Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Diplom-Biologe Tobias Hartmut Bauer
born in Limburg a. d. Lahn, Germany

Oral examination: October 25 2011

Deciphering transcriptional regulation in cancer cells and development of a new method to identify key transcriptional regulators and their target genes

Referees: PD Dr. Rainer König
Prof. Dr. Manfred Schwab

To Cathy, my children and my parents,
And in loving memory of Rev. Dr. Robert D. Johnson
For their sincerity, commitment, and their love

“We are caught in an inescapable network of mutuality,
Tied in a single garment of destiny.
Whatever affects one directly, affects all indirectly.”

Martin Luther King Jr.

“This is our moment here at the crossroads of time
We hope our children carry our dreams down the line
They are the vintage, what kind of life will they live?
Is this a curse or a blessing that we give?”

Billy Joel (Two Thousand Years)

Erklärung

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 13. September 2011 .....

(Tobias Bauer)

Acknowledgements

This doctoral thesis is the result of my research spanning the past five years of my life and would not have been possible without the aid, support, and contribution of many people to whom I wish to express my gratitude here.

I want to thank PD Dr. Rainer König and Prof. Dr. Roland Eils for scientific guidance, discussion and funding of my work, as well as for allowing me to participate in a congenial work group providing an inspiring professional environment.
I feel very grateful to Prof. Dr. Peter Angel and PD Dr. Jochen Hess for opening opportunities to fruitful collaboration, scientific discussion, as well as for funding throughout the completion of my thesis.

Special thanks go to Prof. Dr. Manfred Schwab and PD Dr. Frank Westermann, for being part of my thesis advisory committee and refereeing my thesis, for scientific discussion, cooperation, and *pro rata* funding in the beginning of my work. I also want to thank Prof. Dr. Ulrike Müller for her willingness to referee my thesis disputation.
Prof. Dr. Achim Tresch, another member of my thesis advisory committee, has provided me with helpful insights and input on questions on scientific and statistic problems, as has Dr. Benedikt Brors.

I would like to acknowledge all my collaboration partners that I had the pleasure to work with in excellently interactivity on many projects in- and outside of dkfz: PD Dr. Matthias Fischer, Dr. Kai-Oliver Henrich, and Dr. Astrid Riehl.

Many members of the department of theoretical bioinformatics have been very supportive and made it easy to feel welcome. I have enjoyed meeting people with large-spread expertise, their professionalism, the internationality of this group, and -plainly expressed- the inherent coolness and honest friendship of some people. Some members have particularly earned my gratefulness: Manuela Schäfer and Corinna Sprengart for their vital administrative support and personal kindness; Dr. Gunnar Schramm, Moritz Aschoff, and Il-Han Kim for personal friendship and scientific discussion; Richa Batra, Dr. Phillip Hundeshagen, Marti Bernardo, Dr. Yara Reis, Daniela Reis, Dr. Anna-Lena Kranz, Jens Keienburg, Kitiporn Plaimas, and Apichat Suratanee for all the fun (not only working) together; Dr. Karl-Heinz Groß and Rolf Kabbe for excellent IT and infrastructure support.

I want to thank all the people that I have had the honor to become friends with that have helped me to keep a healthy work-life-balance and make the leisure time count. Among these people are Michael Finkenzeller, Christian Schiller, Thomas Klein, Oliver Kaltheier, Dr. Markus Moosmeier, Thomas Holz, Dr. Florian Sonntag, Roman Jowanowitsch, Dr. Lars Krüger, Florian Kress, Mike Fischer, Matthias Brehm, Alexander Radke, Jörg Schönweiß, Florian Müller, Daniel Schulz, and many others.

Nobel laureate Prof. Dr. Harald Zur Hausen has inspired me personally and scientifically, and I appreciated his informality and respectfulness on the occasion where we met.

My loving parents, my siblings, my grandmother, as well as my godparents and American host parents have always given me a warm and loving shelter and I feel privileged to honor them here.

Finally, I want to express to my wife Cathy and to my children Florian and Christian that they are everything I could have ever wished for. Their unconditional love and their invaluable trust, support and loyalty are the most beautiful gift that I will always keep in my heart.

