-O-methylated at the predicted positions(198). Recently, numerous C/D box snoRNAs were discovered that show no clear sequence complementarity to other non-mRNAs, suggesting that C/D box snoRNAs might have functions other than 2'-O-methylation(199)). One of these 'orphan' C/D box snoRNAs is HBII-52 (SNORD 115). It is expressed from the SNURF-SNRPN locus. Loss of expression from this locus is the most likely cause for Prader-Willi syndrome (PWS)(200), which was supported by the recent finding that a microdeletion containing only snoRNAs causes PWS(201). HBII-52 exhibits sequence complementarity to an alternative exon of the human serotonin receptor 2C mRNA and changes alternative splicing of this pre-mRNA (HTR2C) in transfection experiments. This change has also been observed in brain tissue from PWS patients(191) and a mouse model lacking MBII-52 snoRNAs shows differences in pre-mRNA processing of the serotonin receptor(202). Finally, it was reported that an increase of C/D box snoRNA expression from the 15q11-q13 region leads to autistic phenotypes in mice, which further suggests that snoRNAs play an important role in

gene regulation(203).

PWS is a congenital disease with an incidence of about 1 in 8000–20 000 live births. PWS is the most common genetic cause of marked obesity in humans. The excess weight makes PWS the most frequent genetic cause for type II diabetes(200). Early PWS is characterized by a failure to thrive, feeding difficulties and hypogonadism. Later, the patients are characterized by short stature and develop mild to moderate mental retardation, behavioral problems and hyperphagia that lead to severe obesity. Children with PWS show low levels of growth hormone, IGF-I and insulin as well as elevated levels of ghrelin(204; 205; 206) and often exhibit central adrenal insufficiency(207). Subsequently, growth hormone substitution was approved for the treatment of children with PWS(208).

PWS is caused by the loss of gene expression from a maternally imprinted region on chromosome 15q11–q13 (reviewed in (200)). The SNURF–SNRPN locus in the 15q11–q13 region plays a major role in PWS, and its deletion causes PWS-like symptoms in mouse models(209). The SNURF–SNRPN locus spans more than 460 kb and contains at least 148 exons(210). Ten exons in the 5' part of the gene are transcribed into a bicistronic mRNA that encodes the SNURF (SmN upstream reading frame) and the SmN (small RNP in neurons) protein. The locus harbors a bipartite imprinting center that silences most maternal genes of the PWS critical region. Owing to this imprinting, the SNURF–SNRPN pre-mRNA is expressed only from the paternal allele. The large 3'-UTR region of the SNURF–SNRPN locus harbors clusters of the C/D box snoRNAs HBII-85 and HBII-52 that are present in 24 and 47 copies, respectively. In addition, the region harbors single copies of other C/D box snoRNAs: HBII-13, HBII-436, HBII-437, HBII-438A and HBII-438B. Recent evidence suggests that the HBII-85 and HBII-52 snoRNA clusters are expressed as two transcriptional units(211). The highly conserved snoRNAs are flanked by poorly conserved non-coding exons, suggesting that the functional relevant products of the locus are snoRNAs, not the flanking exons. The expression of these snoRNAs is tissue-specific. HBII-52 could be detected only in brain, whereas other snoRNAs from the SNURF–SNRPN locus are also expressed in non-brain tissues (reviewed in (212)).

Here, we analyzed the function of the mouse ortholog of HBII-52, MBII-52. We found that it regulates alternative pre-mRNA processing of at least five more genes. The unit expressing MBII-52 expresses smaller RNAs that appear to be nuclease processing products of the full-length MBII-52 snoRNA. We termed these shorter RNAs psnoRNAs for processed snoRNAs. psnoRNAs associate with hnRNPs and not with the known C/D box snoRNA binding proteins. We postulate that psnoRNAs recognize target RNAs by sequence complementarity and influence splice-site selection by interfering with splicing regulatory proteins acting on pre-mRNA.

8.2 Results

8.2.1 New targets for MBII-52

The recent finding that HBII-52 regulates alternative splicing of the 5-HT_{2C} receptor (191) raised the question whether there are other targets for this snoRNA. The antisense boxes of the 47 human copies of HBII-52 show up to three sequence variations from their 18 nt consensus sequences(212). We tested HBII-52 variants with one, two, three and five mutations in their antisense box for their ability to change alternative splicing of exon Vb of the serotonin receptor. We found that a snoRNA with three mismatches can still promote exon Vb inclusion (Fig. 8.1). There is no statistically significant change when five mismatches are present in the antisense box. This argues that naturally occurring HBII-52 variants with up to three mismatches between antisense box and target region can influence pre-mRNA processing of the serotonin receptor.

In order to uncover additional targets of HBII-52, we performed a computational screen. Because the mode of interaction between HBII-52 and its targets is not yet known, we based our analysis on the constraints on snoRNA:rRNA interactions leading to ribose methylation in ribosomal targets(193)). Concretely, we started by extracting an 18-nt-long antisense element upstream of the D box of MBII-52. We defined as a putative target site of MBII-52 a genomic region that can either form a perfect stem of length at least 10 bp or form a duplex of low free energy (below -15 kcal/mol) with the MBII-52 antisense element, with the duplex satisfying additional constraints. Minimum free energy duplexes were predicted with RNAhybrid(213) allowing G:U wobble in addition to canonical base pairing. The constraints on the duplexes were that (i) loops in the duplex were limited to maximum two nucleotides in either the target sequence or in antisense element and (ii) only up to three unpaired nucleotides in any of the sequences was allowed. Finally, similar to approaches previously employed to predict miRNA targets, we required that the predicted target site be conserved across mammalian species. More specifically, we extracted the regions in the human, rhesus macaque, cow and dog that are orthologous to the predicted HBII-52 target sites in human and we determined whether they would also be predicted as target sites. As the antisense box of HBII-52 is highly conserved in mammals, we compared all orthologous genes to the human antisense box sequence. Our final set of predictions included only putative MBII-52 target sites that were conserved in all of these other species. We obtained 457 such sites, 222 of which are in close proximity (200 nt) or within known exons. The predictions are available at <http://www.mirz.unibas.ch/MBII-52/>.

We next tested more than 100 computational predictions experimentally. Neuro2A cells were transfected with either MBII-52 or MBII-52mut, an MBII-52 variant with a scrambled antisense box(191), and the isolated RNA was analyzed by RT-PCR,

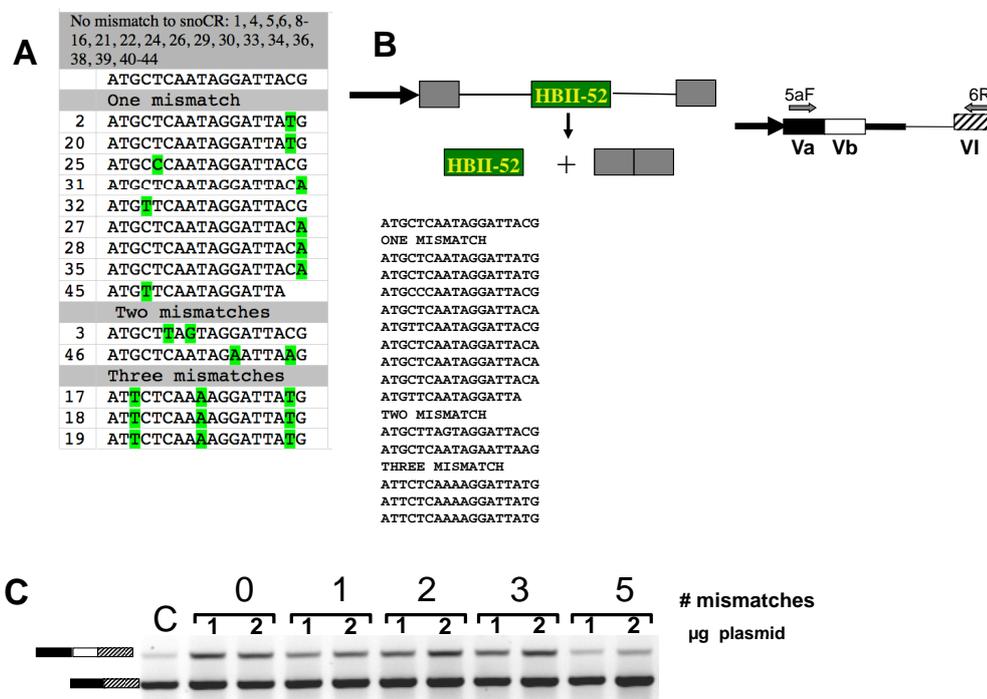


Figure 8.1: MBII-52 with three mismatches can still promote exon Vb inclusion

using primers in the flanking constitutive exons. As shown in Figure 1, we observed a change in alternative splicing patterns in the DPM2, TAF1, RALGPS1, PBRM1 and CRHR1 pre-mRNAs. MBII-52 overexpression promoted either inclusion or skipping of the different exons. Their sequences and the complementarity to MBII-52 are shown in Table 8.1. These data suggest that MBII-52 expression changes alternative splicing of several endogenous pre-mRNAs.

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

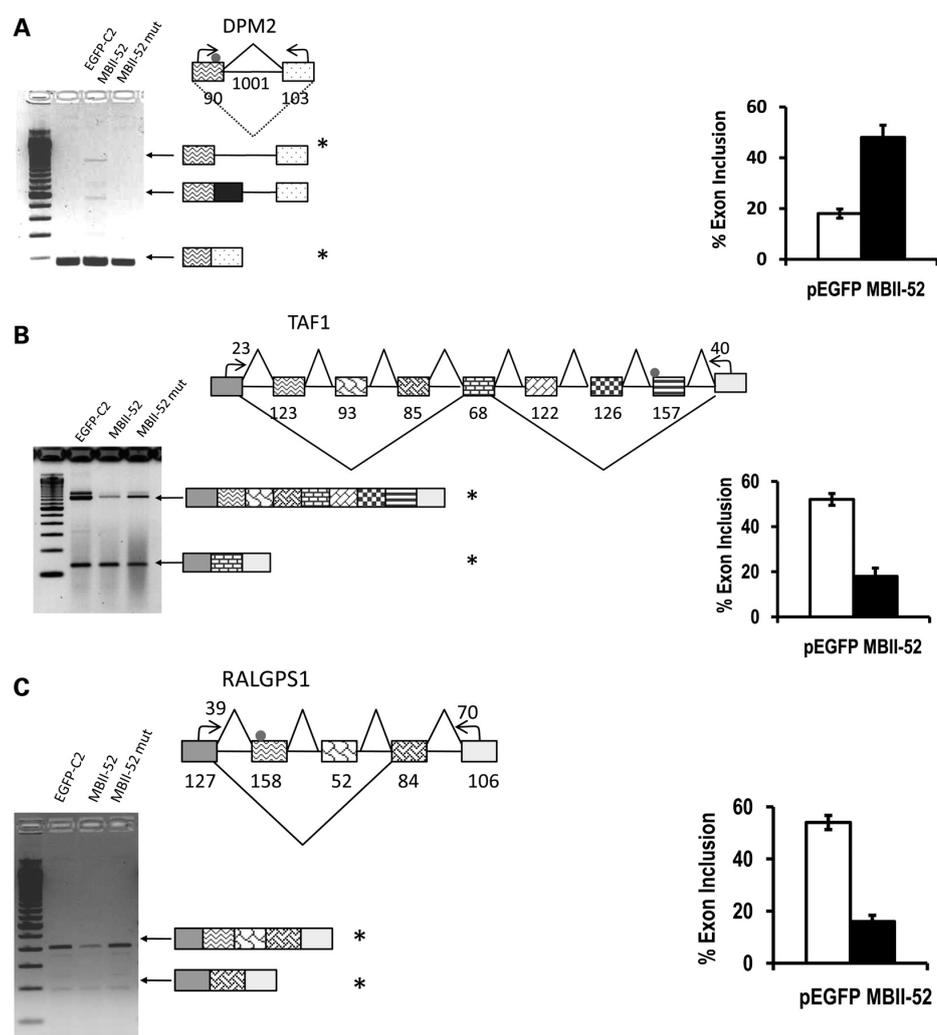


Figure 8.2: MBII-52 changes the alternative splicing pattern of predicted targets. Computationally predicted MBII-52 target genes expressed in Neuro2A cells were analyzed by RT-PCR. Cells were transfected with 1 μ g pEGFP-C2, 1 μ g of the MBII-52 expressing construct pCMV/MBII-52 (MBII-52)(214) and 1 μ g MBII-52 consensus box mutation, MBII-52 cm, (MBII-52 mut) expressing an antisense box mutation of MBII-52 (191). A representative ethidium bromide-stained agarose gel is shown. The adjacent diagram shows the part of the genes that was analyzed. Small arrows indicate the location of the primers used. The MBII-52 complementarity region is indicated by a dot. Numbers in boxes indicate the length of the exons and numbers next to PCR primers indicate the length of the amplified exon fragment. The structure of the PCR products is indicated by similar shading of exons in cDNA and genomic DNA. The statistical analysis of at least four independent experiments is shown on the right. Stars indicate the bands that were used for quantification. The sequences of the regulated exons are shown in Table 8.1.

Name (exon size)	Description	Exon sequence	Function	Complementarity
DPM2 (91 nt)	Dolichol phosphate-mannose biosynthesis regulatory protein	gagacctcccttttctccagggccacgggacagaccagggtggtgggactggcctgctgctggcgttagcctgatactcttcaccTACTACACCGCTGGGTGATTCCTTCgtatvgtcaatctccccgctccgctcaccctccccagcctccgaccgccagcgaactactatata	Frame shift	target 5' G C 3' GUCAUUCU UUG GUAU CAUUAGGA AAC CGUA snoRNA 3' G U U 5'
TAF1 (157 nt)	TAF1 RNA polymerase II, TBP-associated factor	tgacccccactggctctcattcagAAAGTCAAGATGGAGATGGTGATCTTGAGATAGAGAGAACTGTACAACAGCCTCAAGCCAGTCTCCTGTATGAGGA TTTGCTTATGTCTGAAGGA GAAGATGATGAGGAAGATGCTGGGAGTGTCAAGAAGGAGACAATCCTTTCTCTGtagggc	Unknown domain	target 5' A A G 3' GU GUCCUGU UCAG AU CA UAGGAUA ACUC UA snoRNA 3' G U G 5'
RALGPS1 (158 nt)	Ral guanine nucleotide exchange factor RalGPS1A	ctttccattttccagTATGATGTCTCAGTTCCAGTGTAGTTGAGTAAAGTCCGACATCCCATCGGAGAAAGCAAGGCACCTACTGGACGACAGTCTCTAGAGTCCCGCAGCCCCGGAAGGGCCTGGCTCTGACCTCCTCTCTGTGTCAACAAATGGACTCTCCTAGtaagcg	Frame shift	target 5' A A 3' UGU GUC GUUGAGUGU GCA UAG UAACUCGUA snoRNA 3' U GA 5'
PBRM1 (165 nt)	Polybromo 1	tgtttttaactagtcctgtgaaatgcaatggatggatgttttgaattcctgggttaaaaacacagaattgaaaatctgaaatgcccctttacagAGGGGACGCTAAAGTTGCTGAGCAGCAGGAGAGAGAGCGAGCA CAGCAACAGCAGCCGAGTCTCTCCCGGACGACCCCTGTGGGGCTCTCATGGGGTGGTGCACCCACAACACCAATGGGATGCTCAATCAGCAGTTGACACCTGTGCA Ggtaaaaacaggaagctaaag	Unknown domain	target 5' C AU A 3' UAGUCCUGUGA GCA AUUAGGAUAACU CGU snoRNA 3' GC A 5'
CRHR1 (120 nt)	Corticotropin hormone releasing receptor	ctgtgccttaccagccgctctctgccccagACTGTCAGTGCAACGCATCCGTTGGACCTCATTTGGCACTCTGTGGCCCGCAGCCCTGGGGGACGCTAGTGGTTGGCCCTGCTGCCCTGCTTTTCTATGGTGTCCGCTACAATACCACAagtaagga	Deletion of hormone binding domain	target 5' C A C 3' CGUGG CCU AUUG GCA GCAUU GGA UAAC CGU snoRNA 3' A U A 5'

Table 8.1: Genes that showed a dependency on MBII-52 expression both on endogenous and reporter gene level are listed using their HUGO nomenclature (columns 1 and 2). Numbers in parentheses indicate the exon length. The sequence of the regulated exon and its surrounding sequence is shown in column 3. Introns are in small letters, exons in capital letters. The snoRNA complementary region is highlighted in grey and underlined. Column 5 shows the alignment between the MBII-52 antisense box (snoRNA) and its target region.

8.2.2 MBII-52 changes alternative splicing of targeted pre-mRNAs in reporter gene assays

In the next step of the analysis, we determined whether alternative exons that are influenced by MBII-52 expression show this dependency also in a heterologous system, where they are surrounded by a different RNA context. We cloned the MBII-52 regulated exons into an exon-trap vector, where they were flanked by constitutively spliced insulin exons. All constructs were cloned into pSpliceExpress, a system that we developed previously(215).

The reporter genes were cotransfected with MBII-52 expression constructs into Neuro2A cells and the splicing patterns were analyzed by RT-PCR. As shown in Figure 8.3, we observed for the five splicing events identified in endogenous genes a similar dependency on MBII-52 expression. The expression of MBII-85 snoRNA and C and D box mutants of MBII-52 (MBII-52cC,cD) did not show a significant effect on the alternative exons, suggesting that the effect is specific for MBII-52. With the exception of PBRM1, the reporter minigenes followed the splicing pattern of the endogenous genes. In the endogenous PBRM1 gene, MBII-52 promoted both inclusion and skipping of two exons located in a cluster of alternatively spliced cassette exons. In the heterologous system, we observe only the skipping event for PBRM1. This difference is most likely due to the presence of strong insulin exons in pSpliceExpress that interfere with the arrangement of regulatory sequences in this cluster of multiple alternative cassette exons. Finally, we created a series of compensatory mutations in the antisense box of MBII-52 and the snoRNA complementarity regions (snoCR) of its targets. These experiments proved inconclusive, as in most cases mutating the snoCR resulted in strong exon activation that was no longer susceptible to regulation (data not shown). Together, these data suggest that after being transferred into a heterologous gene context at least five alternative exons are influenced by MBII-52 expression.

8.2.3 A mouse model of PWS shows changes in the predicted exons

To address the physiological significance of our data, we asked whether MBII-52 influences alternative splicing of the identified target genes in vivo and analyzed RNA samples from the TgPWS mouse model. TgPWS mice have a paternally derived deletion of the PWS critical region that contains the SNURF-SNRPN locus. They show hormonal and metabolic defects resembling those of human newborns with PWS(209). As a larger locus is deleted, in addition to MBII-52, the mice do not express MBII-85 and other snoRNAs from the Prader-Willi critical region.

We compared RNA from newborn TgPWS mice with RNA from littermates expressing the region. As shown in Figure 8.4, we found that the mouse knockout

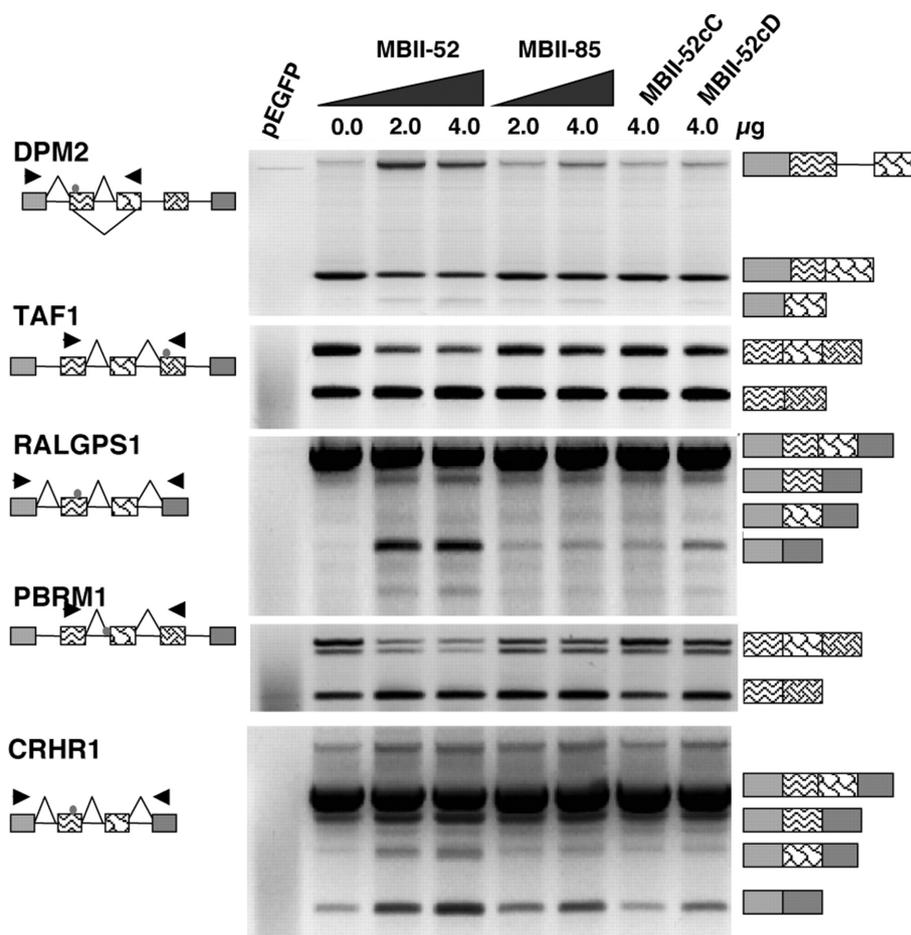


Figure 8.3: Minigene analysis of MBII-52 target genes. The exons harboring the MBII-52 complementary region were subcloned into the exon trap vector pSplice-Express. The structure of the resulting constructs pSE-RALGPS1, pSE-CRHR1, pSE-DPM2, pSE-PBRM1 and pSE-TAF1 as well as the location of the primers used for RT-PCR analysis is indicated on the left. pEGFP: only an expression construct for GFP is transfected. All other lanes contain 1 μg of pSE-reporter. MBII-52: cotransfection with 2 and 4 μg of MBII-52 expression construct, MBII-85: cotransfection with 2 and 4 μg of an MBII-85 expression construct, MBII52cC: cotransfection with 4 μg of a C-box mutant of MBII-52; MBII52cD: cotransfection with 4 μg of a D-box mutant of MBII-52. The structure of the products is shown schematically on the right, using the same shading scheme as in Figure 1. The usage of alternative exons indicated with a triangle was statistically evaluated. The comparison between MBII-52 and MBII-85 transfected cells showed statistically significant differences, the P-values of the Student's t-test were: DPM2: 0.001, TAF1: 0.023; RALGPS: 0.021; PBRM1: 0.076 and CRHR1: 0.002; ($n = 4$).

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

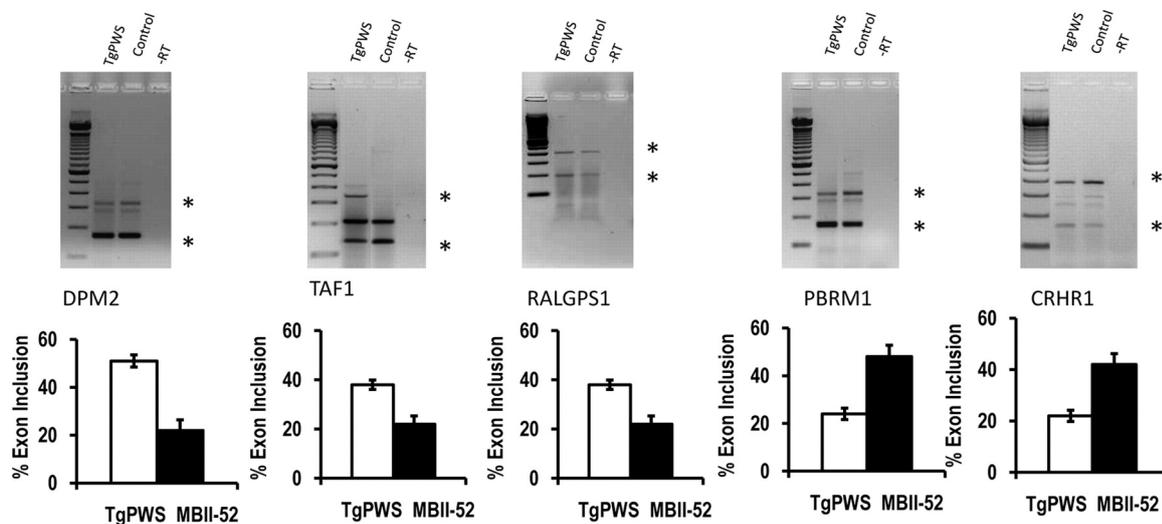


Figure 8.4: Comparison of RNA from TgPWS mice lacking MBII-52 expression and control littermates. Total brain samples from TgPWS mice lacking expression of the Prader-Willi critical region were compared with normal littermates expressing all the snoRNAs from the PWS critical region (control). Primers similar to Figure 8.2 were used. The structure of the gene products is indicated similar to Figures 8.2 and 8.3. Stars indicate the bands that were used for comparison. $n = 6$, other statistical evaluations: $P = 0.093$, <0.001 , 0.0023 , 0.001 , 0.05 for DPM2, TAF1, RALGPS1, PBRM1 and CRHR1, respectively.

systems recapitulates a dependency of alternative splicing on the presence of MBII-52. However, the overall splicing patterns of the endogenous genes are different in mouse brain and Neuro2A cells. This most likely reflects the presence of numerous cell types in brain that show different splicing patterns. Despite this limitation, the presence of MBII-52 promotes exon inclusion in the alternative exons with a complementarity to MBII-52 of the DPM2 and PBRM1 pre-mRNAs and promotes skipping of the RALGPS1 and TAF1 exons, similar to the effect seen in Neuro2A cells. The only discrepancy between the MBII-52 effects in brain and Neuro2A cells was an alternative exon of CRHR1 that showed an increase in exon usage in brain tissue, whereas it showed a decrease in response to MBII-52 in Neuro2A cells. The regulated alternative CRHR1 exon is in a cluster of alternative exons and the discrepancy could be due to differences in splicing regulators between brain and Neuro2A cells. Collectively, the data suggest that the loss of MBII-52 expression influences alternative splicing of target genes in a physiological context.

8.2.4 MBII-52 is processed into smaller RNAs

The data indicate that MBII-52 expression influences usage of multiple exons that contain regions with sequence complementarity to the antisense-box of MBII-52. Four recent studies reported that H/ACA snoRNAs give rise to smaller RNAs(216; 217; 218) and Chapter 9. We therefore tested whether the C/D box snoRNA MBII-52 also gives rise to other RNAs by RNase protection analysis.

Whereas humans have 47 HBII-52 copies, there are at least 130 copies of MBII-52 snoRNAs in mouse. We used an antisense probe against the MBII-52 copy employed in transfection experiments described above. This isoform is 87 nt in length and its sequence is shown in Figure 8.5D as form A. *In silico* analysis shows that this copy shares only an uninterrupted stretch of 20 nt in the antisense box region with other snoRNA isoforms of the MBII-52 cluster. All other regions show single nucleotide differences that prevent longer protected fragments. For the analysis, we used an *in vitro* transcribed, ^{32}P labeled RNA-antisense probe that detects the 87 nt encompassing the full-length snoRNA. Together with linker and vector sequences, the probe is 175 nt in length. After hybridization, RNase A and T1 digestion, the fragments were separated on 15% acrylamide/TBE/8 m urea gels. As shown in Figure 8.5A, lane 1, we observed additional fragments when total mouse brain RNA was analyzed with this probe. In agreement with earlier studies, we do not detect expression in liver(214)(Fig 8.5A, lane 9). We then asked whether the snoRNA expression construct used in Figure 8.2 is processed in a similar way. We analyzed total RNA from Neuro2A cells transfected with the pCMV/MBII-52 expression construct (Fig. 8.5A, lane 2) and found a similar RNA pattern. Importantly, the most abundant RNA species from both the expression construct and brain is shorter than 80 nt (form B, Fig. 8.5A). SnoRNAs contain C and D boxes that stabilize the snoRNP. Mutation of these RNA elements abolished the effect on splicing (Fig. 8.3). We therefore tested expression from constructs expressing this mutants and could not detect any RNA expression (Fig. 8.5A, lanes 6 and 7), suggesting that the smaller RNAs (form B, C, D) derive from a precursor with intact C and D boxes.

It is possible that MBII-52 undergoes nucleotide modifications that would result in mismatching of an RNase protection probe and subsequent generation of smaller fragments. To rule out this possibility, we performed northern blot analysis, using denaturing 15% PAGE gels. Total RNA from brain, liver and spleen was probed with MBII-52 antisense RNA corresponding to the sequence in Figure 8.5D, form A. Even after stringent washing, we see cross-hybridization of MBII-52 with RNAs from liver, spleen and HEK293 cells (Fig. 8.5B). This is to be expected, as there are numerous copies of sequence-related snoRNAs expressed in these tissues(214). To detect the specific hybridization between MBII-52 form A and brain RNA, we treated the membrane with RNase A and RNase T1. The RNase treatment reduced the overall signal strength, as we had to use a 3-fold longer exposure time. As shown in

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

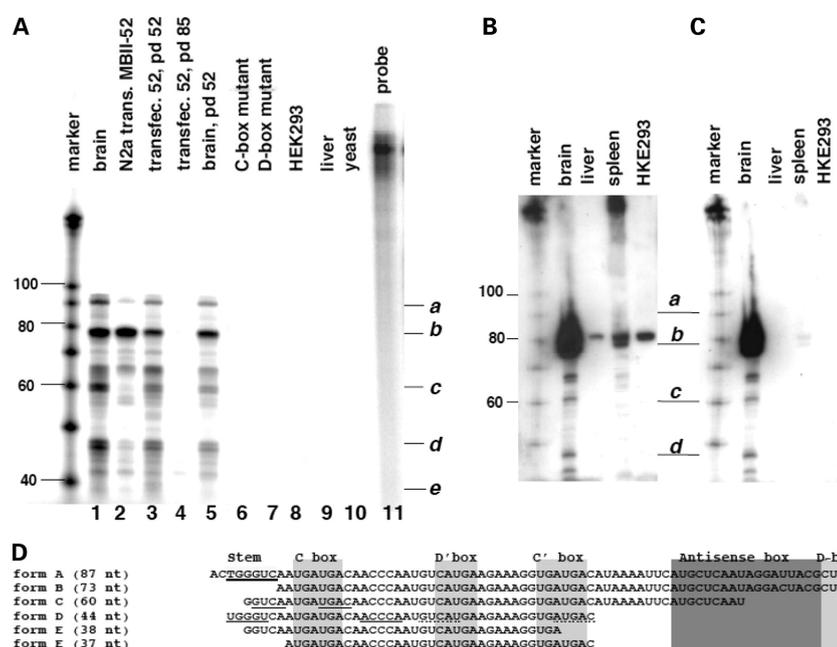


Figure 8.5: MBII-52 is processed into smaller RNAs. (A) RNase protection analysis using a probe detecting the MBII-52 copy used in transfection studies in Figures 8.2 and 8.3. Five microgram of the following total RNAs was hybridized to an MBII-52 antisense probe: (1): total mouse brain, (2): Neuro2A cells transfected with pCMV/MBII-52. Lanes (3–5) are protections from RNPs captured with oligonucleotides against the antisense box (Fig. 8.6). (3): Affinity captured RNA from Neuro2A cells expressing MBII-52 using a MBII-52 probe for pull down (pd), (4): affinity captured RNA from Neuro2A expressing MBII-52 using an MBII-85 probe (negative control) and (5): affinity captured RNA from brain using an MBII-52 probe for pull down. Lane 6: RNA from Neuro2A cells transfected with an expression construct containing a C-box mutant, (7): RNA from Neuro2A cells transfected with an expression construct containing a D-box mutant, (8): HEK293 cells non-transfected, (9, 10): RNA from liver and yeast. (11): Undigested probe. The marker is a 100 nt RNA base ladder. (B) Northern blot analysis of MBII-52. Fifteen microgram total RNA from brain, liver, spleen and HEK293 cells was separated on 15% polyacrylamide gels and probed with a ^{32}P labeled probe for MBII-52. After stringent washing, cross-hybridizing bands in liver, spleen and HEK293 cells still remain. Exposure was overnight. (C) The filter from (B) was treated with RNase A/T1 and again washed. The cross-hybridizing bands disappear, but the signals corresponding to smaller bands remain. Exposure was for 3 days. (D) Sequences of the shorter RNAs. The stems and functional boxes are indicated. The clones are ordered according to their length. Form A corresponds to the published snoRNA MBII-52. Underlined nucleotides in forms C and D indicated predicted stems.

Figure 8.5C, this treatment abolishes the cross-hybridization with non-brain RNAs. However, this treatment does not abolish the signal from brain RNA tissue that corresponds in length to RNA forms B–D. Similar to the RNA protection experiment, the major RNA species is shorter than 80 nt. This indicates that the protection pattern observed in the protection assay is due to shorter RNAs and not the result of nucleotide editing. Unexpectedly, we observe a distinct pattern of shorter RNAs and not a continuous smear of bands. This finding implies that all of the estimated MBII-52 copies are processed in a similar way, giving rise to specific metabolically stable short RNAs.

To determine the identity of the novel short RNAs, we cloned the protected fragments. Total mouse brain RNA was subjected to RNase protection. Subsequently, the RNases were removed by Proteinase K treatment and phenol extraction. The double-stranded RNA was phosphorylated using T4 kinase, and an adenylated-linker was ligated in the absence of ATP(219). After gel purification and isolation, an adapter linker was ligated using T4 DNA ligase. The reaction was subsequently reverse transcribed, amplified and cloned. The positive clones are shown in Figure 8.5D. All shorter RNAs lack the sequences forming the stem of the snoRNA, but contain the C and C' box. The stem conveys complementarity between the snoRNA ends and stabilizes the snoRNP. The remaining cloned RNAs are shortened from the 5' and 3' ends, indicating that they are generated by 3'→5' and 5'→3' exonuclease activity that stops at the C and C' boxes.

Together, these data suggest that the expression unit consisting of MBII-52 and its flanking intron and exons gives rise to several RNAs. These RNAs include the previously described MBII-52 snoRNA (form A), as well as shorter RNA species. The major RNA species (form B) expressed from the MBII-52 cluster lacks the stem box, but still contain C and D boxes.

8.2.5 MBII-52 derived RNAs do not bind to classical snoRNA-associated proteins

As we found that the MBII-52 locus gives rise to previously not described products, we identified the proteins that associate with these RNAs. We used the affinity between a biotinylated 2'-O-methylated oligonucleotide and the antisense box of MBII-52 to isolate RNAs derived from the MBII-52 locus (Fig. 8.6A). Using streptavidin beads, we isolated the MBII-52 snoRNA particle (snoRNP) from nuclear extracts generated from cells transfected with the MBII-52 expression construct. Nuclear extract was generated by a scaled-down Dignam procedure(220). After washing with 100 and 200 mM NaCl, the captured material was separated by SDS-PAGE and proteins were identified by mass spectrometry and database matching. An oligonucleotide against the snoCR of MBII-85 was used as the control. As shown in Figure 8.6B, we found

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

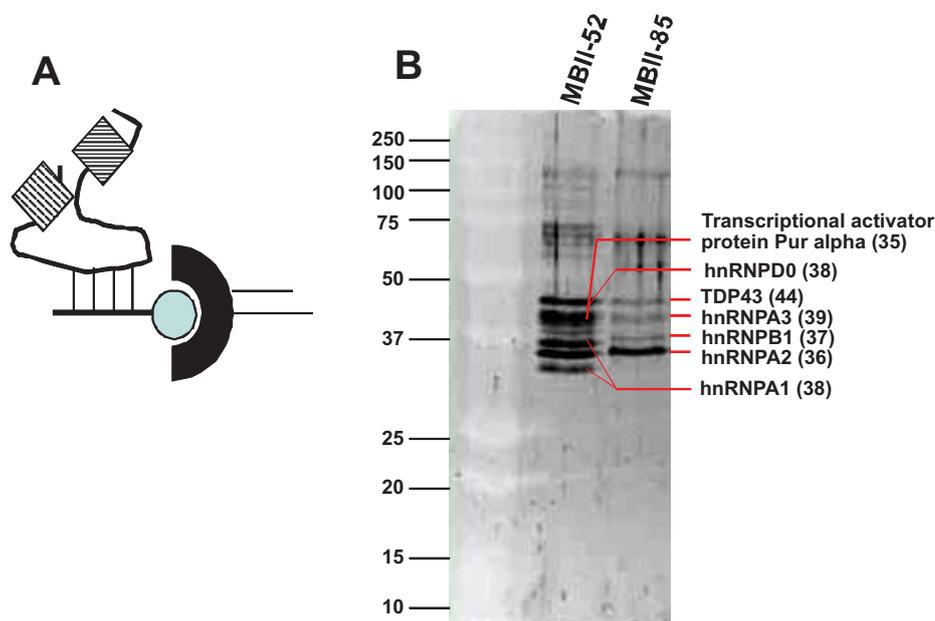


Figure 8.6: Analysis of proteins associated with MBII-52. (A) Experimental strategy: a biotinylated oligonucleotide is used to capture the snoRNP complex. The oligonucleotide is shown as a line, the complementarity is dashed, biotin is shown as a circle. The RNP complex (boxes) is isolated by streptavidin (half-circle)-capture from extracts expressing MBII-52 and washed in non-denaturing buffer. The extracts were prepared by transfecting MBII-52 expression constructs and performing Dignam mini-extracts. (B) Silver stain of a gel with affinity purified RNPs. MBII-52: the RNPs were isolated with a capture-oligonucleotide against MBII-52. MBII-85: the RNP was isolated with a capture oligonucleotide against MBII-85. Proteins were determined by mass spectrometry and are indicated. Sizes in kDa are shown in parentheses.

that hnRNPs were associated with the expressed snoRNA. Similar results were seen with samples obtained from mouse brain nuclear extracts (data not shown). Repeated experiments using different washing and isolation methods to find canonical snoRNP proteins, such as fibrillarin or NOP56, in pulled-down material from MBII-52 affinity material failed to identify known snoRNP-associated proteins.

We determined which RNAs are present in the pulled-down material and performed RNase protection. As shown in Figure 8.5A, lane 4 and 5, we found that the isolates contained the smaller MBII-52-related RNAs, as well as the full-length MBII-52 snoRNA. No MBII-52 RNA was pulled-down with the probe against MBII-85, suggesting the selectivity of the pull-down.

In summary, the findings indicate that the shorter RNAs assemble with hnRNPs,

but not with proteins that have previously been described to associate with C/D box snoRNAs. Although the major RNA isoform B contains C and D boxes, structural hallmarks of C/D box snoRNAs, the composition of the RNP formed is different from a C/D box snoRNP.

8.3 Discussion

8.3.1 The MBII-52 expression unit generates processed snoRNAs (psnoRNAs)

MBII-52 snoRNAs are expressed from a cluster containing multiple copies of tandemly arranged snoRNA expression units. Each unit contains phylogenetically poorly conserved exons that flank an intron which hosts the snoRNA(210). Humans contain 47 HBII-52 copies and mice at least 130 copies. Using RNase protection assays, we analyzed the mouse MBII-52 copy that is most closely related to the copy 27 of human HBII-52 snoRNA cluster. There is enough sequence heterogeneity between the different MBII-52 snoRNA copies that allows their discrimination in protection assays. Unexpectedly, the RNase protection assay indicates that the snoRNA gives rise to other smaller RNAs and that the full-length C/D box snoRNA is a minor form. The presence of the smaller RNAs could be verified by northern blot analysis, which further rules out that signals corresponding to shorter RNAs are caused by the protection of unrelated RNAs or are caused by RNA editing events that introduce mismatches to the probe. Finally, we tested ectopical expression of MBII-52 in HEK293 cells that do not express this snoRNA. The expression construct gives a similar pattern of shorter RNAs, indicating that they are derived from the transfected single MBII-52 expressing unit. The cloning of the shorter RNAs indicates that the major RNA form expressed from the MBII-52 expression unit is a 73 nt long RNA (form B) that lacks the sequences that form the snoRNA stem. However, this RNA contains other C/D box snoRNA elements, such as the C box, D box and antisense box. This RNA appears to be further shortened by exonuclease trimming, giving rise to smaller RNAs. The shorter RNAs can be detected both by northern blot and RNase protection analyses, indicating that they are metabolically stable. It is possible that these RNAs are protected from further endonuclease action by predicted secondary structures. The RNA form D forms a 5 bp stem on its 5' and 3' ends and RNA form C contains a short stem at its 5' end (Fig. 8.5D, underlined region). In addition, the formation of protein complexes is likely to stabilize the RNAs.

Ectopic expression of snoRNA mutants suggests that the formation of shorter RNAs depends on intact C and D boxes, which suggests that they derive from a C/D box snoRNA or pre-snoRNA structure. A possible scenario is that an unknown RNase initially removes the stem of the C/D box RNA, which gives rise to the predominant

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

form B. This form is stabilized by the presence of the C and D boxes, most likely by binding to other proteins. Form B is shortened by exonucleases, giving rise to forms C, D and E that are most likely stabilized by another stem-loop structure and/or associated proteins.

To obtain insight into proteins associated with these novel RNAs, we isolated them by affinity purification of RNP complexes, using a probe that is complementary to the antisense box of MBII-52. We used nuclear extract generated by the Dignam procedure as starting material. In this method, most of the nucleolar material is separated in a high-speed centrifugation step. As the MBII-52-derived snoRNAs are present in this fraction, they are most likely present in the nucleoplasm. The major form RNA form B derived from MBII-52 does not contain the characteristic k-turn, which most likely prevents its association to Snu13p/15.5 kDa(197). In agreement with this RNA structure, we could not detect C/D box snoRNA-associated proteins, such as fibrillarin, or NOP56(192) in the isolated material. In contrast, we identified hnRNPs, including hnRNP A1, A2, TDP-43 and D0 that have been reported to be involved in splice-site selection. Unexpectedly, in the pull-downed material, we could still detect RNA forms C and D. These RNAs lack a complete snoCR that is complementary to the pull-down probe. Relative to the starting brain material, the RNA forms C and D are reduced in the pulled down material (compare Fig. 8.5A, lanes 1 and 5), but are still detectable. This suggests that the different RNA forms could be present within a complex.

We propose to name these shorter RNAs psnoRNAs for processed small nucleolar RNAs. PsnoRNAs could represent a new class of nuclear small RNAs. The psnoRNAs described here are the first to be derived from C/D box snoRNAs.

8.3.2 MBII-52-derived psnoRNAs regulate splicing of several pre-mRNAs

We previously found that the expression of the snoRNA HBII-52 promotes inclusion of exon Vb of the serotonin receptor 5-HT_{2C}. To investigate whether this represents a special, unique case or is part of a new regulatory mechanism, we developed a computational screen that predicted more than 400 putative snoRNA targets. We tested some of these predicted targets by RT-PCR in transfection assays and further concentrated on five splicing events that showed consistent dependency on MBII-52 expression. In contrast to the 5-HT_{2C} receptor pre-mRNA, the pre-mRNAs harboring the MBII-52-dependent exons are expressed in Neuro2A and HEK293 cells, which allows us to determine the influence of MBII-52 expression on the endogenous genes. Also in contrast to the neuron-specific 5-HT_{2C} system where a splice site had to be optimized to detect the dependency on MBII-52(191), the new alternative exons showed the dependency on MBII-52 expression when analyzed in their endogenous

HBII-52	A	U	G	C	U	C	A	A	U	A	G	G	A	U	U	A	C	G
5HTC	+	+	+	+	+	+	+	+	E	+	+	+	+	E	E	+	+	+
PB1	L	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	M	M
DPM2	+	+	+	+	L	+	+	+	M	+	+	+	+	+	+	+	M	M
RALGPS	+	+	+	+	+	+	+	+	+	L	M	+	+	+	L	+	+	+
TAF1	+	+	M	+	+	+	+	+	+	+	+	+	+	+	L	+	+	M
CRHR1		+	+	+	G	+	+	+	+	+	+	+	M	+	+	+	+	+
CONSENSUS	4	6	5	6	4	6	6	6	4	5	5	6	5	6/5	3	6	4	3

Table 8.2: Complementarity between MBII-52 antisense box and its experimentally confirmed targets.

gene context.

The alternative exons were next tested in a heterologous exon trap system and showed the dependency of MBII-52 when flanked by insulin exons that are controlled by a CMV promoter. These experiments suggest that MBII-52 RNAs act on defined parts of the pre-mRNA, in a mechanism that is independent of the promoter usage and genomic context. Together, these data strongly suggest that MBII-52 expression influences alternative pre-mRNA splicing events.

Expression of MBII-52 causes a small, but statistically significant changes in multiple targets. This modest influence on numerous targets has been observed for other splicing factors, such as SMN(221) and NOVA(222). Detailed work in the NOVA system(222) suggested that a splicing factor can control biological processes by coordinating numerous small changes and it is possible that MBII-52 fulfills a similar function. An alignment of the antisense box of MBII-52 and its experimentally confirmed targets is shown in Table 8.2. The complementarity between the MBII-52 antisense box and its targets can be interrupted in multiple positions. With the exception of the serotonin receptor 5HT2C, there are always three mismatches in the alignment of the snoCR and the MBII-52 antisense box. It is interesting that the serotonin pre-mRNA can be edited at three positions within the snoCR. Taking these editing events into account, the data suggest that preferably 15 of the 18 nucleotides of the antisense box show complementarity towards its target. It is noteworthy that we initially concentrated on targets with only one or two mismatches, but did not find a dependency of these exons on MBII-52 expression. The data indicate that MBII-52-derived RNAs need a defined degree of sequence complementarity towards their targets. This scenario is reminiscent of the action of U1 snRNP on the 5' splice site, where natural occurring exons rarely show 100% complementarity towards the U1 snRNA, but usually have several mismatches, which cluster in certain position of the 5' splice site(223).

The existence of psnoRNAs could explain the influence of MBII-52 expression on splice-site selection in a model illustrated in Figure 8.7. We postulate that the MBII-

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

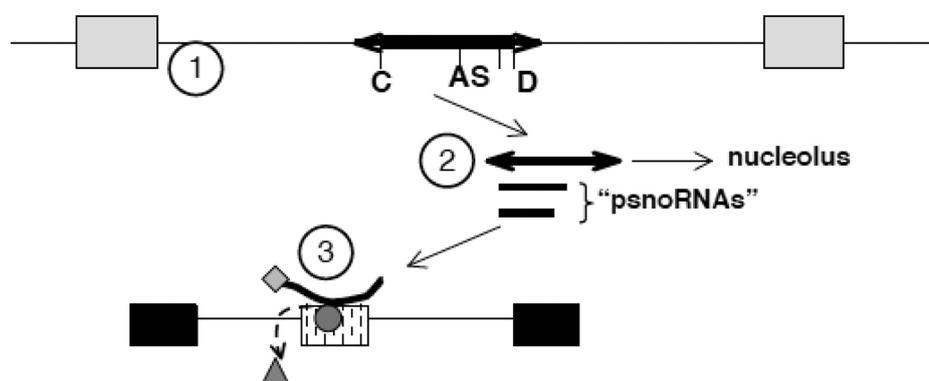


Figure 8.7: Model for MBII-52 action on RNA processing. (1) The PWS critical region contains snoRNAs (thick line) located in introns between non-coding exons (grey boxes). The snoRNA is characterized by a C box (C), D box (D) and an antisense box (AS), as well as stem-forming sequences (arrows). (2) This unit generates several RNAs, including the full-length snoRNA that shows its highest concentration in the nucleolus and Cajal bodies (224) as well as several shorter psnoRNAs (for processed snoRNAs). PsnoRNAs are present in the nucleoplasm where they associate with hnRNPs. (3) PsnoRNAs can change splice-site selection, most likely by binding to complementary sequences. We postulate that they either remove regulatory proteins from their targets (triangle) or bring in associated proteins to the exon recognition complex (diamond associated with the small RNA).

52 expressing unit consisting of two non-coding exons flanking an intron that hosts an snoRNA gives rise to several RNAs. The major form derived from the expression unit is form B that lacks the snoRNA stem-structure and is associated with hnRNPs, but not C/D box snoRNA-typical proteins. Form B contains the antisense box that targets it to other RNAs, including pre-mRNAs identified in this study. Form B RNA can influence splice-site selection by competing with existing splicing regulatory factors on the pre-mRNA or by bringing the associated hnRNPs to the targets, similar to a bifunctional oligonucleotide. The longest RNA (form A) has all the hallmarks of a traditional C/D box snoRNA, but is only a minor fraction of the RNA expressed. It is likely that this RNA is transported into the nucleolus, where it can be detected by *in situ* hybridization (224). It is not clear what function this RNA has in the nucleolus, but it could represent a storage form for the formation of the active RNA form B that is released from the nucleolus according to the physiological needs.

8.3.3 Relevance for PWS

The loss of C/D box snoRNA expression has been postulated as the underlying mechanism for the development of PWS(225). This hypothesis was recently supported by a patient with a 174 584 bp large microdeletion that encompassed only snoRNAs and their flanking hosting introns and exons. The deletion led to a Prader–Willi phenotype (201). To date, the only published RNAs expressed from the 174 584 bp region are snoRNAs and fragments of their surrounding non-coding exons.

The idea that the loss of snoRNA expression is central to PWS is further supported by genetic evidence that ruled out proteins encoded by MKRN3, MAGEL2 and NDN genes expressed in the Prader–Willi critical region (226). The 174 584 bp microdeletion removes the snoRNAs HBII-438A, -85 and 23 of the 47 HBII-52 copies from the paternally expressed allele. The only snoRNA that was totally removed by the microdeletion was MBII-85, which led to the suggestion that MBII-85 loss is the major reason for PWS. However, there is evidence that HBII-85 and HBII-52 are expressed by two transcriptional units (211). As the 174 584 bp microdeletion contains the 5' end of the HBII-52 cluster, it could harbor transcriptional elements necessary for proper HBII-52 expression. Furthermore, in the majority of cases, the complete SNURF–SNRPN locus is deleted (200). The contribution of HBII-85 and HBII-52 loss to PWS is therefore not clear.

Our findings indicate that the SNURF–SNRPN locus not only gives rise to typical C/D box snoRNAs, but generates shorter psnoRNAs. The northern blot analysis indicates that all of the at least 130 MBII-52 copies are processed in a similar manner. The major RNA form from the MBII-52 cluster is not the canonical C/D box snoRNA, but a shorter RNA form, most likely similar to psnoRNA form B. psnoRNAs are associated with hnRNPs and could have multiple functions by targeting these proteins to other RNAs. It is not clear whether several psnoRNAs lacking the antisense box use other RNA parts for targeting or have non-related functions.

The loss of the regulatory psnoRNAs could be a significant contribution to the etiology of PWS and substitution of the short psnoRNAs could be a therapeutic principle for the disease.

8.4 Materials and methods

Transfection experiments were performed using Ca-phosphate method as described (227).

The construction of reporter minigenes was done by using recombination between pSpliceExpress, an exon-trap vector and BAC-derived PCR fragments that encompassed the alternative exons, as previously described (215).

Pull-down experiments were performed using Dignam-derived miniextracts (220).

The snoRNA MBII-52 is processed into smaller RNAs and regulates...

RNase protection analysis was performed using the Ambion RNase protection kit using uniformly ^{32}P -labeled probes.

Cloning of the psnoRNAs was performed as follows: 100 μg of total brain RNA, isolated by the Trizol method was protected using $3 \cdot 10^6$ cpm of an MBII-52 antisense probe. Hybridization was overnight. Single-stranded RNA was digested with RNase A/T1 (Ambion, dilution 1:100) for 1 h at 37°C and RNases were subsequently removed by 100 $\mu\text{g}/\text{ml}$ Proteinase K treatment for 1 h. Following phenol extraction and precipitation, RNAs were separated on a 15% acrylamide, 8 m urea gel, exposed overnight and the appropriate bands were excised, crushed, eluted overnight in 3 m ammonium acetate/1% SDS and recovered by precipitation. The first RNA linker was 5'rAppCTGTAGGCACCATCAAT/3ddC. It was ligated for 2 h in a 20 μl reaction in 50 mm HEPES pH 8.3, 10 mm MgCl_2 , 3.3 mm DTT, 10 $\mu\text{g}/\text{ml}$ BSA, 8.3 v/v glycerol and 20 U RNA ligase. The reaction was again separated by a 15% acrylamide, 8 m urea gel, bands excised, crushed and eluted overnight. The second RNA linker was 5'-AmMC6/GCTCCAGAATTCGGACCCGArGrUrGrCrCrUrArCrArG. It was ligated at 18°C in $1 \times$ ligase buffer using T4 DNA ligase overnight. The reaction was reverse transcribed, PCR amplified and subcloned. Positive clones were isolated using colony hybridization and sequenced.

Primers for PCR detection were:

CRHR1:

MmNEWCRHR1F CCAGGATCAGCAGTGTGAGA;

MmNEWCRHR1R AGTGGCCCAGGTAGTTGATG;

TAF1:

TAF1NewF TCTGCGATGAAAACTCAAAGA;

TAF1NewR TCCACATCAGAGTCACTTCCA;

DPM2:

F CAGACCAAGCAGTAGGATTT;

R ACAAACAGGAGCAGCAGGAG;

RALGPS1:

F AGTCCCCAGACACAGGAAGA;

R TCTCAGAGGCCCTCCAT;

PB1:

F TGGCTACATTTTGTTCAGCAG;

R ATGGGGGCTACTCCTTGATT.

Chapter 9

Expression and Processing of a Small Nucleolar RNA from the Epstein-Barr Virus Genome

Roland Hutzinger, Regina Feederle, Jan Mrazek, Natalia Schiefermeier, Piotr J. Balwierz, Mihaela Zavolan, Norbert Polacek, Henri-Jacques Delecluse, Alexander Hüttenhofer

PLoS Pathogens 5(8): e1000547 2009

Small nucleolar RNAs (snoRNAs) are localized within the nucleolus, a sub-nuclear compartment, in which they guide ribosomal or spliceosomal RNA modifications, respectively. Up until now, snoRNAs have only been identified in eukaryal and archaeal genomes, but are notably absent in bacteria. By screening B lymphocytes for expression of non-coding RNAs (ncRNAs) induced by the Epstein-Barr virus (EBV), we here report, for the first time, the identification of a snoRNA gene within a viral genome, designated as v-snoRNA1. This genetic element displays all hallmark sequence motifs of a canonical C/D box snoRNA, namely C/C'- as well as D/D'-boxes. The nucleolar localization of v-snoRNA1 was verified by *in situ* hybridisation of EBV-infected cells. We also confirmed binding of the three canonical snoRNA proteins, fibrillarin, Nop56 and Nop58, to v-snoRNA1. The C-box motif of v-snoRNA1 was shown to be crucial for the stability of the viral snoRNA; its selective deletion in the viral genome led to a complete down-regulation of v-snoRNA1 expression levels within EBV-infected B cells. We further provide evidence that v-snoRNA1 might serve as a miRNA-like precursor, which is processed into 24 nt sized RNA species, designated as v-snoRNA1^{24pp}. A potential target site of v-snoRNA1^{24pp} was identified within the 3'-UTR of BALF5

mRNA which encodes the viral DNA polymerase. v-snoRNA1 was found to be expressed in all investigated EBV-positive cell lines, including lymphoblastoid cell lines (LCL). Interestingly, induction of the lytic cycle markedly up-regulated expression levels of v-snoRNA1 up to 30-fold. By a computational approach, we identified a v-snoRNA1 homolog in the rhesus lymphocryptovirus genome. This evolutionary conservation suggests an important role of v-snoRNA1 during γ -herpesvirus infection.

9.1 Introduction

The Epstein-Barr virus (EBV), a member of the γ -herpesvirus subfamily, possesses a large (170 to 180 kb) double-stranded DNA genome. EBV infection is etiologically linked with various cancers of the lymphoid and epithelial lineages that include Burkitt's lymphoma (BL), Hodgkin's disease, nasopharyngeal carcinoma (NPC) and post-transplant lymphoproliferate disease (PTLD)(228) (229)(230)(231) *In vitro and in vivo*, EBV transforms normal B cells through establishment of a type III latency during which a restricted set of viral genes is expressed (eight Epstein-Barr nuclear antigens and two latent membrane proteins)(232). More restricted expression patterns such as latency type II in NPC and latency type I in BL have also been characterized. In fact, recent work on Burkitt's lymphoma has shown that a subset of these tumours display a latency pattern intermediate between latency I and III showing that the boundaries between the latency types are not always sharply established as initially thought (233).

More than two decades ago, the group of J. Steitz discovered two highly abundant 170-nt long non-coding RNAs (ncRNAs) in the EBV genome, designated as Epstein-Barr encoded RNAs (EBER1 and EBER2) (234). EBER RNAs have subsequently been shown to bind to human ribosomal protein L22. However, no unequivocal biological functions could be assigned to EBER transcripts, up till now (235). The list of non-coding RNAs encoded by EBV has since rapidly expanded with the recent discovery of 25 microRNAs (miRNAs) (236; 237; 238; 239; 240; 241).

In addition to miRNAs, numerous other ncRNAs have been discovered in all three domains of life, i.e. Archaea, Bacteria and Eukarya, as well as in various viruses (242; 243). A large number of these ncRNA species were found to be involved in multiple regulatory functions including cellular differentiation and development, chromatin architecture, transcription and translation, alternative splicing, RNA editing, virulence and stress responses (191; 244; 245; 246).

Small nucleolar RNAs (snoRNAs) consist of more than 200 stable ncRNA species in Eukarya of about 60 to 300 nt in size which are located in a sub-nuclear compartment, the nucleolus (192; 247). SnoRNAs guide nucleotide modifications within ribosomal RNAs (rRNAs) or spliceosomal RNAs (snRNAs), i.e. 2'-O-ribose methy-

lation or pseudouridylation, respectively. The snoRNA class has been identified in Archaea and Eukarya, but not in Bacteria, and is subdivided into box C/D and box H/ACA snoRNAs. In Eukarya, the majority of snoRNAs is located within introns of protein-coding genes and is processed by splicing followed by endo- and exonucleolytic cleavage (246; 248; 249).

Each member of the box C/D snoRNA family possesses characteristic sequence elements called box C (PuUGAUGA) and box D (CUGA), optional degenerate C'/D' boxes and a short 5'-3' terminal stem structure (193; 249). 10–21 nt long sequence-specific antisense elements upstream of the boxes D/D' guide the box C/D snoRNA core proteins fibrillarin, a RNA methyltransferase, Nop56, Nop58 and the 15.5 kD protein to the target RNA. 2'-O-methylation of the ribose at the fifth nucleotide upstream of the D/D' box on the target RNA is carried out by the fibrillarin core protein (249). Box H/ACA snoRNAs possess a distinctive common ACA sequence motif at their 3'-terminus and one to two stem-loop structures linked by a hinge (the so-called H-box motif: ANANNA, with N being any nucleotide), and guide the conversion of uridine to pseudouridine within the RNA target (250; 251). The large number of conserved modifications in functionally conserved regions of rRNAs, such as the peptidyl-transferase centre, has suggested an important role for rRNA modifications in fine-tuning the structure and/or function of rRNAs (252). It is important to note that a significant number of so-called “orphan” snoRNAs, lacking rRNA or snRNA targets, have been identified in Eukarya (253; 254). However, the biological functions of orphan snoRNAs are still elusive.

In this study, we report, for the first time, the identification of a functional C/D box snoRNA within the EBV genome. We demonstrate that this viral snoRNA exhibits all bona fide box C/D snoRNA features with respect to its processing and expression, nucleolar localization as well as to canonical core protein binding partners. We also provide evidence that v-snoRNA1 is processed into a 24 nt long miRNA-like species which might target the 3'-UTR of the viral DNA polymerase mRNA.

9.2 Results

9.2.1 Identification of v-snoRNA1 by cDNA cloning and expression analysis

We have established an experimental strategy, designated as SHORT, to identify viral-induced ncRNAs in cord blood lymphocytes (CBL) infected with the EBV strain B95.8 (255). The SHORT method is based on subtractive hybridisation of ncRNA populations of virus-infected cells from non-infected cells. NcRNAs, selectively expressed in the infected cell population, were subsequently converted into cDNAs. Sequencing of a small number, i.e. about 500 cDNA clones, allowed identification of

Expression and processing of a small nucleolar RNA...

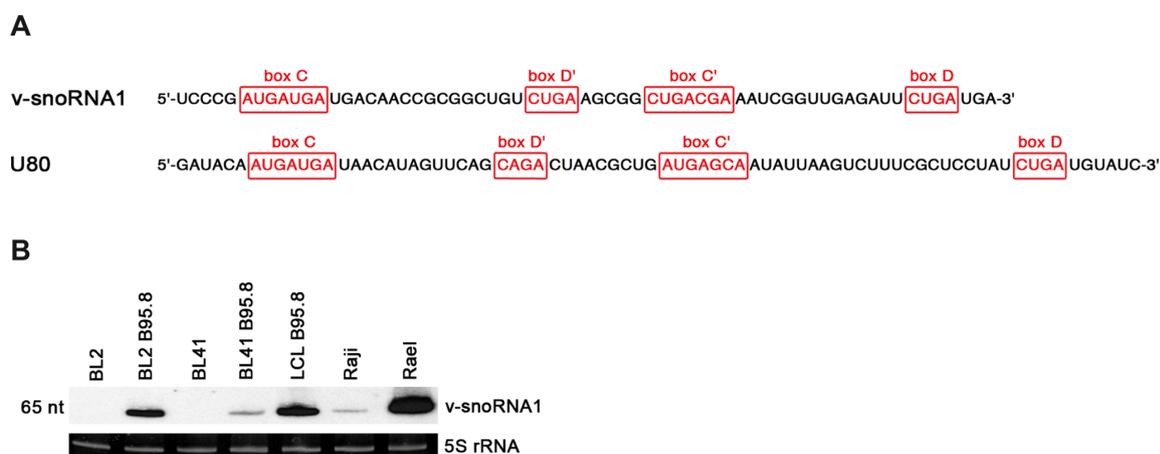


Figure 9.1: Sequence and expression profile of v-snoRNA1.

(A) Sequences of the newly identified 65 nt long v-snoRNA1 and the canonical box C/D from snoRNA U80 used as a control are shown here. The position of C/D boxes and C'/D' boxes are indicated in red. (B) Northern blot analysis showing expression of v-snoRNA1 in a panel of EBV-positive (BL2-B95.8, BL41-B95.8, LCL-B95.8, Raji, Rael) and EBV-negative (BL2, BL41) cell lines. 5S rRNA served as internal loading control.

several ncRNAs from the human as well as from the EBV genome whose expression was up-regulated upon viral infection (256).

Deep-sequencing analysis of 40'000 cDNA clones from this subtracted cDNA library further extended the list of differentially expressed ncRNAs (257). Interestingly, one of these sequences was represented by 95 cDNAs and exhibited all defining features of canonical C/D box snoRNA sequence motifs, i.e. C, C', D' and D boxes (193; 249). Crucially, this potentially novel snoRNA species mapped to the EBV genome and was therefore designated as v-snoRNA1 (Fig. 9.1A, and see above Accession number FN376861). It is noteworthy that the canonical terminal stem-structure, formed by the 5' and 3' ends of eukaryal snoRNAs, was absent in the viral snoRNA, a feature shared with snoRNAs identified from archaeal or fungal species (258; 259). To assess expression of v-snoRNA1, northern blot analysis was performed employing RNA from EBV-positive cell lines (Rael, Raji, BL2-B95.8, BL41-B95.8 and a LCL generated *in vitro* with the B95.8 virus) or EBV-negative cell lines (BL2 and BL41; 9.1B). As expected, v-snoRNA1 could only be detected in infected cells but not in the EBV-negative control cells. Comparison with an internal RNA marker showed that the hybridized RNA species was 65 nt in size, which fully matched the size suggested by the original sequence obtained by cDNA cloning (see above and Fig. 9.1B). Repeated attempts to identify v-snoRNA1-precursor transcripts by northern blot analysis were unsuccessful (unpublished data), suggesting that they are subjected

to rapid processing.

The v-snoRNA1 gene is located within the BamHI A rightward transcripts, known as BARTs, on the sense strand of the viral genome and maps about 100 nt downstream of the EBV mir-BART2 (Fig. 9.2A and 9.2B). The BARTs represent abundant RNA species in EBV that are expressed in all latently infected EBV-B cell lines, in peripheral blood B cells of EBV-positive individuals and, at higher levels, in nasopharyngeal carcinoma (260; 261). They do not encode for proteins but are processed into 22 different BART miRNAs (Fig. 9.2A) (241). Thereby, v-snoRNA1 as well as mir-BART2 arise from the same intron, which was found to be 4.9 kb in size in the AG876 strain (Accession number AJ507799) (260).

BART transcripts were previously shown to be predominantly transcribed from the P1 promoter (261). However, P2 promoter-initiated BARTs were also detected in different B-cell lines with the exception of the EBV-positive BL cell line Raji. As shown in Fig 9.1B, v-snoRNA1 expression was verified in all tested EBV cell lines, including Raji cells, although expression levels varied considerably. In particular, v-snoRNA1 was expressed in Raji cells at barely detectable levels. Therefore, we infer that v-snoRNA1 transcription can be initiated at the P1 promoter but that the P2 promoter might be required to obtain full expression.

9.2.2 Co-Immunoprecipitation and FISH analysis of v-snoRNA1

To determine the sub-cellular location of v-snoRNA1 within EBV-infected cells, we employed fluorescent *in situ* hybridization (FISH) with dye-labeled antisense oligonucleotides complementary to v-snoRNA1. As a control, we also investigated the localization of U3 snoRNA, which is known to be localized in the nucleolus (262; 263). Examination of EBV-infected BL2 cells by confocal microscopy revealed that both v-snoRNA1 and U3 snoRNA in fact co-localized to the nucleolus (Fig. 9.3A). In contrast, a v-snoRNA1 hybridization signal was absent in non-infected B cells.

Canonical C/D box snoRNAs have previously been shown to bind to four snoRNA core proteins: fibrillarin, Nop56, Nop58, and the 15.5 K protein, respectively. These proteins have previously been shown to be strictly required for RNA maturation, stabilization and function (192; 264). The C/D box proteins assemble with snoRNAs thus forming ribonucleo-protein complexes (snoRNPs) that localize to the nucleolus. In order to assess whether v-snoRNA1 assembles into a canonical C/D box snoRNP, binding of v-snoRNA1 to three of these canonical snoRNA-binding proteins (fibrillarin, Nop56 and Nop58) was assessed by co-immunoprecipitation using specific antibodies. Immuno-precipitated samples were subsequently analyzed for the presence of v-snoRNA1 by northern blot analysis. These assays demonstrated that v-snoRNA1 and the canonical U81 snoRNA, used as a positive control, were both co-immunoprecipitated with similar efficiencies with antibodies against all three snoRNA-binding protein (Fig. 9.3B). In contrast, none of the snoRNAs was pre-

Expression and processing of a small nucleolar RNA...

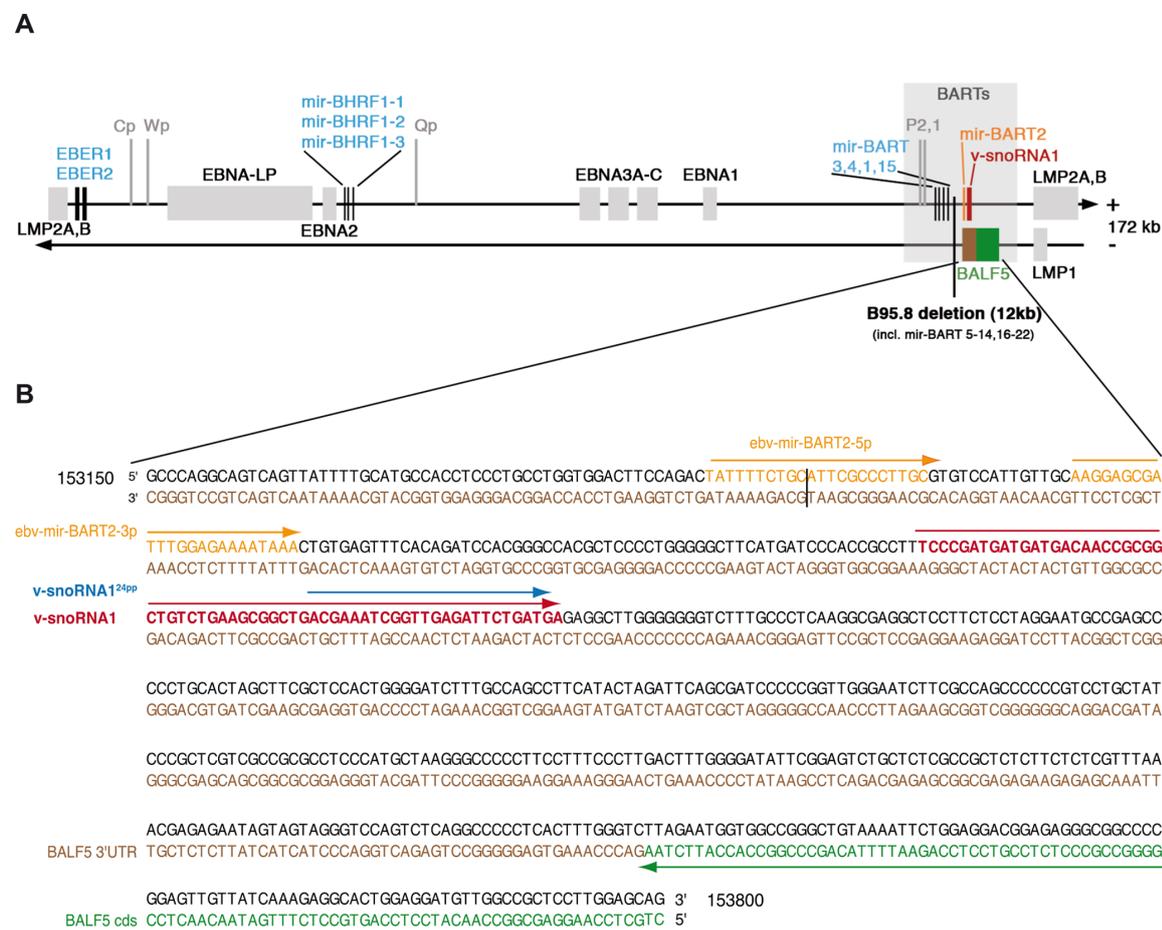


Figure 9.2: Schematic representation of the Epstein-Barr-virus genome. The location of ncRNA genes, latent genes and the precise location of v-snoRNA1 is indicated. (A) Location and transcription of EBV ncRNA genes (black lines with blue lettering) and EBV latent genes (grey bars with black lettering). The v-snoRNA1 is indicated in red, the neighboring miRNA BART2 in orange and the viral DNA polymerase BALF5 is depicted in green (for coding region) and brown (for 3'-UTR). The promoters are shown in grey lines and lettering, the BARTs region as a grey bar and the B95.8 deletion are also indicated. (B) Close-up of v-snoRNA1 location within the 3'-UTR of the viral DNA polymerase gene. The v-snoRNA1 is located on the sense strand about 60 nt downstream of the mir-BART2 precursor transcript and complementary to the BALF5 3'UTR that is situated on the antisense strand. v-snoRNA1^{24pp} is indicated in blue, other transcripts are indicated in the same colors as described above. The black line illustrates the cleavage site of mir-BART2. Corresponding EBV coordinates refer to the EBV B95.8 deletion strain (Accession number V01555.2).

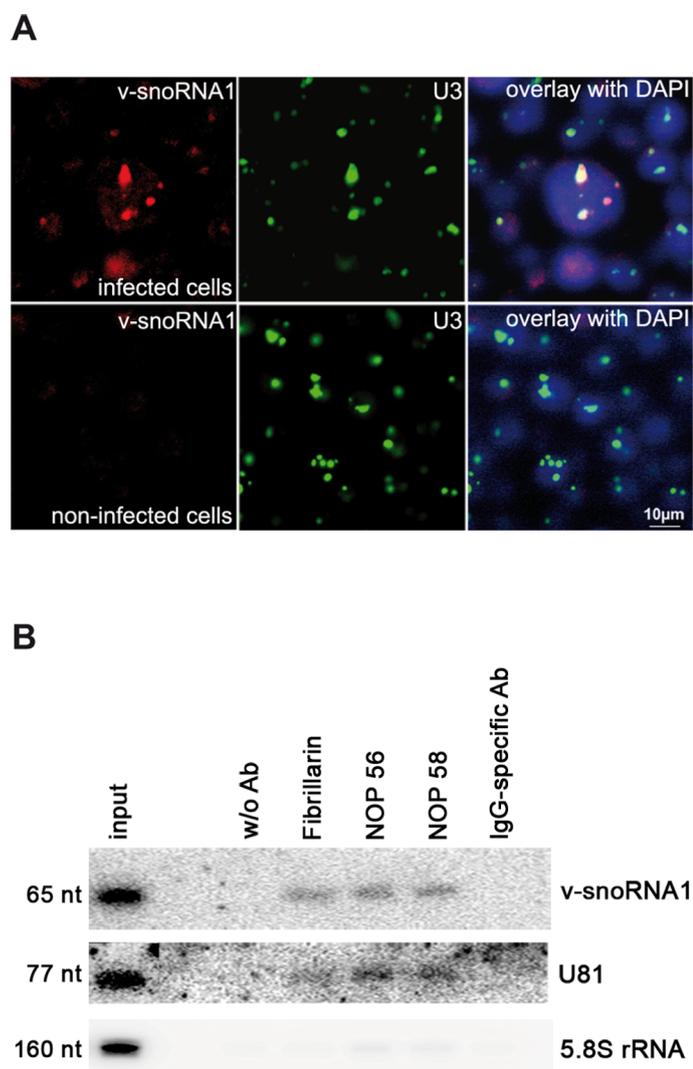


Figure 9.3: Fluorescent in situ hybridization and Co-immunoprecipitation of v-snoRNA1.

(A) The box C/D v-snoRNA1 (red) localizes in the nucleolus of EBV-positive BL2-B95.8 cells. Box C/D snoRNA U3 (green) was used as a nucleolar marker. In EBV-infected cells both v-snoRNA1 and U3 co-localize in the nucleoli. In EBV-negative cells only U3 is expressed. The nucleus was stained with DAPI for visualization of nuclei and the scale bar is 10 μm . (B) Co-immunoprecipitation of v-snoRNA1 with fibrillarin, NOP56 and NOP58 snoRNP proteins. Following immunoprecipitation employing antibodies specific to fibrillarin, NOP56 and NOP58, the v-snoRNA1 was co-precipitated and detected via northern blot analysis. Box C/D snoRNA U81 and 5.8 rRNA were used as positive and negative controls, respectively.

precipitated in controls without antibodies or employing an IgG-specific antibody. Hybridization with an oligonucleotide specific for 5.8S rRNA was used to test for the specificity of the employed antibodies. Thereby, a faint, unspecific signal was detected in all samples after antibody addition, except the control without an antibody. This is likely caused by the high expression levels of 5.8S rRNA in our samples. From these results we conclude that the newly identified 65 nt long viral RNA transcript displays all hallmark features of a genuine box C/D snoRNA.

9.2.3 v-snoRNA1 expression is strongly stimulated in the lytic cycle

A common trait shared by all herpesviruses is their ability to infect their target cells under several modes; cells can support lytic replication during which new virus progeny is replicated or instead induce virus latency. Viral proteins used in both modes are usually, but not always, distinct. We therefore assayed v-snoRNA1 expression in latently infected cells or in cells undergoing lytic replication. We took advantage of LCLs established with viruses that are devoid of the lytic immediate early gene *BZLF1* (Δ BZLF1) and therefore cannot initiate lytic replication (265) and examined v-snoRNA1 expression in these cells by northern blot analysis (Fig. 9.4). Northern blot signals were clearly visible in these cells thereby demonstrating that v-snoRNA1 is a latent transcript. We then performed the same experiment with replication-competent 293/EBV-wt cells lytically induced by transfection of the *BZLF1* gene (Fig. 9.4). Comparison with non-induced cells showed that the v-snoRNA1 expression levels were up-regulated up to 30-fold following induction (Fig. 9.4). v-snoRNA1 is therefore especially part of the EBV lytic expression program.

9.2.4 Phenotypic traits of a recombinant virus lacking v-snoRNA1

In an attempt to discover the function of v-snoRNA1 during the EBV life cycle, we constructed a recombinant virus that lacks a functional v-snoRNA1. To this aim, the C-box motif of v-snoRNA1 from the B95.8 strain was exchanged against the sequence of the kanamycin resistance gene flanked by two FLP recombinase recognition sites (Fig. 9.5A). Excision of this cassette left an unrelated bacterial sequence containing a HindIII restriction site in place of the box C of v-snoRNA1 (Fig. 9.5A and 9.5B, lane 2). DNA from the recombinant virus was stably transfected into 293 cells to generate a virus producer cell line, here referred to as 293/ Δ v-snoRNA1. Multiple clones were screened for their ability to support virus replication. One of the replication-competent clones was chosen at random for further experiments. Recombinant episomes purified from this producer cell line and transformed into *E. coli* cells were found to be intact as assessed by restriction

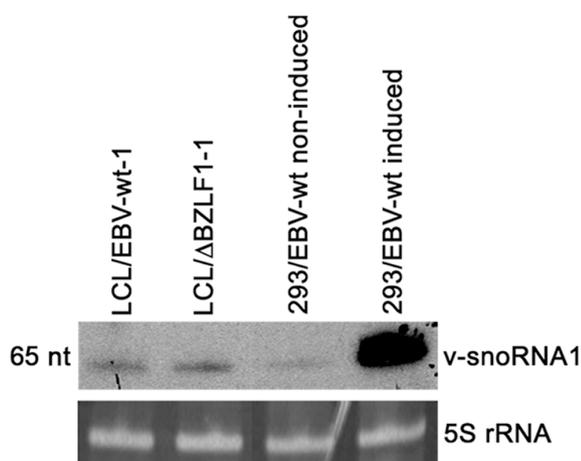


Figure 9.4: Expression of v-snoRNA1 during latency and lytic replication. Expression of v-snoRNA1 was investigated in LCLs infected with either the wild type or the replication-defective Δ BZLF1 EBV strain. The expression of v-snoRNA1 in 293 cells that stably carry the EBV-wt genome was monitored before and after induction with a BZLF1 expression plasmid. 5S rRNA was used as an internal loading control.

analysis (Fig. 9.5B, lane 3). Sequencing of the recombination site on these rescued episomes confirmed exchange of the Box C against unrelated DNA **TTTCCC-GCGCCAAGCTTCAAAGCGCTCTGAAGTTCCTATACTTTCTAGAGAATAG-GAACTTCGGAATAGGAACTTCCAACC** (EBV DNA around the insertion is indicated in bold). A northern blot, performed on 293/ Δ v-snoRNA1 cells using a v-snoRNA1-specific probe, yielded negative results while signals could be clearly identified in the 293/EBV-wt positive control (Fig. 9.5E, left panel). We therefore conclude that the Δ v-snoRNA1 virus is devoid of the viral snoRNA and that destruction of the putative C box of v-snoRNA1 is sufficient to exert this effect.

We then conducted a series of experiments aiming at defining phenotypic traits of the mutant strain. We first assessed the ability of the 293/ $\Delta\Delta$ v-snoRNA1 to support viral replication. Viral titres were quantified either as packaged viral genome-equivalents (physical titres) or as green Raji units, i.e. as the concentration of viruses able to infect the Raji cell line determined by exposure to a limiting dilution of the viral supernatants (functional titres). Both assays revealed nearly identical titres for both the mutant and the wild type control (Figure 5D). The Δ v-snoRNA1 viruses and producer cell line were then examined in electron microscopy; both displayed normal morphological features: encapsidation, primary and secondary egress were unchanged in the absence of the viral snoRNA (unpublished data). We further evaluated viral gene expression by western blot or immunostains (BZLF1, EA/D-BMRF1, gp350). Again, we could not discern any differences between the mutant and its wild type

Expression and processing of a small nucleolar RNA...

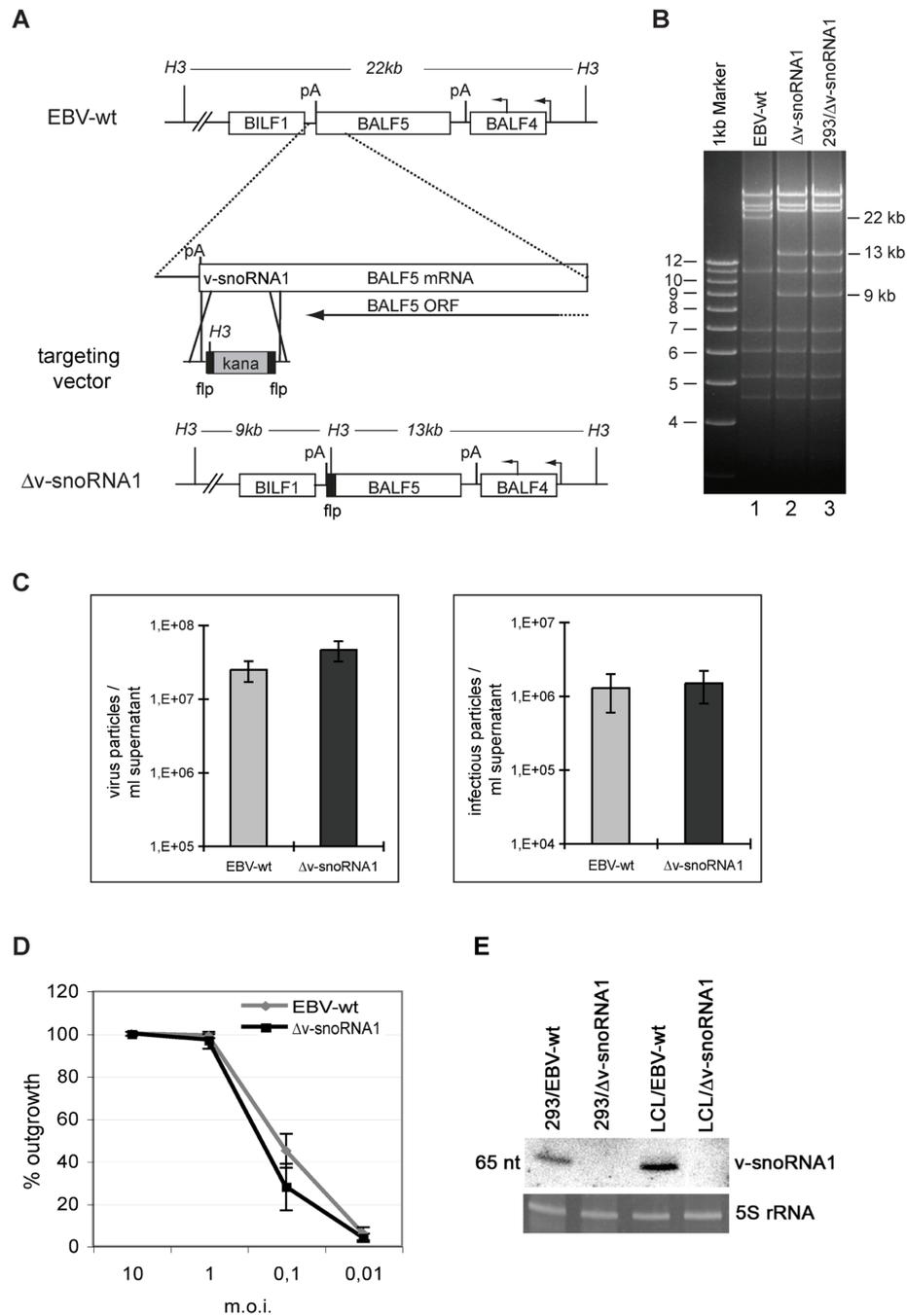


Figure 9.5: Construction of a v-snoRNA1 null recombinant virus. Continued on the next page ...

Figure 9.5: Construction of a v-snoRNA1 null recombinant virus (continued from the previous page)

(A) Schematic map of the EBV genome segment that encompasses the v-snoRNA1 in EBV-wt before and after homologous recombination with the targeting vector carrying the kanamycin resistance gene flanked by flp recombinase recognition sites. The kanamycin cassette was excised in a second step. The restriction sites for HindIII (H3) and the expected fragment sizes after cleavage of EBV-wt and Δ v-snoRNA1 genomes with this enzyme are given. pA: polyadenylation site, kana: kanamycin. **(B)** HindIII restriction fragment analysis of EBV-wt (lane 1) and Δ v-snoRNA1 mutant genomes directly after construction in *E. coli* (lane 2) or after rescue from stably transfected 293 cells (293/ Δ v-snoRNA1) (lane 3). The result is fully consistent with the predicted restriction pattern (see A). **(C)** v-snoRNA1 is not required for virus production. Titres in supernatants from cells induced to produce viruses were determined either by measuring the concentration of viral genomes or by infecting the Raji B cell line in a limiting dilution assay. The concentration of viral genome equivalents and infectious particles is given for wild type and Δ v-snoRNA1 viruses. Shown are mean values from three independent experiments. **(D)** Δ v-snoRNA1 viruses show intact transforming properties. Primary B cells from three different normal donors were exposed to wild type and Δ v-snoRNA1 viruses at various multiplicities of infection in a limiting dilution assay in cluster plates. The percentage of wells showing cell outgrowth is indicated. The presented results represent the average values from three experiments with the corresponding standard deviations. **(E)** v-snoRNA1 is expressed in cell lines infected with wild type EBV but not in cell lines infected with the Δ v-snoRNA1 null-mutant. A northern blot analysis using a v-snoRNA1-specific probe was performed on 293 and B cells infected with either wild type EBV or with the Δ v-snoRNA1-null mutant. 5S rRNA served as a loading control.

counterpart (unpublished data).

We then exposed various established cell lines or primary cells to the Δ v-snoRNA1 mutant and monitored the efficiency of infection by counting the percentage of GFP-positive (293 cell line, primary epithelial cells) or EBNA2-positive (primary B cells) lymphocytes three days post-infection. The rate of infection was nearly identical in both wild type and mutant viruses (unpublished data). We finally investigated the transforming capacity of the mutant by performing infections of normal resting B cells from three different normal individuals at decreasing multiplicity of infections (Fig. 9.5D). Wild type and mutant viruses both exhibited a transforming potential that resulted in a very similar number of outgrowing cell clones. We confirmed the identity of the viruses present in the growing LCLs by northern blot analysis; only LCLs generated by infection with wild type B95.8 virus expressed the snoRNA while those infected with Δ v-snoRNA1 remained negative (Fig. 9.5E, right panel).

9.2.5 Computational and functional analysis of v-snoRNA1

The majority of snoRNAs have been found to target rRNAs or snRNAs by guiding ribose methylation or pseudouridylation, respectively. In contrast, a number of snoRNAs lack telltale complementarities to canonical targets and hence are designated as “orphan” snoRNAs (246; 249; 254). We therefore examined 18S and 28S rRNAs for putative v-snoRNA1 target sites using criteria established by Cavaille and Bachellerie (193): the putative target sites were required to display at least a 7 nucleotides-long perfect complementarity with a region that ended within 3 nucleotides of the end of the snoRNA antisense boxes, and at most one nucleotide should be involved in a bulge or loop (193). In particular we searched for putative target sites of the v-snoRNA1 box D antisense elements and for two potential alternative box D' antisense elements (see Fig. 9.6A). Using a program that was successfully used to predict targets of bacterial ncRNAs (266) we identified two putative ribose methylation site within the 18S rRNA and 23 sites within the 28S rRNA for box D' (Tab. 9.1). However, none of the predicted target sites coincided with known methylated nucleotides within 18S and 28S rRNA. The same strategy applied to box D failed to reveal any putative ribose methylation sites within rRNAs. Nevertheless, we experimentally tested the ribose methylation status of the highest-scoring predictions for rRNA targets (Fig. 9.6B) by primer extension analysis (267; 268). However, no methylation at the predicted nucleotide positions C617 of human 18S rRNA and C3140 and C3152 of human 28S rRNA was observed in EBV-infected LCL B95.8 cells (data not shown), suggesting that v-snoRNA1 is a member of the still growing class of orphan snoRNAs.

9.2.6 Processing of v-snoRNA1 into v-snoRNA1^{24pp}: potential v-snoRNA1^{24pp} targets

In addition to full-length cDNA clones encoding v-snoRNA1, we also identified nine identical partial cDNA clones of 24 nt in size in our cDNA library derived from the very 3'-end of v-snoRNA1 (Fig. 9.2B). Previously, two studies were able to demonstrate processing of specific snoRNA species into functional miRNAs (216; 217). Attempts to verify expression of the 24 nt long v-snoRNA1-derived processing product, designated as v-snoRNA1^{24pp}, by northern blot analysis with conventional DNA oligonucleotide probes or by splinter ligation (217; 269) were unsuccessful (data not shown). In contrast, by applying a locked nucleic acid (LNA) probe, complementary to v-snoRNA1^{24pp}, we were able to verify its expression (Fig. 9.7). An additional hybridization signal at 40 nt was also observed that might represent a processing intermediate. All hybridization signals, except for full length v-snoRNA1, were only detected in the 293/EBV-wt cells induced with *BZLF1*, likely due to the high expression level of v-snoRNA1 within this strain. Notably, v-snoRNA1^{24pp} was not detected in the snoRNA knock-out strain (Fig. 9.7).

Since the 3'-UTR of the BALF5 mRNA exhibits full complementarity to v-snoRNA^{24pp} (Fig. 9.8) we investigated whether it might serve as a potential target site for cleavage by applying a 5'-RACE approach, as previously described (270; 271). 5'-RACE products from the predicted 3'-UTR cleavage site were amplified by specific primers and sequenced (Fig. 9.8). Indeed, we detected two clones corresponding exactly to a predicted cleavage site by v-snoRNA^{24pp} 11 nt from its 5'-end in 293/EBV-wt cells induced with *BZLF1* which exhibits highest expression levels of v-snoRNA1 (Fig. 9.4 and 9.7). Remaining clones from this region exhibited shorter sequences likely due to exonucleolytic degradation of the BALF5 mRNA following initial cleavage by v-snoRNA^{24pp} as described previously for plant miRNAs (270). Notably, not a single sequence was observed that was longer than the expected size, which is indicative of a specific cleavage event triggered by v-snoRNA^{24pp} and followed by exonucleolytic degradation. In contrast, no fragments cleaved within the 3'-UTR of BALF5 mRNA were observed in the snoRNA knock-out strain.

9.2.7 Conservation of v-snoRNA1 in other viral genomes

The identification of a snoRNA species in a viral genome raised two obvious questions: is v-snoRNA1 conserved among the different herpesvirus subfamilies or even among several EBV strains and do v-snoRNA1 homologs exist in other virus families? This prompted us to perform a BLAST alignment search using all available databases. This search showed that the v-snoRNA1 sequence is 100% conserved among the tested EBV strains (B95.8, AG876, M81, GD1, Raji). It further revealed that the distantly related rhesus lymphocryptovirus (rLCV) genome (exhibiting an overall sequence identity of 65% with the EBV genome; Accession number NC_006146) contains a 65 base pair sequence that shows 86% identity with v-snoRNA1 (Accession number FN376863). In particular, the canonical D, D' and C, C' boxes were universally conserved as well as antisense elements, preceding D or D' boxes. This high degree of sequence identity did not extend to the v-snoRNA1 flanking regions; these showed only 69% sequence identity and were therefore clearly less conserved (Fig. 9.9A). Northern blot analysis, employing an rLCV-specific antisense oligonucleotide, confirmed that the rLCV sequence homolog of v-snoRNA1 is actively transcribed and processed into an RNA species of 65 nt in simian B cells (Fig. 9.9B). Despite the high degree of sequence identity between human and rLCV v-snoRNA1s, hybridization with the rLCV-specific probe did not detect its EBV counterpart. Altogether, these findings strongly indicate that rLCV also encodes a box C/D snoRNA homolog to v-snoRNA1.

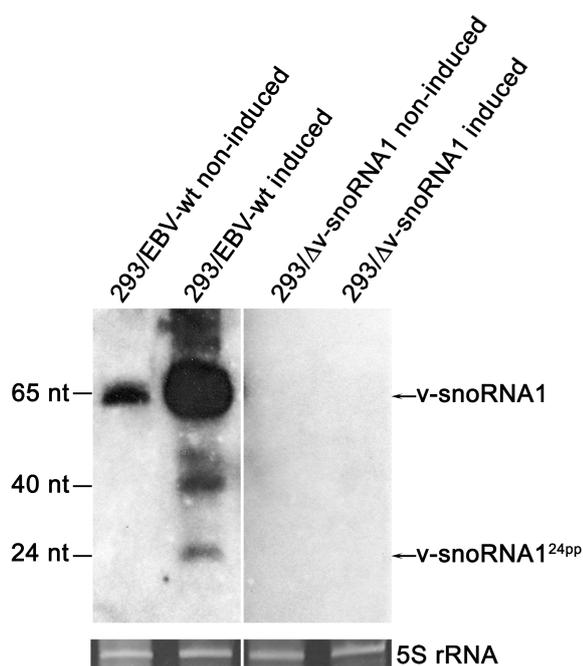


Figure 9.7: Expression analysis of v-snoRNA1^{24pp}.

Northern blot analysis demonstrating expression of the 24 nt long processing product v-snoRNA1^{24pp}, derived from v-snoRNA1, by employing a specific LNA oligonucleotide probe in 293/EBV-wt or in 293/Δv-snoRNA1 knock-out strain cells without or upon *BZLF1*-induction. Expression of full length v-snoRNA1 (65 nt) and a potential cleavage intermediate (40 nt) are also shown. 5S rRNA serves as an internal loading control.

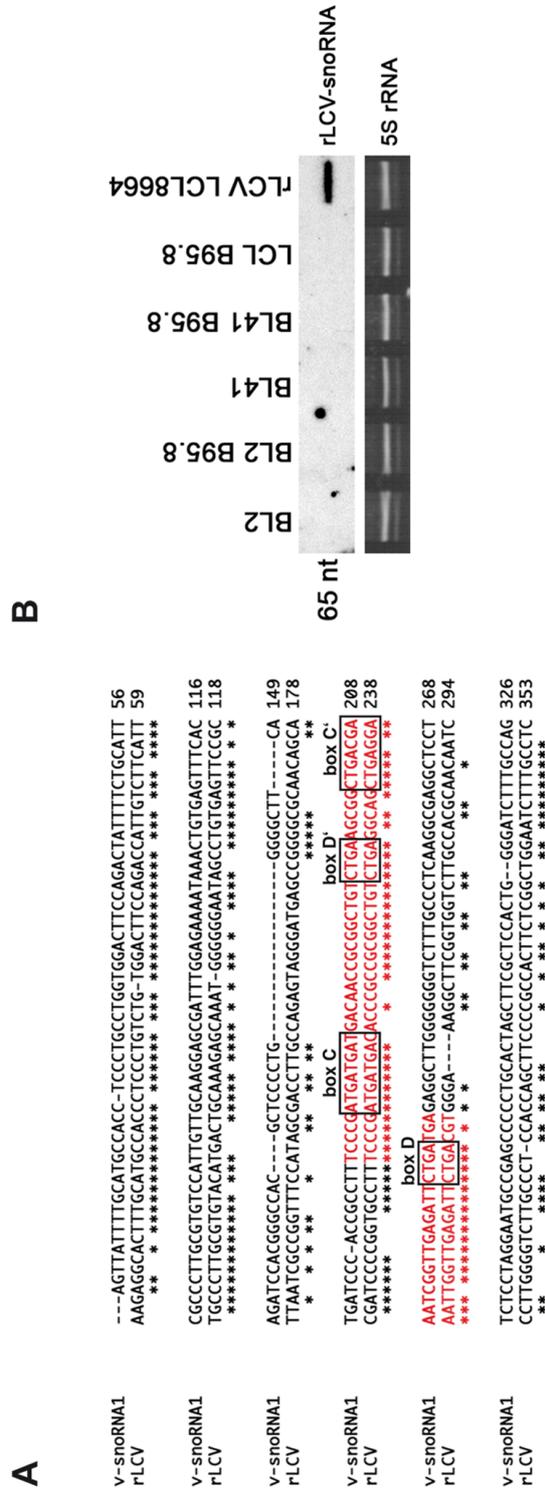


Figure 9: Expression of v-snoRNA1 in rLCV. (A) Alignment of v-snoRNA1 and rLCV including their flanking regions. v-snoRNA1 and rLCV snoRNA sequences are marked in red, flanking nucleotides in black. Stars in red and black indicate conservation of the nucleotides between EBV and rLCV sequences. The boxes C/D and C'/D' are encircled by black rectangles. (B) The putative rLCV snoRNA is expressed in simian LCLs. A northern blot analysis using a labeled rLCV snoRNA oligonucleotide was performed on the LCL8664 cell line that was generated by infection with rLCV. A panel of EBV-negative and EBV-positive human LCLs were used as controls. 5S rRNA was used as loading control.

9.3 Discussion

Herpes virus genomes carry numerous cellular gene homologs (272). Many of these genes encode house keeping proteins but others serve more specialized functions e.g. within the host immune system. This is particularly true of γ -herpesviruses whose genomes encode homologs of cytokines (e.g. CSF-1 and IL10 for EBV, IL6 for Kaposi's sarcoma-associated herpesvirus (KSHV) or of anti-apoptotic mediators (e.g. BCL2 in EBV and KHSV). These striking homologies between a virus and a cellular genome were reinforced by the discovery that herpesviruses encode multiple miRNA clusters. Here we report that herpesviruses and their host share yet another fundamental ncRNA species.

Deep-sequencing analysis of a subtracted cDNA library that was constructed to specifically identify transcripts expressed in EBV-infected B cells allowed discovery of a viral transcript that exhibited all defining features of a C/D box snoRNA. Indeed, v-snoRNA1 comprises canonical C/C' as well as D/D' boxes. It is of note that v-snoRNA1 is lacking the canonical terminal stem-structure usually encountered in eukaryal snoRNAs. In this respect, v-snoRNA1 appears to be closer to snoRNA species previously identified in fungi or in the domain of Archaea (258; 273). In addition to the EBV-encoded v-snoRNA1, the genome of the Herpesvirus saimiri (HVS), a member of the γ -herpesvirus family, was recently reported to encode seven small nuclear RNAs (274; 275). Thereby, in latently infected HVS-transformed T cells, the Herpesvirus saimiri U RNAs (HSURs) represent the most abundant viral transcripts. Similar to EBERs, HSURs are not essential for viral replication or transformation, but are involved in the activation of specific genes in virus-transformed T cells during latency (274).

v-snoRNA1 was found to be expressed in all samples of a panel of EBV-positive cell lines that included several BLs and in particular the latency I Rael cell line, LCLs and the 293/EBV-wt producer cell line (Fig. 9.1). Detection of reduced levels of v-snoRNA1 in LCLs, generated with the BZLF1-null virus that therefore cannot undergo lytic replication, demonstrated that v-snoRNA1 is an integral part of the EBV latent transcription program (Fig. 9.4). However, expression levels of v-snoRNA1 increased significantly up to 30-fold upon induction of the lytic replication cycle. This is consistent with a model that v-snoRNA1 serves, presumably different, functions in both the latent and the lytic mode of infection.

Three findings demonstrated that v-snoRNA1 is likely to represent a fully functional ncRNA species. v-snoRNA1 was found to co-localize with canonical snoRNA to the nucleolus (Fig. 9.3). Furthermore, we could show that v-snoRNA1 assembles into a canonical snoRNP that at least includes the fibrillarin, Nop56 and Nop58 proteins. Finally, selective destruction of the C box resulted in a complete down-regulation of steady state levels of v-snoRNA1 (Fig. 9.5E). This is consistent with previous work that ascribed an essential role in the regulation of the stability of snoRNA to this

sequence motif (247; 276; 277).

v-snoRNA1 could be localized to the BARTs region which follows a complex splicing pattern and also encodes a cluster of non-coding miRNA genes (Fig. 9.2). v-snoRNA1 was located outside the putative BARTs open reading frame and is therefore, as previously observed for canonical eukaryal snoRNAs, likely processed from an intron. The BARTs transcripts can be initiated from two promoters P1 and P2 (261). Analysis of v-snoRNA1 expression levels showed a large degree of variation within the tested cell lines, as was also observed for EBV's miRNAs (278). In principle, this could be related to the highly variable virus copy numbers among different EBV-positive cell lines. Alternatively, it may be related to the propensity of some of these cell lines to undergo lytic replication. The low expression levels of v-snoRNA1 in Raji are probably due to an inactive BART P2 promoter; this suggests that the P2 promoter initiates most of the v-snoRNA1 transcripts.

The discovery of a snoRNA in a Herpesvirus genome prompted us to search for homologs in other viral or cellular genomes. This search revealed that the v-snoRNA1 is strictly conserved across five distinct EBV strains. It further led to the identification of a transcript within the rLCV genome that displays a high degree of homology to v-snoRNA1. This genetic element comprises perfectly conserved canonical C/D and C'/D' boxes and was expressed in a simian LCL which suggests that rLCV also encodes a snoRNA. Discovery of a v-snoRNA1 homolog in rLCV is not entirely unexpected; rLCV is the closest EBV relative as both genomes exhibit 65% sequence identity and, therefore, display more than 80% sequence identity for protein-coding genes and ncRNA genes. Indeed, seven rLCV miRNA were found to be closely related to their EBV counterparts (238). The relatively crude approach (BLAST) we initially took failed to reveal further v-snoRNA1 relatives; we nevertheless consider that this question is still open and hope that our work will stimulate research in this direction.

The strict conservation of v-snoRNA1 domains within various EBV strains and among evolution strongly suggests that this element serves an essential role in the natural history of EBV infection. We therefore initiated a series of experiments that aimed at defining potential functions of v-snoRNA1. We thereby combined a computational with an experimental approach to determine putative ribosomal or spliceosomal RNA targets for v-snoRNA1 using previously identified criteria (see section 9.2). However, both attempts failed to identify any obvious rRNA candidates. Hence, v-snoRNA1 can be assigned in all probability to the class of so-called "orphan" snoRNAs that lack rRNA or snRNA targets (see below).

Another strategy to discover the function of v-snoRNA1 consisted in constructing a v-snoRNA1-null mutant and defining its phenotypic traits using well-characterized *in vitro* assays. As of now, the Δ v-snoRNA1 mutant remained indistinguishable from its wild type counterparts in terms of lytic replication, infection and B cell transformation (Fig. 9.5). However, this does not exclude that v-snoRNA1 serves an important function during the virus life cycle; unraveling miRNAs contributions to EBV infec-

Expression and processing of a small nucleolar RNA...

tion has also proven a difficult enterprise. Aside from a few notable exceptions such as miR-BART5 and miR-BART2 that respectively target the cellular gene *PUMA* (279) and the viral gene *BALF5* (280) or the BART cluster 1 and BHRF1-2 that respectively modulate LMP1 expression and BHRF1 mRNA processing (281; 282), the essential functions served by these ncRNAs remain unclear. Indeed, the B95.8 strain that lacks a large number of miRNAs perfectly replicates and immortalizes primary B cells with high efficiency.

Recently, specific snoRNA species have been characterized as miRNA precursors, which are processed to mature miRNAs and assemble into a functional RNA induced silencing complex (283; 284). Indeed, by deep-sequencing we identified nine identical cDNA clones of 24 nt in size, that mapped to the very 3'-end of v-snoRNA1. The expression of v-snoRNA1^{24pp} was verified by northern blot analysis employing a specific LNA oligonucleotide antisense probe (Fig. 9.7). Thereby, the hybridization signal was especially apparent in 293/EBV cells induced by *BZLF1*, which results in a 30-fold up-regulation of v-snoRNA1 expression; the hybridization signal was absent, however, in non-induced wild type cells. This could be explained by lower v-snoRNA1 expression levels in non-induced 293/EBV cells, compared to *BZLF1*-induced cells (Fig. 9.7), resulting in reduced processing of v-snoRNA1^{24pp} below the northern blot detection limit. Alternatively, this finding could result from preferential processing of v-snoRNA1 into v-snoRNA1^{24pp} during lytic replication.

Subsequently, by a 5'-RACE approach we also investigated a potential target for snoRNA1^{24pp}. Since the RNA species maps in antisense orientation to the 3'-UTR of the BALF5 mRNA, which encodes the viral DNA polymerase, BALF5 mRNA might represent a likely target site. As has been shown previously, the 3'-UTR of the BALF5 mRNA encodes in antisense orientation, in addition to v-snoRNA1^{24pp}, a bona fide EBV miRNA, designated as mir-BART2. Thereby, it has been reported that mir-BART2 down-regulates the mRNA levels by cleavage within the BALF5 3'-UTR (280). According to the proposed model, mir-BART2 thereby inhibits the transition from latent to lytic viral replication. By 5'-RACE analysis, we provide evidence that v-snoRNA1^{24pp} might also target BALF5 mRNA for cleavage and subsequent degradation. In contrast to mir-BART2, however, expression of v-snoRNA1^{24pp} was only apparent upon induction of the viral lytic cycle by *BZLF1* (Fig. 9.7). Future experiments will focus on the function of v-snoRNA1 and v-snoRNA1^{24pp} especially in respect to its function in the latent and lytic cycles of EBV infection.

9.4 Parts of Materials and Methods

9.4.1 Computational prediction of target sites in rRNAs

We predicted putative rRNA target sites for the snoRNAs in this study as follows. We first downloaded from Genbank the sequences of the human 18S (Accession NR_003286) and 28S (Accession NR_003287) rRNAs. The sequences of the antisense D-box (TGACGAAATCGGTTGAGATT) and D'-box (TGACAACCGCGGCTGT) were used to search for subsequences with good complementarity to the rRNAs with the program described in Mandin P et al. (266). As the study of Cavaille & Bachellerie (193) indicated that snoRNA-rRNA interactions involve regions of at least 7 nucleotides complementarity that are located at most 3 nucleotides from the end of the snoRNA antisense box, and that bulges and loops of more than 1 nucleotide are disfavored, we implemented these constraints in our programs. That is, we first used relatively large penalties for the introduction and extension of bulges and loops (a score penalty of 8), and we restricted the maximum size of loops and bulges to 1 nucleotide. The energy parameters of nucleotide-nucleotide interactions were kept with their default values coded in the program. We then extracted only hybrids that contained at least 7 nucleotide-nucleotide pairs, that ended within 3 nucleotides of the end of the antisense box, and that did not contain more than one bulge or loop.

Accession numbers

v-snoRNA1: FN376861; BZLF1: NC_007605.1; 18S rRNA: NR_003286; 28S rRNA: NR_003287; Epstein-Barr-Virus genome, strain AG876: AJ507799; Rhesus lymphocryptovirus genome: NC_006146

Expression and processing of a small nucleolar RNA...

Table 9.1: Potential target site of 18S and 28S rRNA complementary to v-snoRNA1

18S target predictions: box-D' long TGACAACCGCGCTGT

Homo sapiens 18S ribosomal RNA-w617; Score=227; snoRNA start=6; snoRNA end=16
target start=617 target end=627

```
ncRNA:    TGTCGGCGCCA
hybrid:    |||||
target:    GCAGCCGCGGT
```

Homo sapiens 18S ribosomal RNA-w873; Score=131; snoRNA start=5; snoRNA end=14
target start=873 target end=883

```
ncRNA:    TC.GGCGCCAA
hybrid:    || |||||
target:    GGACCGCGGTT
```

28S target predictions: box-D' short TGACAACCGCGG

Homo sapiens 28S ribosomal RNA-w2688; Score=130; snoRNA start=5; snoRNA end=12
target start=2688 target end=2695

```
ncRNA:    GGCGCCAA
hybrid:    |||||
target:    TCGCGGTT
```

28S target predictions: box-D' long TGACAACCGCGCTGT

Homo sapiens 28S ribosomal RNA-w3536; Score=181; snoRNA start=5; snoRNA end=14
target start=3536 target end=3545

```
ncRNA:    TCGGCGCCAA
hybrid:    |||||
target:    GGCCGCGGTT
```

Homo sapiens 28S ribosomal RNA-w3498; Score=156; snoRNA start=7; snoRNA end=14
target start=3498 target end=3505

```
ncRNA:    TCGGCGCC
hybrid:    |||||
target:    GGCCGCGG
```

Homo sapiens 28S ribosomal RNA-w3140; Score=155; snoRNA start=7; snoRNA end=16
target start=3140 target end=3150

```
ncRNA:    TGTCGGC.GCC
hybrid:    ||||| |||
target:    GCGGCCGCGG
```

Homo sapiens 28S ribosomal RNA-w3152; Score=155; snoRNA start=7; snoRNA end=16
target start=3152 target end=3162

```
ncRNA:    TGTCGG.CGCC
hybrid:    ||||| |||
target:    GCGGCCGCGG
```

Homo sapiens 28S ribosomal RNA-w2926; Score=149; snoRNA start=7; snoRNA end=16
target start=2926 target end=2934

```
ncRNA:    TGTCGGCGCC
hybrid:    || |||||
target:    GC.GCCGCGG
```

9.9.4 Parts of Materials and Methods

Homo sapiens 28S ribosomal RNA-w4704; Score=149; snoRNA start=7; snoRNA end=16
target start=4704 target end=4712
ncRNA: TGTCGGCGCC
hybrid: || |||||
target: GC.GCCGCGG

Homo sapiens 28S ribosomal RNA-w2685; Score=142; snoRNA start=5; snoRNA end=16
target start=2685 target end=2695
ncRNA: TGTCGGCGCAA
hybrid: || |||||
target: GC.GTCGCGGTT

Homo sapiens 28S ribosomal RNA-w845; Score=138; snoRNA start=6; snoRNA end=16
target start=845 target end=855
ncRNA: TGTCGGCGCCA
hybrid: ||||| |||||
target: GCGGCGGCGGT

Homo sapiens 28S ribosomal RNA-w2275; Score=138; snoRNA start=6; snoRNA end=16
target start=2275 target end=2285
ncRNA: TGTCGGCGCCA
hybrid: || |||||
target: GCCGCTGCGGT

Homo sapiens 28S ribosomal RNA-w3273; Score=138; snoRNA start=9; snoRNA end=16
target start=3273 target end=3280
ncRNA: TGTCGGCG
hybrid: |||||
target: GCGGCCG

Homo sapiens 28S ribosomal RNA-w3372; Score=138; snoRNA start=6; snoRNA end=16
target start=3372 target end=3382
ncRNA: TGTCGGCGCCA
hybrid: ||||| |||||
target: GCGGCGGCGGT

Homo sapiens 28S ribosomal RNA-w4878; Score=138; snoRNA start=9; snoRNA end=16
target start=4878 target end=4885
ncRNA: TGTCGGCG
hybrid: |||||
target: GCGGCCG

Homo sapiens 28S ribosomal RNA-w3256; Score=134; snoRNA start=7; snoRNA end=15
target start=3256 target end=3264
ncRNA: GTCGGCGCC
hybrid: |||||
target: CAGCTGCGG

Expression and processing of a small nucleolar RNA...

Homo sapiens 28S ribosomal RNA-w2102; Score=123; snoRNA start=8; snoRNA end=16
target start=2102 target end=2111
ncRNA: TGTCG.GCGC
hybrid: ||||| ||||
target: GCGGCGCGCG

Homo sapiens 28S ribosomal RNA-w3131; Score=123; snoRNA start=8; snoRNA end=14
target start=3131 target end=3137
ncRNA: TCGGCGC
hybrid: |||||
target: GGCCGCG

Homo sapiens 28S ribosomal RNA-w241; Score=122; snoRNA start=8; snoRNA end=16
target start=241 target end=250
ncRNA: TGTCGG.CGC
hybrid: ||||| |||
target: GCGGCCGCGCG

Homo sapiens 28S ribosomal RNA-w857; Score=122; snoRNA start=7; snoRNA end=16
target start=857 target end=866
ncRNA: TGTCGGCGCC
hybrid: ||||| ||||
target: GCGGCGGCGG

Homo sapiens 28S ribosomal RNA-w1865; Score=122; snoRNA start=7; snoRNA end=16
target start=1865 target end=1873
ncRNA: TGTCGGCGCC
hybrid: || |||||
target: GC.GCTGCGG

Homo sapiens 28S ribosomal RNA-w2140; Score=122; snoRNA start=7; snoRNA end=16
target start=2140 target end=2149
ncRNA: TGTCGGCGCC
hybrid: ||||| ||||
target: GCGGCGGCGG

Homo sapiens 28S ribosomal RNA-w3469; Score=122; snoRNA start=7; snoRNA end=16
target start=3469 target end=3478
ncRNA: TGTCGGCGCC
hybrid: ||||| ||||
target: GCGGCGGCGG

Homo sapiens 28S ribosomal RNA-w4013; Score=122; snoRNA start=6; snoRNA end=14
target start=4013 target end=4022
ncRNA: TCGGC.GCCA
hybrid: ||||| ||||
target: GGCCGCCGGT

Homo sapiens 28S ribosomal RNA-w488; Score=121; snoRNA start=7; snoRNA end=16
target start=488 target end=496
ncRNA: TGTCGGCGCC
hybrid: ||||| |||
target: GCGGCC.CGG

Chapter 10

Conclusions

By a computational and experimental approach we have extended the list of transcripts regulated by MBII-52 by five new members. By RT-PCR in transfection assays we have shown that the splicing events are consistently dependent of MBII-52 expression. Moreover, using a heterologous exon trap system which helps to separate splice events from the genomic context of the gene, we have proven that MBII-52 RNAs act on defined parts of the pre-mRNA in a mechanism that is independent of the promoter usage.

Interestingly, all the functional hybrid structures between MBII-52 and its targets contain three mismatches. Prior to testing these targets, we have focused on the predictions with one or two mismatches, but failed to observe a dependency of these exons on the MBII-52 expression.

We could confirm in various ways the presence of processed forms of snoRNAs lacking sequences which form the snoRNA stem. This form appears to be further shortened giving rise to smaller RNAs. We hypothesize that an unknown RNase initially removes the stem of the C/D box snoRNA. This dominant form might be stabilized by binding to other proteins. This form subsequently is shortened by exonucleases, and the resulting forms are stable, either stabilized by secondary structure formation or by other proteins.

In addition to MBII-52, we have analyzed, in a similar fashion a more diverse cluster of snoRNAs (MBII-85) coming from a neighboring locus. Given hundreds of putative target sites in the transcriptome, it was impossible to test a substantial fraction of them. Yet, our effort targeted to manually selected subset of these did not result in confirmation of any significant splicing form alternations. It is not unlikely that this cluster has a different function which still needs to be discovered.

By deep sequencing of a library of ncRNAs selectively expressed in EBV-infected cells we have identified sequences which exhibit the defining features of box C/D snoRNA. This potentially novel snoRNA species mapped to the EBV genome and was therefore designated as v-snoRNA1. Subsequently, we have shown that v-snoRNA1 is

Conclusions

a present in the latent transcriptome, however is 30-fold upregulated upon induction of the lytic replication cycle. The following analysis has shown nucleolar localization of v-snoRNA1 and snoRNP formation with the canonical partner proteins: fibrillarin, Nop56 and Nop58.

The combined computational and experimental approach failed to determine any putative ribosomal or spliceosomal RNA targets for v-snoRNA1, thus it can be classified to a substantial group of “orphan” snoRNAs. The exact function of (unprocessed) v-snoRNA1 remains unknown as the Δ v-snoRNA1 mutant we constructed did not show any obvious phenotypic traits.

By deep sequencing we have shown, however, that v-snoRNA1 gives rise to a shorter product, 24 nt in size, that maps to the very 3'-end of the v-snoRNA1. This processing product could only be observed during lytic replication. Since this RNA species maps in the antisense orientation to the 3'-UTR of the BALF5 transcript (which encodes the viral DNA polymerase), we postulate that the processing product functions as miRNA.

A common feature of v-snoRNA1 and the dominant form B of MBII-52 is the lack of the canonical terminal stem-structure, which is usually encountered in eukaryal snoRNAs. Given that both snoRNAs are evolutionarily conserved and functional (MBII-52) or most likely functional (v-snoRNA1), we hypothesize an existence of common processing steps, which might be also apply to a broader subclass of the “orphan” box C/D snoRNAs.

Part III
Acknowledgments

First of all, I would like to express my gratitude to my supervisor Prof. Erik van Nimwegen for his great ideas, teaching about Bayesian formalism and lots of insightful discussions. I am also grateful to Prof. Mihaela Zavolan for giving me an opportunity to work with her, patience and interesting projects.

Next, I would also like to thank Mikhail for the amount of effort in implementing various methods described in this work and making them publicly available.

I would also like to thank my friends Philipp, Biter, Jean, Evgeniy, Phil, Mohsen, Luise and Shivendra for challenging discussions, their help and especially for their good company during my stay in Basel.

Last but not least, I would like to thank my parents, Renata and Edward. Without their support and encouragement I would never have gotten so far.

Bibliography

- [1] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci U S A* 2003, **100**(26):15776–15781, [<http://dx.doi.org/10.1073/pnas.2136655100>].
- [2] Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117**(2):185–198.
- [3] Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat. Genet.* 2001, **27**:167–171.
- [4] Nguyen DH, D'haeseleer P: **Deciphering principles of transcription regulation in eukaryotic genomes.** *Mol Syst Biol* 2006, **2**:2006.0012, [<http://dx.doi.org/10.1038/msb4100054>].
- [5] Bussemaker HJ, Foat BC, Ward LD: **Predictive modeling of genome-wide mRNA expression: from modules to molecules.** *Annu Rev Biophys Biomol Struct* 2007, **36**:329–347, [<http://dx.doi.org/10.1146/annurev.biophys.36.040306.132725>].
- [6] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature Methods* 2008, **5**(7):621–628.
- [7] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.** *Science* 2008, **320**:1344–1349.
- [8] Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453**:1239–1245.

BIBLIOGRAPHY

- [9] Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H: **Characterizing the mouse ES cell transcriptome with Illumina sequencing.** *Genomics* 2008, **187**:1871–194.
- [10] Maeda N, Nishiyori H, Nakamura M, Kawazu C, Murata M, Sano H, Hayashida K, Fukuda S, Tagami M, Hasegawa A, Murakami K, Schroder K, Hume KID, Hayashizaki Y, Carninci P, Suzuki H: **Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra high-throughput sequencer.** *Biotechniques* 2008, **45**:95–97.
- [11] Carninci P: **Tagging mammalian transcription complexity.** *Trends Genet* 2006, **22**(9):501–510.
- [12] Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38**(6):626–635, [<http://dx.doi.org/10.1038/ng1789>].
- [13] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Gatta GD, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ieko K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SPT, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Babu MM, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin

- A, Schneider C, Schönbach C, Sekiguchi K, Sempé CAM, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusica V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y, FANTOM Consortium, RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group): **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**(5740):1559–1563.
- [14] Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA microarray measurements.** *Trends Genet* 2006, **22**(2):65–66.
- [15] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer FM: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.** *Journal of the American Statistical Association* 2004, **99**(468):909–917.
- [16] Lassmann T: **Manuscript in preparation, available at <http://genome.gsc.riken.jp/osc/english/software>.** [Unpublished].
- [17] Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM: **A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE.** *Genomics* 2008, **91**:281–288.
- [18] Ueda HR, Hayashi S, Matsuyama S, Yomo T, Hashimoto S, Kay SA, Hogenesch JB, Iino M: **Universality and flexibility in gene expression from bacteria to human.** *PNAS* 2004, **101**(11):3765–3769.
- [19] Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M: **Modeling ChIP sequencing in silico with applications.** *PLoS Comput Biol* 2008, **4**(8):e1000158, [<http://dx.doi.org/10.1371/journal.pcbi.1000158>].
- [20] Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**(2):185–193.

BIBLIOGRAPHY

- [21] Lu T, Costello CM, Croucher PJP, Häsler R, Deuschl G, Schreiber S: **Can Zipf's law be adapted to normalize microarrays?** *BMC Bioinformatics* 2005, **6**:37.
- [22] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4**(8):651–657, [<http://dx.doi.org/10.1038/nmeth1068>].
- [23] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res.* 2008, **18**:1509–1517.
- [24] Jaynes ET: *Probability Theory: The Logic of Science.* Cambridge University Press 2003.
- [25] FANTOM Consortium, RIKEN Omics Science Center: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet* 2009, **41**(5):553–562, [<http://dx.doi.org/10.1038/ng.375>].
- [26] Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A: **A code for transcription initiation in mammalian genomes.** *Genome Res* 2008, **18**:1–12, [<http://dx.doi.org/10.1101/gr.6831208>].
- [27] Siepel A, Bejerano G, Pedersen JS, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res.* 2005, **15**:1034–1050.
- [28] Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36**(Database issue):D773–D779, [<http://dx.doi.org/10.1093/nar/gkm966>].
- [29] Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.** *PNAS* 2006, **103**(5):1412–1417.

- [30] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P: **CAGE: cap analysis of gene expression**. *Nat Methods* 2006, **3**(3):211–222, [<http://dx.doi.org/10.1038/nmeth0306-211>].
- [31] Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, Marstrand TT, Tang MHE, Zhao X, Krogh A, Winther O, Arakawa T, Kawai J, Wells C, Daub C, Harbers M, Hayashizaki Y, Gustincich S, Sandelin A, Carninci P: **Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE**. *Genome Res* 2009, **19**(2):255–265, [<http://dx.doi.org/10.1101/gr.084541.108>].
- [32] Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins**. *J Mol Biol* 1961, **3**:318–356.
- [33] Bartel D: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**:215–233.
- [34] Fabian MR, Sonenberg N, Filipowicz W: **Regulation of mRNA translation and stability by microRNAs**. *Annu. Rev. Biochem.* 2010, **79**:351–379.
- [35] Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**:374–378.
- [36] Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**(4):276–287, [<http://dx.doi.org/10.1038/nrg1315>].
- [37] Pachkov M, Erb I, Molina N, van Nimwegen E: **SwissRegulon: a database of genome-wide annotations of regulatory sites**. *Nucleic Acids Res* 2007, **35**(Database issue):D127–D131, [<http://dx.doi.org/10.1093/nar/gkl857>].
- [38] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution**. *Nat Rev Genet* 2009, **10**(4):252–263, [<http://dx.doi.org/10.1038/nrg2538>].
- [39] van Nimwegen E: **Finding regulatory elements and regulatory motifs: a general probabilistic framework**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S4, [<http://dx.doi.org/10.1186/1471-2105-8-S6-S4>].

BIBLIOGRAPHY

- [40] Friedman RC, Farh KKH, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs.** *Genome Res* 2009, **19**:92–105, [<http://dx.doi.org/10.1101/gr.082701.108>].
- [41] Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E: **MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences.** *Bioinformatics* 2012, **28**(4):487–494, [<http://dx.doi.org/10.1093/bioinformatics/btr695>].
- [42] de Hoon M, Hayashizaki Y: **Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference.** *Biotechniques* 2008, **44**(5):627–8, 630, 632, [<http://dx.doi.org/10.2144/000112802>].
- [43] Gao F, Foat BC, Bussemaker HJ: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.** *BMC Bioinformatics* 2004, **5**:31, [<http://dx.doi.org/10.1186/1471-2105-5-31>].
- [44] Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ: **Expression analysis of G Protein-Coupled Receptors in mouse macrophages.** *Immunome Res* 2008, **4**:5, [<http://dx.doi.org/10.1186/1745-7580-4-5>].
- [45] Thangue NBL: **DRTF1/E2F: an expanding family of heterodimeric transcription factors implicated in cell-cycle control.** *Trends Biochem Sci* 1994, **19**(3):108–114.
- [46] Trimarchi JM, Lees JA: **Sibling rivalry in the E2F family.** *Nat. Rev. Mol. Cell Biol.* 2002, **3**:11–20.
- [47] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8—a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**(Database issue):D412–D416, [<http://dx.doi.org/10.1093/nar/gkn760>].
- [48] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29, [<http://dx.doi.org/10.1038/75556>].

- [49] Campanero MR, Armstrong M, Flemington E: **Distinct cellular factors regulate the c-myc promoter through its E2F element.** *Mol. Cell. Biol.* 1999, **19**(12):8442–8450.
- [50] Longworth MS, Wilson R, Laimins LA: **HPV31 E7 facilitates replication by activating E2F2 transcription through its interaction with HDACs.** *EMBO J.* 2005, **24**(10):1821–1830.
- [51] Kuo CJ, Conley PB, Hsieh CL, Francke U, Crabtree GR: **Molecular cloning, functional expression, and chromosomal localization of mouse hepatocyte nuclear factor 1.** *Proc. Natl. Acad. Sci. U.S.A.* 1990, **87**:9838–9842.
- [52] Serfas MS, Tyner AL: **HNF-1 alpha and HNF-1 beta expression in mouse intestinal crypts.** *Am. J. Physiol.* 1993, **265**:G506–513.
- [53] Pontoglio M, Barra J, Hadchouel M, Doyen A, Kress C, Bach JP, Babinet C, Yaniv M: **Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria, and renal Fanconi syndrome.** *Cell* 1996, **84**:575–585.
- [54] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101**(16):6062–6067, [<http://dx.doi.org/10.1073/pnas.0400782101>].
- [55] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, de Rijn MV, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**(3):227–235, [<http://dx.doi.org/10.1038/73432>].
- [56] Semenza GL: **Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics.** *Oncogene* 2010, **29**(5):625–634.
- [57] Meyer N, Penn LZ: **Reflecting on 25 years with MYC.** *Nat. Rev. Cancer* 2008, **8**(12):976–990.
- [58] Chen HZ, Tsai SY, Leone G: **Emerging roles of E2Fs in cancer: an exit from cell cycle control.** *Nat. Rev. Cancer* 2009, **9**(11):785–797.
- [59] Gandellini P, Folini M, Longoni N, Pennati M, Binda M, Colecchia M, Salvioni R, Supino R, Moretti R, Limonta P, Valdagni R, Daidone MG, Zaffaroni N: **miR-205 Exerts tumor-suppressive functions in human prostate**

BIBLIOGRAPHY

- through down-regulation of protein kinase Cepsilon. *Cancer Res.* 2009, **69**(6):2287–2295.
- [60] Majid S, Dar AA, Saini S, Yamamura S, Hirata H, Tanaka Y, Deng G, Dahiya R: **MicroRNA-205-directed transcriptional activation of tumor suppressor genes in prostate cancer.** *Cancer* 2010, **116**(24):5637–5649.
- [61] Dar AA, Majid S, de Semir D, Nosrati M, Bezrookove V, Kashani-Sabet M: **miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein.** *J. Biol. Chem.* 2011, **286**(19):16606–16614.
- [62] Wu H, Zhu S, Mo YY: **Suppression of cell growth and invasion by miR-205 in breast cancer.** *Cell Res.* 2009, **19**(4):439–448.
- [63] Liu S, Tetzlaff MT, Liu A, Liegl-Atzwanger B, Guo J, Xu X: **Loss of microRNA-205 expression is associated with melanoma progression.** *Lab. Invest.* 2012, **92**(7):1084–1096.
- [64] Kota J, Chivukula RR, O'Donnell KA, Wentzel EA, Montgomery CL, Hwang HW, Chang TC, Vivekanandan P, Torbenson M, Clark KR, Mendell JR, Mendell JT: **Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model.** *Cell* 2009, **137**(6):1005–1017.
- [65] He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM: **A microRNA polycistron as a potential human oncogene.** *Nature* 2005, **435**(7043):828–833.
- [66] Chhabra R, Dubey R, Saini N: **Cooperative and individualistic functions of the microRNAs in the miR-23a 27a 24-2 cluster and its implication in human diseases.** *Mol. Cancer* 2010, **9**:232.
- [67] To KH, Pajovic S, Gallie BL, Theriault BL: **Regulation of p14ARF expression by miR-24: a potential mechanism compromising the p53 response during retinoblastoma development.** *BMC Cancer* 2012, **12**:69.
- [68] Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, Becker KG, Gorospe M, Hide W, Lieberman J: **miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements.** *Mol. Cell* 2009, **35**(5):610–625.

- [69] Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, Inoue K, Komura D, Kitakami J, Oshida N, Papantonis A, Izumi A, Kobayashi M, Meguro H, Kanki Y, Mimura I, Yamamoto K, Mataka C, Hamakubo T, Shirahige K, Aburatani H, Kimura H, Kodama T, Cook PR, Ihara S: **A wave of nascent transcription on activated human genes.** *Proc Natl Acad Sci U S A* 2009, **106**(43):18357–18361, [<http://dx.doi.org/10.1073/pnas.0902573106>].
- [70] Inoue K, Kobayashi M, Yano K, Miura M, Izumi A, Mataka C, Doi T, Hamakubo T, Reid PC, Hume DA, Yoshida M, Aird WC, Kodama T, Minami T: **Histone deacetylase inhibitor reduces monocyte adhesion to endothelium through the suppression of vascular cell adhesion molecule-1 expression.** *Arterioscler Thromb Vasc Biol* 2006, **26**(12):2652–2659, [<http://dx.doi.org/10.1161/01.ATV.0000247247.89787.e7>].
- [71] Kempe S, Kestler H, Lasar A, Wirth T: **NF-kappaB controls the global pro-inflammatory response in endothelial cells: evidence for the regulation of a pro-atherogenic program.** *Nucleic Acids Res* 2005, **33**(16):5308–5319, [<http://dx.doi.org/10.1093/nar/gki836>].
- [72] Harada H, Takahashi E, Itoh S, Harada K, Hori TA, Taniguchi T: **Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system.** *Mol Cell Biol* 1994, **14**(2):1500–1509.
- [73] Ten RM, Blank V, Le Bail O, Kourilsky P, Israël A: **Two factors, IRF1 and KBF1/NF-kappa B, cooperate during induction of MHC class I gene expression by interferon alpha beta or Newcastle disease virus.** *C R Acad Sci III* 1993, **316**(5):496–501.
- [74] Martins G, Calame K: **Regulation and functions of Blimp-1 in T and B lymphocytes.** *Annu Rev Immunol* 2008, **26**:133–169, [<http://dx.doi.org/10.1146/annurev.immunol.26.021607.090241>].
- [75] Seifert U, Bialy LP, Ebstein F, Bech-Otschir D, Voigt A, Schröter F, Prozorovski T, Lange N, Steffen J, Rieger M, Kuckelkorn U, Aktas O, Kloetzel PM, Krüger E: **Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress.** *Cell* 2010, **142**(4):613–624, [<http://dx.doi.org/10.1016/j.cell.2010.07.036>].
- [76] Glimcher LH: **XPB1: the last two decades.** *Ann Rheum Dis* 2010, **69** Suppl 1:i67–i71, [<http://dx.doi.org/10.1136/ard.2009.119388>].
- [77] Gargalovic PS, Gharavi NM, Clark MJ, Pagnon J, Yang WP, He A, Truong A, Baruch-Oren T, Berliner JA, Kirchgessner TG, Luscis AJ: **The unfolded**

BIBLIOGRAPHY

- protein response is an important regulator of inflammatory genes in endothelial cells.** *Arterioscler Thromb Vasc Biol* 2006, **26**(11):2490–2496, [<http://dx.doi.org/10.1161/01.ATV.0000242903.41158.a1>].
- [78] Civelek M, Manduchi E, Riley RJ, Stoeckert CJ Jr, Davies PF: **Chronic endoplasmic reticulum stress activates unfolded protein response in arterial endothelium in regions of susceptibility to atherosclerosis.** *Circ Res* 2009, **105**(5):453–461, [<http://dx.doi.org/10.1161/CIRCRESAHA.109.203711>].
- [79] Kitamura M: **Control of NF- κ B and inflammation by the unfolded protein response.** *Int Rev Immunol* 2011, **30**:4–15, [<http://dx.doi.org/10.3109/08830185.2010.522281>].
- [80] Kaser A, Lee AH, Franke A, Glickman JN, Zeissig S, Tilg H, Nieuwenhuis EES, Higgins DE, Schreiber S, Glimcher LH, Blumberg RS: **XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease.** *Cell* 2008, **134**(5):743–756, [<http://dx.doi.org/10.1016/j.cell.2008.07.021>].
- [81] Li J, Wang JJ, Zhang SX: **Preconditioning with endoplasmic reticulum stress mitigates retinal endothelial inflammation via activation of X-box binding protein 1.** *J Biol Chem* 2011, **286**(6):4912–4921, [<http://dx.doi.org/10.1074/jbc.M110.199729>].
- [82] Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K: **XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor.** *Cell* 2001, **107**(7):881–891.
- [83] Calton M, Zeng H, Urano F, Till JH, Hubbard SR, Harding HP, Clark SG, Ron D: **IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA.** *Nature* 2002, **415**(6867):92–96.
- [84] Ross AJ, Dailey LA, Brighton LE, Devlin RB: **Transcriptional profiling of mucociliary differentiation in human airway epithelial cells.** *Am J Respir Cell Mol Biol* 2007, **37**(2):169–185, [<http://dx.doi.org/10.1165/rcmb.2006-0466OC>].
- [85] Bonnafe E, Touka M, AitLounis A, Baas D, Barras E, Ucla C, Moreau A, Flamant F, Dubruille R, Couble P, Collignon J, Durand B, Reith W: **The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification.** *Mol Cell Biol* 2004, **24**(10):4417–4427.
- [86] Zein LE, Ait-Lounis A, Morlé L, Thomas J, Chhin B, Spassky N, Reith W, Durand B: **RFX3 governs growth and beating efficiency of motile cilia**

- in mouse and controls the expression of genes involved in human ciliopathies. *J Cell Sci* 2009, **122**(Pt 17):3180–3189, [<http://dx.doi.org/10.1242/jcs.048348>].
- [87] Horvath GC, Kistler MK, Kistler WS: **RFX2 is a candidate downstream amplifier of A-MYB regulation in mouse spermatogenesis.** *BMC Dev Biol.* 2009, **9**:63.
- [88] Scheel C, Eaton EN, Li SHJ, Chaffer CL, Reinhardt F, Kah KJ, Bell G, Guo W, Rubin J, Richardson AL, Weinberg RA: **Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast.** *Cell* 2011, **145**(6):926–940, [<http://dx.doi.org/10.1016/j.cell.2011.04.029>].
- [89] Polyak K, Weinberg RA: **Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits.** *Nat Rev Cancer* 2009, **9**(4):265–273, [<http://dx.doi.org/10.1038/nrc2620>].
- [90] Xiong M, Jiang L, Zhou Y, Qiu W, Fang L, Tan R, Wen P, Yang J: **The miR-200 family regulates TGF- β 1-induced renal tubular epithelial to mesenchymal transition through Smad pathway by targeting ZEB1 and ZEB2 expression.** *Am J Physiol Renal Physiol* 2012, **302**(3):F369–F379, [<http://dx.doi.org/10.1152/ajprenal.00268.2011>].
- [91] Burk U, Schubert J, Wellner U, Schmalhofer O, Vincan E, Spaderna S, Brabletz T: **A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells.** *EMBO Rep* 2008, **9**(6):582–589, [<http://dx.doi.org/10.1038/embor.2008.74>].
- [92] Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, Vadas MA, Khew-Goodall Y, Goodall GJ: **The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1.** *Nat Cell Biol* 2008, **10**(5):593–601, [<http://dx.doi.org/10.1038/ncb1722>].
- [93] Hajra KM, Chen DYS, Fearon ER: **The SLUG zinc-finger protein represses E-cadherin in breast cancer.** *Cancer Res* 2002, **62**(6):1613–1618.
- [94] Grootclaes ML, Frisch SM: **Evidence for a function of CtBP in epithelial gene regulation and anoikis.** *Oncogene* 2000, **19**(33):3823–3828, [<http://dx.doi.org/10.1038/sj.onc.1203721>].
- [95] F T, R Z, Y H, M Z, L G, et al: **MicroRNA-125b Induces Metastasis by Targeting STARD13 in MCF-7 and MDA-MB-231 Breast Cancer Cells.** *PLoS ONE* 2012, **7**(5):e35435.

BIBLIOGRAPHY

- [96] Arnold P, Scholer A, Pachkov M, Balwierz P, Jørgensen H, Stadler MB, van Nimwegen E, Schubeler D: **Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting.** *Genome Res.* 2012.
- [97] Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**(7345):43–49.
- [98] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat. Genet.* 2008, **40**(7):897–903.
- [99] Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M: **Histone modification levels are predictive for gene expression.** *Proc. Natl. Acad. Sci. U.S.A.* 2010, **107**(7):2926–2931.
- [100] Tippmann SC, Ivanek R, Gaidatzis D, Scholer A, Hoerner L, van Nimwegen E, Stadler PF, Stadler MB, Schubeler D: **Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels.** *Mol. Syst. Biol.* 2012, **8**:593.
- [101] Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigo R, Birney E, Weng Z: **Modeling gene expression using chromatin features in various cellular contexts.** *Genome Biol.* 2012, **13**(9):R53.
- [102] Heintzman ND, Ren B: **Finding distal regulatory elements in the human genome.** *Curr. Opin. Genet. Dev.* 2009, **19**(6):541–549.
- [103] Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome Res.* 2011, **21**(10):1757–1767.
- [104] Schuettengruber B, Cavalli G: **Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice.** *Development* 2009, **136**(21):3531–3542.

- [105] Yuan LW, Gambée JE: **Histone acetylation by p300 is involved in CREB-mediated transcription on chromatin.** *Biochim. Biophys. Acta* 2001, **1541**(3):161–169.
- [106] Masternak K, Peyraud N, Krawczyk M, Barras E, Reith W: **Chromatin remodeling and extragenic transcription at the MHC class II locus control region.** *Nat. Immunol.* 2003, **4**(2):132–137.
- [107] Gan Q, Thiebaud P, Theze N, Jin L, Xu G, Grant P, Owens GK: **WD repeat-containing protein 5, a ubiquitously expressed histone methyltransferase adaptor protein, regulates smooth muscle cell-selective gene activation through interaction with pituitary homeobox 2.** *J. Biol. Chem.* 2011, **286**(24):21853–21864.
- [108] Kizer KO, Phatnani HP, Shibata Y, Hall H, Greenleaf AL, Strahl BD: **A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation.** *Mol. Cell. Biol.* 2005, **25**(8):3305–3316.
- [109] Yuan W, Xie J, Long C, Erdjument-Bromage H, Ding X, Zheng Y, Tempst P, Chen S, Zhu B, Reinberg D: **Heterogeneous nuclear ribonucleoprotein L is a subunit of human KMT3a/Set2 complex required for H3 Lys-36 trimethylation activity in vivo.** *J. Biol. Chem.* 2009, **284**(23):15701–15707.
- [110] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation.** *Nature* 2008, **451**(7178):535–540, [<http://dx.doi.org/10.1038/nature06496>].
- [111] He X, Samee MAH, Blatti C, Sinha S: **Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression.** *PLoS Comput Biol* 2010, **6**(9), [<http://dx.doi.org/10.1371/journal.pcbi.1000935>].
- [112] Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126**(4):663–676, [<http://dx.doi.org/10.1016/j.cell.2006.07.024>].
- [113] Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D: **Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors.** *Cell* 2010, **142**(3):375–386, [<http://dx.doi.org/10.1016/j.cell.2010.07.002>].

BIBLIOGRAPHY

- [114] Kim HD, Shay T, O’Shea EK, Regev A: **Transcriptional regulatory circuits: predicting numbers from alphabets.** *Science* 2009, **325**(5939):429–432.
- [115] Ruan J: **A top-performing algorithm for the DREAM3 gene expression prediction challenge.** *PLoS One* 2010, **5**(2):e8944, [<http://dx.doi.org/10.1371/journal.pone.0008944>].
- [116] Summers KM, Raza S, van Nimwegen E, Freeman TC, Hume DA: **Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome.** *Eur. J. Hum. Genet.* 2010, **18**(11):1209–1215.
- [117] Aceto N, Sausgruber N, Brinkhaus H, Gaidatzis D, Martiny-Baron G, Mazzarol G, Confalonieri S, Quarto M, Hu G, Balwierz PJ, Pachkov M, Elledge SJ, van Nimwegen E, Stadler MB, Bentires-Alj M: **Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop.** *Nat Med* 2012, **18**(4):529–537, [<http://dx.doi.org/10.1038/nm.2645>].
- [118] Pérez-Schindler J, Summermatter S, Salatino S, Zorzato F, Beer M, Balwierz PJ, van Nimwegen E, Feige JN, Auwerx J, Handschin C: **The Corepressor NCoR1 Antagonizes PGC-1 α and ERR α in the Regulation of Skeletal Muscle Function and Oxidative Metabolism.** *Mol Cell Biol* 2012, [<http://dx.doi.org/10.1128/MCB.00877-12>].
- [119] Arner E, Mejhert N, Kulyté A, Balwierz PJ, Pachkov M, Cormont M, Lorente-Cebrián S, Ehrlund A, Laurencikiene J, Hedén P, Dahlman-Wright K, Tanti JF, Hayashizaki Y, Rydén M, Dahlman I, van Nimwegen E, Daub CO, Arner P: **Adipose tissue microRNAs as regulators of CCL2 production in human obesity.** *Diabetes* 2012, **61**(8):1986–1993, [<http://dx.doi.org/10.2337/db11-1508>].
- [120] Hasegawa R, Tomaru Y, de Hoon M, Suzuki H, Hayashizaki Y, Shin JW: **Identification of ZNF395 as a novel modulator of adipogenesis.** *Exp. Cell Res.* 2012.
- [121] Cui Q, Yu Z, Purisima E, Wang E: **Principles of microRNA regulation of a human cellular signaling network.** *Mol. Syst. Biol.* 2006, **2**:46.
- [122] Hornstein E, Shomron N: **Canalization of development by microRNAs.** *Nat. Genet.* 2006, **38** Suppl:S20–S24.

- [123] Zhou Y, Ferguson J, Chang J, Kluger Y: **Inter- and intra-combinatorial regulation by transcription factors and microRNAs.** *BMC Genomics* 2007, **8**:396.
- [124] Bauer DC, Buske FA, Bailey TL: **Dual-functioning transcription factors in the developmental gene network of *Drosophila melanogaster*.** *BMC Bioinformatics* 2010, **11**:366, [<http://dx.doi.org/10.1186/1471-2105-11-366>].
- [125] Bulyk ML: **DNA microarray technologies for measuring protein-DNA interactions.** *Curr Opin Biotechnol* 2006, **17**(4):422–430, [<http://dx.doi.org/10.1016/j.copbio.2006.06.015>].
- [126] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459**(7243):108–112, [<http://dx.doi.org/10.1038/nature07829>].
- [127] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D: **DNA-binding factors shape the mouse methylome at distal regulatory regions.** *Nature* 2011, **480**(7378):490–495.
- [128] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**(2):311–322.
- [129] Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205–217, [<http://dx.doi.org/10.1006/jmbi.2000.4042>].
- [130] Kent WJ: **BLAT—The BLAST-Like Alignment Tool.** *Genome Research* 2002, **12**:656–664.
- [131] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ, of California Santa Cruz U: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31**:51–54.
- [132] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic Acids Res.* 2010, **38**:D105–D110.

BIBLIOGRAPHY

- [133] Finn RD, Tate J, Mistry J, Coghill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database**. *Nucleic Acids Research* 2008, **36**:D281–D288.
- [134] Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**:755–763.
- [135] Siddharthan R, Siggia ED, van Nimwegen E: **PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny**. *PLoS Comput Biol* 2005, **1**(7):e67, [<http://dx.doi.org/10.1371/journal.pcbi.0010067>].
- [136] Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Airey M, Anagnostopoulos A, Babiuk R, Baldarelli R, Beal J, Bello S, Butler N, Campbell J, Corbani L, Giannatto S, Dene H, Dolan M, Drabkin H, Forthofer K, Knowlton M, Lewis J, McAndrews-Hill M, McClatchy S, Miers D, Ni L, Onda H, Ormsby JE, Recla J, Reed D, Richards-Smith B, Shaw R, Sinclair R, Sitnikov D, Smith C, Washburn L, Zhu Y: **The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse**. *Nucleic Acids Res.* 2012, **40**(Database issue):D881–886.
- [137] Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**:103–107, [<http://dx.doi.org/10.1101/gr.809403>].
- [138] Molina N, van Nimwegen E: **Universal patterns of purifying selection at noncoding positions in bacteria**. *Genome Res* 2008, **18**:148–160, [<http://dx.doi.org/10.1101/gr.6759507>].
- [139] Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, Pong SLK, Nekrutenko A, Giardine B, Harris RS, Tyekucheva S, Diekhans M, Pringle TH, Murphy WJ, Lesk A, Weinstock GM, Lindblad-Toh K, Gibbs RA, Lander ES, Siepel A, Haussler D, Kent WJ: **28-way vertebrate alignment and conservation track in the UCSC Genome Browser**. *Genome Res* 2007, **17**(12):1797–1808, [<http://dx.doi.org/10.1101/gr.6761107>].
- [140] Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy—analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**(3):307–315.
- [141] Carvalho BS, Irizarry RA: **A Framework for Oligonucleotide Microarray Preprocessing**. *Bioinformatics* 2010.
- [142] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays**. *Journal of the American Statistical Association* 2004, **99**:909.

- [143] Fraley C, Raftery AE: **Model-Based Clustering, Discriminant Analysis and Density Estimation**. *Journal of the American Statistical Association* 2002, **97**:611–631.
- [144] Fraley C, Raftery AE: **MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering**. Technical Report 504, University of Washington, Department of Statistics 2006, revised in 2009.
- [145] Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710–3715, [<http://dx.doi.org/10.1093/bioinformatics/bth456>].
- [146] Pachkov M, Balwiercz PJ, Arnold P, Ozonov E, van Nimwegen E: **SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates**. *Nucleic Acids Res.* 2012.
- [147] Tsuchiya S, Kobayashi Y, Goto Y, Okumura H, Nakae S, Konno T, Tada K: **Induction of maturation in cultured human monocytic leukemia cells by a phorbol diester**. *Cancer Res* 1982, **42**(4):1530–1536.
- [148] Abrink M, Gobl AE, Huang R, Nilsson K, Hellman L: **Human cell lines U-937, THP-1 and Mono Mac 6 represent relatively immature cells of the monocyte-macrophage cell lineage**. *Leukemia* 1994, **8**(9):1579–1584.
- [149] Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, Li B, Gilchrist M, Gold ES, Johnson CD, Litvak V, Navarro G, Roach JC, Rosenberger CM, Rust AG, Yudkovsky N, Aderem A, Shmulevich I: **Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics**. *PLoS Comput Biol* 2008, **4**(3):e1000021, [<http://dx.doi.org/10.1371/journal.pcbi.1000021>].
- [150] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166–176, [<http://dx.doi.org/10.1038/ng1165>].
- [151] Das D, Nahlé Z, Zhang MQ: **Adaptively inferring human transcriptional subnetworks**. *Mol Syst Biol* 2006, **2**:2006.0029, [<http://dx.doi.org/10.1038/msb4100067>].

BIBLIOGRAPHY

- [152] The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**(7146):799–816, [<http://dx.doi.org/10.1038/nature05874>].
- [153] Roh TY, Cuddapah S, Zhao K: **Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping.** *Genes Dev* 2005, **19**(5):542–552, [<http://dx.doi.org/10.1101/gad.1272505>].
- [154] Sandoval J, Rodríguez JL, Tur G, Serviddio G, Pereda J, Boukaba A, Sastre J, Torres L, Franco L, López-Rodas G: **RNAPol-ChIP: a novel application of chromatin immunoprecipitation to the analysis of real-time gene transcription.** *Nucleic Acids Res* 2004, **32**(11):e88, [<http://dx.doi.org/10.1093/nar/gnh091>].
- [155] Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**(7):613–619, [<http://dx.doi.org/10.1038/nmeth.1223>].
- [156] Vlieghe D, Sandelin A, Bleser PJD, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34**(Database issue):D95–D97, [<http://dx.doi.org/10.1093/nar/gkj115>].
- [157] Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**:238–241.
- [158] Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: **MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model.** *Genome Biol* 2004, **5**(12):R98, [<http://dx.doi.org/10.1186/gb-2004-5-12-r98>].
- [159] Schmitz G, Grandl M: **Lipid homeostasis in macrophages - implications for atherosclerosis.** *Rev Physiol Biochem Pharmacol* 2008, **160**:93–125, [http://dx.doi.org/10.1007/112_2008_802].
- [160] Odero MD, Zeleznik-Le NJ, Chinwalla V, Rowley JD: **Cytogenetic and molecular analysis of the acute monocytic leukemia cell line THP-1 with an MLL-AF9 translocation.** *Genes Chromosomes Cancer* 2000, **29**(4):333–338.

- [161] Martino V, Tonelli R, Montemurro L, Franzoni M, Marino F, Fazzina R, Pession A: **Down-regulation of MLL-AF9, MLL and MYC expression is not obligatory for monocyte-macrophage maturation in AML-M5 cell lines carrying t(9;11)(p22;q23)**. *Oncol Rep* 2006, **15**:207–211.
- [162] Pession A, Martino V, Tonelli R, Beltramini C, Locatelli F, Biserni G, Franzoni M, Freccero F, Montemurro L, Pattacini L, Paolucci G: **MLL-AF9 oncogene expression affects cell growth but not terminal differentiation and is downregulated during monocyte-macrophage maturation in AML-M5 THP-1 cells**. *Oncogene* 2003, **22**(54):8671–8676, [<http://dx.doi.org/10.1038/sj.onc.1207125>].
- [163] Roach JC, Smith KD, Strobe KL, Nissen SM, Haudenschild CD, Zhou D, Vasicsek TJ, Held GA, Stolovitzky GA, Hood LE, Aderem A: **Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells**. *Proc Natl Acad Sci U S A* 2007, **104**(41):16245–16250, [<http://dx.doi.org/10.1073/pnas.0707757104>].
- [164] Biggs JR, Ahn NG, Kraft AS: **Activation of the mitogen-activated protein kinase pathway in U937 leukemic cells induces phosphorylation of the amino terminus of the TATA-binding protein**. *Cell Growth Differ* 1998, **9**(8):667–676.
- [165] Iyer D, Chang D, Marx J, Wei L, Olson EN, Parmacek MS, Balasubramanyam A, Schwartz RJ: **Serum response factor MADS box serine-162 phosphorylation switches proliferation and myogenic gene programs**. *Proc Natl Acad Sci U S A* 2006, **103**(12):4516–4521, [<http://dx.doi.org/10.1073/pnas.0505338103>].
- [166] Morton S, Davis RJ, Cohen P: **Signalling pathways involved in multisite phosphorylation of the transcription factor ATF-2**. *FEBS Lett* 2004, **572**(1-3):177–183, [<http://dx.doi.org/10.1016/j.febslet.2004.07.031>].
- [167] Trejo J, Massamiri T, Deng T, Dewji NN, Bayney RM, Brown JH: **A direct role for protein kinase C and the transcription factor Jun/AP-1 in the regulation of the Alzheimer's beta-amyloid precursor protein gene**. *J Biol Chem* 1994, **269**(34):21682–21690.
- [168] Kelly LM, Englmeier U, Lafon I, Sieweke MH, Graf T: **MafB is an inducer of monocytic differentiation**. *EMBO J* 2000, **19**(9):1987–1997, [<http://dx.doi.org/10.1093/emboj/19.9.1987>].

BIBLIOGRAPHY

- [169] Krishnaraju K, Hoffman B, Liebermann DA: **The zinc finger transcription factor Egr-1 activates macrophage differentiation in M1 myeloblastic leukemia cells.** *Blood* 1998, **92**(6):1957–1966.
- [170] Mauxion F, Faux C, Séraphin B: **The BTG2 protein is a general activator of mRNA deadenylation.** *EMBO J* 2008, **27**(7):1039–1048, [<http://dx.doi.org/10.1038/emboj.2008.43>].
- [171] Blackshear PJ: **Tristetraprolin and other CCCH tandem zinc-finger proteins in the regulation of mRNA turnover.** *Biochem Soc Trans* 2002, **30**(Pt 6):945–952, [<http://dx.doi.org/10.1042/>].
- [172] Carey JO, Posekany KJ, deVente JE, Pettit GR, Ways DK: **Phorbol ester-stimulated phosphorylation of PU.1: association with leukemic cell growth inhibition.** *Blood* 1996, **87**(10):4316–4324.
- [173] Foster N, Lea SR, Preshaw PM, Taylor JJ: **Pivotal advance: vasoactive intestinal peptide inhibits up-regulation of human monocyte TLR2 and TLR4 by LPS and differentiation of monocytes to macrophages.** *J Leukoc Biol* 2007, **81**(4):893–903, [<http://dx.doi.org/10.1189/jlb.0206086>].
- [174] Xu X, Bieda M, Jin VX, Rabinovich A, Oberley MJ, Green R, Farnham PJ: **A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members.** *Genome Res* 2007, **17**(11):1550–1561, [<http://dx.doi.org/10.1101/gr.6783507>].
- [175] Anfossi G, Gewirtz AM, Calabretta B: **An oligomer complementary to c-myb-encoded mRNA inhibits proliferation of human myeloid leukemia cell lines.** *Proc Natl Acad Sci U S A* 1989, **86**(9):3379–3383.
- [176] Reddy MA, Yang BS, Yue X, Barnett CJ, Ross IL, Sweet MJ, Hume DA, Ostrowski MC: **Opposing actions of c-ets/PU.1 and c-myb protooncogene products in regulating the macrophage-specific promoters of the human and mouse colony-stimulating factor-1 receptor (c-fms) genes.** *J Exp Med* 1994, **180**(6):2309–2319.
- [177] Feng R, Desbordes SC, Xie H, Tillo ES, Pixley F, Stanley ER, Graf T: **PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells.** *Proc Natl Acad Sci U S A* 2008, **105**(16):6057–6062, [<http://dx.doi.org/10.1073/pnas.0711961105>].
- [178] Carter JH, Tourtellotte WG: **Early growth response transcriptional regulators are dispensable for macrophage differentiation.** *J Immunol* 2007, **178**(5):3038–3047.

- [179] Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133**(6):1106–1117, [<http://dx.doi.org/10.1016/j.cell.2008.04.043>].
- [180] Arsenian S, Weinhold B, Oelgeschläger M, Rütther U, Nordheim A: **Serum response factor is essential for mesoderm formation during mouse embryogenesis.** *EMBO J* 1998, **17**(21):6289–6299, [<http://dx.doi.org/10.1093/emboj/17.21.6289>].
- [181] Cooper SJ, Trinklein ND, Nguyen L, Myers RM: **Serum response factor binding sites differ in three human cell types.** *Genome Res* 2007, **17**(2):136–144, [<http://dx.doi.org/10.1101/gr.5875007>].
- [182] Fleige A, Alberti S, Gröbe L, Frischmann U, Geffers R, Müller W, Nordheim A, Schippers A: **Serum response factor contributes selectively to lymphocyte development.** *J Biol Chem* 2007, **282**(33):24320–24328, [<http://dx.doi.org/10.1074/jbc.M703119200>].
- [183] Poser S, Impey S, Trinh K, Xia Z, Storm DR: **SRF-dependent gene expression is required for PI3-kinase-regulated cell proliferation.** *EMBO J* 2000, **19**(18):4955–4966, [<http://dx.doi.org/10.1093/emboj/19.18.4955>].
- [184] Huang S, Ingber DE: **Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks.** *Exp Cell Res* 2000, **261**:91–103, [<http://dx.doi.org/10.1006/excr.2000.5044>].
- [185] Kauffman SA: *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, New York 1993.
- [186] Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA: **DBD-taxonically broad transcription factor predictions: new content and functionality.** *Nucleic Acids Res* 2008, **36**(Database issue):D88–D92, [<http://dx.doi.org/10.1093/nar/gkm964>].
- [187] Gershenzon NI, Stormo GD, Ioshikhes IP: **Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites.** *Nucleic Acids Res* 2005, **33**(7):2290–2301, [<http://dx.doi.org/10.1093/nar/gki519>].

BIBLIOGRAPHY

- [188] Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CWH, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.** *Nat Genet* 2006, **38**(4):431–440, [<http://dx.doi.org/10.1038/ng1760>].
- [189] Beissbarth T, Speed TP: **GStat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20**(9):1464–1465, [<http://dx.doi.org/10.1093/bioinformatics/bth088>].
- [190] Kishore S, Lubner S, Zavolan M: **Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression.** *Brief Funct Genomics* 2010, **9**(5-6):391–404, [<http://dx.doi.org/10.1093/bfgp/elq028>].
- [191] Kishore S, Stamm S: **The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.** *Science* 2006, **311**(5758):230–232, [<http://dx.doi.org/10.1126/science.1118265>].
- [192] Matera AG, Terns RM, Terns MP: **Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs.** *Nat Rev Mol Cell Biol* 2007, **8**(3):209–220, [<http://dx.doi.org/10.1038/nrm2124>].
- [193] Cavallé J, Bachellerie JP: **SnoRNA-guided ribose methylation of rRNA: structural features of the guide RNA duplex influencing the extent of the reaction.** *Nucleic Acids Res* 1998, **26**(7):1576–1587.
- [194] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**(12):1413–1415, [<http://dx.doi.org/10.1038/ng.259>].
- [195] Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**(7221):470–476, [<http://dx.doi.org/10.1038/nature07509>].
- [196] Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing.** *Gene* 2005, **344**:1–20, [<http://dx.doi.org/10.1016/j.gene.2004.10.022>].
- [197] Watkins NJ, Ségault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R: **A common core RNP structure**

- shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP. *Cell* 2000, **103**(3):457–466.
- [198] Steitz JA, Tycowski KT: **Small RNA chaperones for ribosome biogenesis.** *Science* 1995, **270**(5242):1626–1627.
- [199] Filipowicz W, Pogacić V: **Biogenesis of small nucleolar ribonucleoproteins.** *Curr Opin Cell Biol* 2002, **14**(3):319–327.
- [200] Butler MG, Hanchett JM, Thompson T: *Clinical findings and natural history of Prader–Willi syndrome.* Springer 2006. [chapter in Butler MG, Lee PDK, Whitman BY: Management of Prader–Willi Syndrome].
- [201] Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL: **Prader–Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster.** *Nat Genet* 2008, **40**(6):719–721, [<http://dx.doi.org/10.1038/ng.158>].
- [202] Doe CM, Relkovic D, Garfield AS, Dalley JW, Theobald DEH, Humby T, Wilkinson LS, Isles AR: **Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour.** *Hum Mol Genet* 2009, **18**(12):2140–2148, [<http://dx.doi.org/10.1093/hmg/ddp137>].
- [203] Nakatani J, Tamada K, Hatanaka F, Ise S, Ohta H, Inoue K, Tomonaga S, Watanabe Y, Chung YJ, Banerjee R, Iwamoto K, Kato T, Okazawa M, Yamauchi K, Tanda K, Takao K, Miyakawa T, Bradley A, Takumi T: **Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism.** *Cell* 2009, **137**(7):1235–1246, [<http://dx.doi.org/10.1016/j.cell.2009.04.024>].
- [204] Eiholzer U, Stutz K, Weinmann C, Torresani T, Molinari L, Prader A: **Low insulin, IGF-I and IGFBP-3 levels in children with Prader–Labhart–Willi syndrome.** *Eur J Pediatr* 1998, **157**(11):890–893.
- [205] Eiholzer U, Gisin R, Weinmann C, Kriemler S, Steinert H, Torresani T, Zachmann M, Prader A: **Treatment with human growth hormone in patients with Prader–Labhart–Willi syndrome reduces body fat and increases muscle mass and physical performance.** *Eur J Pediatr* 1998, **157**(5):368–377.
- [206] Cummings DE, Clement K, Purnell JQ, Vaisse C, Foster KE, Frayo RS, Schwartz MW, Basdevant A, Weigle DS: **Elevated plasma ghrelin levels**

BIBLIOGRAPHY

- in **Prader Willi syndrome**. *Nat Med* 2002, **8**(7):643–644, [<http://dx.doi.org/10.1038/nm0702-643>].
- [207] de Lind van Wijngaarden RFA, Otten BJ, Festen DAM, Joosten KFM, de Jong FH, Sweep FCGJ, Hokken-Koelega ACS: **High prevalence of central adrenal insufficiency in patients with Prader-Willi syndrome**. *J Clin Endocrinol Metab* 2008, **93**(5):1649–1654, [<http://dx.doi.org/10.1210/jc.2007-2294>].
- [208] Carrel AL, Lee PDK, Mogul HR: *Growth hormone and Prader-Willi syndrome*. Springer 2006. [chapter in Butler MG, Lee PDK, Whitman BY: Management of Prader-Willi Syndrome].
- [209] Stefan M, Ji H, Simmons RA, Cummings DE, Ahima RS, Friedman MI, Nicholls RD: **Hormonal and metabolic defects in a prader-willi syndrome mouse model with neonatal failure to thrive**. *Endocrinology* 2005, **146**(10):4377–4385, [<http://dx.doi.org/10.1210/en.2005-0371>].
- [210] Runte M, Hüttenhofer A, Gross S, Kiefmann M, Horsthemke B, Buiting K: **The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A**. *Hum Mol Genet* 2001, **10**(23):2687–2700.
- [211] Vitali P, Royo H, Marty V, Bortolin-Cavaillé ML, Cavaillé J: **Long nuclear-retained non-coding RNAs and allele-specific higher-order chromatin organization at imprinted snoRNA gene arrays**. *J Cell Sci* 2010, **123**(Pt 1):70–83, [<http://dx.doi.org/10.1242/jcs.054957>].
- [212] Kishore S, Stamm S: **Regulation of alternative splicing by snoRNAs**. *Cold Spring Harb Symp Quant Biol* 2006, **71**:329–334, [<http://dx.doi.org/10.1101/sqb.2006.71.024>].
- [213] Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes**. *RNA* 2004, **10**(10):1507–1517, [<http://dx.doi.org/10.1261/rna.5248604>].
- [214] Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J, Hüttenhofer A: **Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization**. *Proc Natl Acad Sci U S A* 2000, **97**(26):14311–14316, [<http://dx.doi.org/10.1073/pnas.250426397>].
- [215] Kishore S, Khanna A, Stamm S: **Rapid generation of splicing reporters with pSpliceExpress**. *Gene* 2008, **427**(1-2):104–110, [<http://dx.doi.org/10.1016/j.gene.2008.09.021>].

- [216] Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G: **A human snoRNA with microRNA-like functions.** *Mol Cell* 2008, **32**(4):519–528, [<http://dx.doi.org/10.1016/j.molcel.2008.10.017>].
- [217] Saraiya AA, Wang CC: **snoRNA, a novel precursor of microRNA in Giardia lamblia.** *PLoS Pathog* 2008, **4**(11):e1000224, [<http://dx.doi.org/10.1371/journal.ppat.1000224>].
- [218] Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ: **Human miRNA precursors with box H/ACA snoRNA features.** *PLoS Comput Biol* 2009, **5**(9):e1000507, [<http://dx.doi.org/10.1371/journal.pcbi.1000507>].
- [219] Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.** *Science* 2001, **294**(5543):858–862, [<http://dx.doi.org/10.1126/science.1065062>].
- [220] Lee KA, Bindereif A, Green MR: **A small-scale procedure for preparation of nuclear extracts that support efficient transcription and pre-mRNA splicing.** *Gene Anal Tech* 1988, **5**(2):22–31.
- [221] Zhang Z, Lotti F, Dittmar K, Younis I, Wan L, Kasim M, Dreyfuss G: **SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing.** *Cell* 2008, **133**(4):585–600, [<http://dx.doi.org/10.1016/j.cell.2008.03.031>].
- [222] Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB: **HITS-CLIP yields genome-wide insights into brain alternative RNA processing.** *Nature* 2008, **456**(7221):464–469, [<http://dx.doi.org/10.1038/nature07488>].
- [223] Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ: **An alternative-exon database and its statistical analysis.** *DNA Cell Biol* 2000, **19**(12):739–756, [<http://dx.doi.org/10.1089/104454900750058107>].
- [224] Vitali P, Basyuk E, Meur EL, Bertrand E, Muscatelli F, Cavaillé J, Huttenhofer A: **ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs.** *J Cell Biol* 2005, **169**(5):745–753, [<http://dx.doi.org/10.1083/jcb.200411129>].
- [225] Ding F, Prints Y, Dhar MS, Johnson DK, Garnacho-Montero C, Nicholls RD, Francke U: **Lack of Pwcr1/MBII-85 snoRNA is critical for neonatal lethality in Prader-Willi syndrome mouse models.** *Mamm Genome* 2005, **16**(6):424–431, [<http://dx.doi.org/10.1007/s00335-005-2460-2>].

BIBLIOGRAPHY

- [226] Kanber D, Giltay J, Wieczorek D, Zogel C, Hochstenbach R, Caliebe A, Kuechler A, Horsthemke B, Buiting K: **A paternal deletion of MKRN3, MAGEL2 and NDN does not result in Prader-Willi syndrome.** *Eur J Hum Genet* 2009, **17**(5):582–590, [<http://dx.doi.org/10.1038/ejhg.2008.232>].
- [227] Stoss O, Stoilov P, Hartmann AM, Nayler O, Stamm S: **The in vivo minigene approach to analyze tissue-specific splicing.** *Brain Res Brain Res Protoc* 1999, **4**(3):383–394.
- [228] Murray PG, Young LS: **Epstein-Barr virus infection: basis of malignancy and potential for therapy.** *Expert Rev Mol Med* 2001, **3**(28):1–20, [<http://dx.doi.org/doi:10.1017/S1462399401003842>].
- [229] Straathof KCM, Bollard CM, Rooney CM, Heslop HE: **Immunotherapy for Epstein-Barr virus-associated cancers in children.** *Oncologist* 2003, **8**:83–98.
- [230] Hislop AD, Taylor GS, Sauce D, Rickinson AB: **Cellular responses to viral infection in humans: lessons from Epstein-Barr virus.** *Annu Rev Immunol* 2007, **25**:587–617, [<http://dx.doi.org/10.1146/annurev.immunol.25.022106.141553>].
- [231] Rickson AB, Kieff E: *Epstein-Barr Virus*. Lippincott Williams & Wilkins 2006. [chapter in Fields BN, Knipe DM, Howley PM, Griffin DE: Fields' virology 5th edition].
- [232] Miller G, Lipman M: **Release of infectious Epstein-Barr virus by transformed marmoset leukocytes.** *Proc Natl Acad Sci U S A* 1973, **70**:190–194.
- [233] Kelly G, Bell A, Rickinson A: **Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2.** *Nat Med* 2002, **8**(10):1098–1104, [<http://dx.doi.org/10.1038/nm758>].
- [234] Lerner MR, Andrews NC, Miller G, Steitz JA: **Two small RNAs encoded by Epstein-Barr virus and complexed with protein are precipitated by antibodies from patients with systemic lupus erythematosus.** *Proc Natl Acad Sci U S A* 1981, **78**(2):805–809.
- [235] Fok V, Friend K, Steitz JA: **Epstein-Barr virus noncoding RNAs are confined to the nucleus, whereas their partner, the human La protein, undergoes nucleocytoplasmic shuttling.** *J Cell Biol* 2006, **173**(3):319–325, [<http://dx.doi.org/10.1083/jcb.200601026>].

- [236] Kim DN, Chae HS, Oh ST, Kang JH, Park CH, Park WS, Takada K, Lee JM, Lee WK, Lee SK: **Expression of viral microRNAs in Epstein-Barr virus-associated gastric carcinoma.** *J Virol* 2007, **81**(2):1033–1036, [<http://dx.doi.org/10.1128/JVI.02271-06>].
- [237] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, Vita GD, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Lauro RD, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T: **A mammalian microRNA expression atlas based on small RNA library sequencing.** *Cell* 2007, **129**(7):1401–1414, [<http://dx.doi.org/10.1016/j.cell.2007.04.040>].
- [238] Cai X, Schäfer A, Lu S, Bilello JP, Desrosiers RC, Edwards R, Raab-Traub N, Cullen BR: **Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed.** *PLoS Pathog* 2006, **2**(3):e23, [<http://dx.doi.org/10.1371/journal.ppat.0020023>].
- [239] Pfeffer S, Zavolan M, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T: **Identification of virus-encoded microRNAs.** *Science* 2004, **304**(5671):734–736, [<http://dx.doi.org/10.1126/science.1096781>].
- [240] Pfeffer S, Voinnet O: **Viruses, microRNAs and cancer.** *Oncogene* 2006, **25**(46):6211–6219, [<http://dx.doi.org/10.1038/sj.onc.1209915>].
- [241] Zhu JY, Pfuhl T, Motsch N, Barth S, Nicholls J, Grässer F, Meister G: **Identification of novel Epstein-Barr virus microRNA genes from nasopharyngeal carcinomas.** *J Virol* 2009, **83**(7):3333–3341, [<http://dx.doi.org/10.1128/JVI.01689-08>].
- [242] Mattick JS: **RNA regulation: a new genetics?** *Nat Rev Genet* 2004, **5**(4):316–323, [<http://dx.doi.org/10.1038/nrg1321>].
- [243] Sullivan CS: **New roles for large and small viral RNAs in evading host defences.** *Nat Rev Genet* 2008, **9**(7):503–507, [<http://dx.doi.org/10.1038/nrg2349>].
- [244] Romby P, Vandenesch F, Wagner EGH: **The role of RNAs in the regulation of virulence-gene expression.** *Curr Opin Microbiol* 2006, **9**(2):229–236, [<http://dx.doi.org/10.1016/j.mib.2006.02.005>].

BIBLIOGRAPHY

- [245] Amaral PP, Dinger ME, Mercer TR, Mattick JS: **The eukaryotic genome as an RNA machine.** *Science* 2008, **319**(5871):1787–1789, [<http://dx.doi.org/10.1126/science.1155472>].
- [246] Hüttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype?** *Trends Genet* 2005, **21**(5):289–297, [<http://dx.doi.org/10.1016/j.tig.2005.03.007>].
- [247] Samarsky DA, Fournier MJ, Singer RH, Bertrand E: **The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization.** *EMBO J* 1998, **17**(13):3747–3757, [<http://dx.doi.org/10.1093/emboj/17.13.3747>].
- [248] Hüttenhofer A, Schattner P: **The principles of guiding by RNA: chimeric RNA-protein enzymes.** *Nat Rev Genet* 2006, **7**(6):475–482, [<http://dx.doi.org/10.1038/nrg1855>].
- [249] Hüttenhofer A, Brosius J, Bachellerie JP: **RNomics: identification and function of small, non-messenger RNAs.** *Curr Opin Chem Biol* 2002, **6**(6):835–843.
- [250] Ganot P, Bortolin ML, Kiss T: **Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs.** *Cell* 1997, **89**(5):799–809.
- [251] Kiss AM, Jády BE, Bertrand E, Kiss T: **Human box H/ACA pseudouridylation guide RNA machinery.** *Mol Cell Biol* 2004, **24**(13):5797–5807, [<http://dx.doi.org/10.1128/MCB.24.13.5797-5807.2004>].
- [252] Bachellerie JP, Cavallé J, Hüttenhofer A: **The expanding snoRNA world.** *Biochimie* 2002, **84**(8):775–790.
- [253] Hüttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J: **RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse.** *EMBO J* 2001, **20**(11):2943–2953, [<http://dx.doi.org/10.1093/emboj/20.11.2943>].
- [254] Kiss T: **Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs.** *EMBO J* 2001, **20**(14):3617–3622, [<http://dx.doi.org/10.1093/emboj/20.14.3617>].
- [255] Frech B, Zimmer-Strobl U, Suentzenich KO, Pavlish O, Lenoir GM, Bornkamm GW, Mueller-Lantzsch N: **Identification of Epstein-Barr virus terminal protein 1 (TP1) in extracts of four lymphoid cell lines, expression in**

- insect cells, and detection of antibodies in human sera. *J Virol* 1990, **64**(6):2759–2767.
- [256] Mrázek J, Kreutmayer SB, Grässer FA, Polacek N, Hüttenhofer A: **Subtractive hybridization identifies novel differentially expressed ncRNA species in EBV-infected human B cells.** *Nucleic Acids Res* 2007, **35**(10):e73, [<http://dx.doi.org/10.1093/nar/gkm244>].
- [257] Hutzinger R, Mrázek J, Vorwerk S, Hüttenhofer A: **NcRNA-microchip analysis: A novel approach to identify differential expression of non-coding RNAs.** *RNA Biol* 2010, **7**(5):81–90.
- [258] Jöchel C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H, Hüttenhofer A: **Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis.** *Nucleic Acids Res* 2008, **36**(8):2677–2689, [<http://dx.doi.org/10.1093/nar/gkn123>].
- [259] Aspegren A, Hinas A, Larsson P, Larsson A, Söderbom F: **Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development.** *Nucleic Acids Res* 2004, **32**(15):4646–4656, [<http://dx.doi.org/10.1093/nar/gkh804>].
- [260] Edwards RH, Marquitz AR, Raab-Traub N: **Epstein-Barr virus BART microRNAs are produced from a large intron prior to splicing.** *J Virol* 2008, **82**(18):9094–9106, [<http://dx.doi.org/10.1128/JVI.00785-08>].
- [261] Chen H, Huang J, Wu FY, Liao G, Hutt-Fletcher L, Hayward SD: **Regulation of expression of the Epstein-Barr virus BamHI-A rightward transcripts.** *J Virol* 2005, **79**(3):1724–1733, [<http://dx.doi.org/10.1128/JVI.79.3.1724-1733.2005>].
- [262] Dragon F, Gallagher JEG, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF, Baserga SJ: **A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis.** *Nature* 2002, **417**(6892):967–970, [<http://dx.doi.org/10.1038/nature00769>].
- [263] Michienzi A, Cagnon L, Bahner I, Rossi JJ: **Ribozyme-mediated inhibition of HIV 1 suggests nucleolar trafficking of HIV-1 RNA.** *Proc Natl Acad Sci U S A* 2000, **97**(16):8955–8960.
- [264] Kiss T: **Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions.** *Cell* 2002, **109**(2):145–148.

BIBLIOGRAPHY

- [265] Feederle R, Kost M, Baumann M, Janz A, Drouet E, Hammerschmidt W, Delecluse HJ: **The Epstein-Barr virus lytic program is controlled by the co-operative functions of two transactivators.** *EMBO J* 2000, **19**(12):3080–3089, [<http://dx.doi.org/10.1093/emboj/19.12.3080>].
- [266] Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P: **Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets.** *Nucleic Acids Res* 2007, **35**(3):962–974, [<http://dx.doi.org/10.1093/nar/gkl1096>].
- [267] Maden BE, Corbett ME, Heeney PA, Pugh K, Ajuh PM: **Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA.** *Biochimie* 1995, **77**(1-2):22–29.
- [268] Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T: **Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs.** *Cell* 1996, **85**(7):1077–1088.
- [269] Maroney PA, Chamnongpol S, Souret F, Nilsen TW: **A rapid, quantitative assay for direct detection of microRNAs and other small RNAs using splinted ligation.** *RNA* 2007, **13**(6):930–936, [<http://dx.doi.org/10.1261/rna.518107>].
- [270] Yekta S, Shih IH, Bartel DP: **MicroRNA-directed cleavage of HOXB8 mRNA.** *Science* 2004, **304**(5670):594–596, [<http://dx.doi.org/10.1126/science.1097434>].
- [271] Llave C, Xie Z, Kasschau KD, Carrington JC: **Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.** *Science* 2002, **297**(5589):2053–2056, [<http://dx.doi.org/10.1126/science.1076311>].
- [272] Holzerlandt R, Orengo C, Kellam P, Albà MM: **Identification of new herpesvirus gene homologs in the human genome.** *Genome Res* 2002, **12**(11):1739–1748, [<http://dx.doi.org/10.1101/gr.334302>].
- [273] Omer AD, Lowe TM, Russell AG, Ehardt H, Eddy SR, Dennis PP: **Homologs of small nucleolar RNAs in Archaea.** *Science* 2000, **288**(5465):517–522.
- [274] Cook HL, Lytle JR, Mischo HE, Li MJ, Rossi JJ, Silva DP, Desrosiers RC, Steitz JA: **Small nuclear RNAs encoded by Herpesvirus saimiri up-regulate the expression of genes linked to T cell activation in virally transformed T cells.** *Curr Biol* 2005, **15**(10):974–979, [<http://dx.doi.org/10.1016/j.cub.2005.04.034>].

- [275] Cook HL, Mischo HE, Steitz JA: **The Herpesvirus saimiri small nuclear RNAs recruit AU-rich element-binding proteins but do not alter host AU-rich element-containing mRNA levels in virally transformed T cells.** *Mol Cell Biol* 2004, **24**(10):4522–4533.
- [276] Huang GM, Jarmołowski A, Struck JC, Fournier MJ: **Accumulation of U14 small nuclear RNA in *Saccharomyces cerevisiae* requires box C, box D, and a 5', 3' terminal stem.** *Mol Cell Biol* 1992, **12**(10):4456–4463.
- [277] Jarmołowski A, Zagorski J, Li HV, Fournier MJ: **Identification of essential elements in U14 RNA of *Saccharomyces cerevisiae*.** *EMBO J* 1990, **9**(13):4503–4509.
- [278] Pratt ZL, Kuzembayeva M, Sengupta S, Sugden B: **The microRNAs of Epstein-Barr Virus are expressed at dramatically differing levels among cell lines.** *Virology* 2009, **386**(2):387–397, [<http://dx.doi.org/10.1016/j.virol.2009.01.006>].
- [279] Choy EYW, Siu KL, Kok KH, Lung RWM, Tsang CM, To KF, Kwong DLW, Tsao SW, Jin DY: **An Epstein-Barr virus-encoded microRNA targets PUMA to promote host cell survival.** *J Exp Med* 2008, **205**(11):2551–2560, [<http://dx.doi.org/10.1084/jem.20072581>].
- [280] Barth S, Pfuhl T, Mamiani A, Ehses C, Roemer K, Kremmer E, Jäker C, Höck J, Meister G, Grässer FA: **Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5.** *Nucleic Acids Res* 2008, **36**(2):666–675, [<http://dx.doi.org/10.1093/nar/gkm1080>].
- [281] Lo AKF, To KF, Lo KW, Lung RWM, Hui JWY, Liao G, Hayward SD: **Modulation of LMP1 protein expression by EBV-encoded microRNAs.** *Proc Natl Acad Sci U S A* 2007, **104**(41):16164–16169, [<http://dx.doi.org/10.1073/pnas.0702896104>].
- [282] Xing L, Kieff E: **Epstein-Barr virus BHRF1 micro- and stable RNAs during latency III and after induction of replication.** *J Virol* 2007, **81**(18):9967–9975, [<http://dx.doi.org/10.1128/JVI.02244-06>].
- [283] Bernstein E, Caudy AA, Hammond SM, Hannon GJ: **Role for a bidentate ribonuclease in the initiation step of RNA interference.** *Nature* 2001, **409**(6818):363–366, [<http://dx.doi.org/10.1038/35053110>].
- [284] Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R: **Human RISC couples microRNA biogenesis and posttranscriptional gene silencing.** *Cell* 2005, **123**(4):631–640, [<http://dx.doi.org/10.1016/j.cell.2005.10.022>].