List of publications

I) Own publications presented in this thesis and my contributions:

Publication I

Fischer M, **Bauer T**, Oberthuer A, Hero B, Theissen J, Ehrich M, Spitz R, Eils R, Westermann F, Brors B, König R, and Berthold F. **Integrated genomic profiling identifies two distinct molecular subtypes with divergent outcome in neuroblastoma with loss of chromosome 11q.** *Oncogene* 2009, 29(6):865-875.

The portion of my contribution to this publication makes up about 35%. I conducted most of the bioinformatics analyses. My initial studies of the global gene expression profiles raised the formulation of the central hypothesis of the publication. I implemented the application of the centroid distance method in R statistical software as advised by Dr. Benedikt Brors. This approach was conducted among several others and eventually delivered the most reliable results. (Other attempts included integration of Support Vector Machines and multidimensional scaling and led to similar results as those presented here.) Furthermore, I conducted the SAM analysis on my own account to judge the differences between clinico-genetic subgroups in more detail. Finally, I decided to relate the differential expression to chromosomal location and I used Fisher's exact tests to estimate enrichments of genes located on 11q. The revised tumorigenesis model was jointly developed mainly by PD Dr. Matthias Fischer and me, also considering the input of other co-authors of the publication.

Publication II

Henrich KO, **Bauer T**, Schulte J, Ehemann V, Deubzer H, Gogolin S, Muth D, Fischer M, Benner A, König R, Schwab M, and Westermann F. **CAMTA1, a 1p36 Tumor Suppressor Candidate, Inhibits Growth and Activates Differentiation Programs in Neuroblastoma Cells.** *Cancer Res.* 2011 Apr 15;71(8):3142-51. Epub 2011 Mar 8.

In this project, I analyzed time series of gene expression profiles from CAMTA1-induced cell lines. In addition to normalizing and clustering the data, I used both standard and more intuitive gene filtering to select genes that reflected best the impact of the experimental setup. I performed follow-up analyses and determined descriptive GO terms separately for up- and down-regulated genes upon CAMTA1 induction, enabling iterative lab experiments for validation of selected candidate genes. The portion of my contribution to this publication makes up about 20%.

Publication III

Westermann F, Muth D, Benner A, **Bauer T**, Henrich KO, Oberthuer A, Brors B, Beissbarth T, Vandesompele J, Pattyn F, Hero B, König R, Fischer M, and Schwab M. **Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas.** *Genome Biol* 2008, 9(10):R150.

I experimented with two different approaches to identify putative transcription factor binding sites (TFBSs) in DNA sequences: position weighted matrix (PWM) scans and consensus sequence matching. I analyzed promoter sequences for a large number of genes, applied the two techniques and optimized parameters for the PWM scans. Subsequently, I implemented and performed enrichment analyses for binding motifs within selected gene sets. The portion of my contribution to this publication makes up roughly 10%.

Publication IV

Riehl A, **Bauer T**, Brors B, Busch H, Mark R, Németh J, Gebhardt C, Bierhaus A, Nawroth P, Eils R, König R, Angel P, and Hess J. **Identification of the Rage-dependent gene regulatory network in a mouse model of skin inflammation.** *BMC Genomics* 2010 Oct 5;11:537.

I conducted a clustering analysis subsequent to the analysis of differentially expressed genes by Dr. Benedikt Brors. I extended the PWM scans and enrichment analysis described in the previous study to a large number of transcription factors (TFs) to predict comprehensive associations of TFs to gene subsets and therefore provided hypotheses for further lab investigation by cooperation partners. My contribution to this study was ≥20%.

Publication V

Bauer T, Eils R, and König R. **RIP: The regulatory interaction predictor - a machine learning based approach for predicting target genes of transcription factors.** *Bioinformatics.* 2011 Aug 15;27(16):2239-47. Epub 2011 Jun 20.

This work was fully conducted by me. Conception and design of the algorithm was done by PD Dr. Rainer König and me. Critical advice and discussion were contributed from Prof. Dr. Roland Eils. I drafted the manuscript and incorporated suggestions for improvements from PD Dr. Rainer König and Prof. Dr. Roland Eils. My contribution was 90%.

II) Other publications

Publication VI

Heinemann T, Kramer S, Velten L, Kranz AL, Bauer T, Faura MB, König R, Keienburg J, Eils R, and Iwamoto N. **Design of Specific Mammalian Promoters by in silico Prediction.** BBF RFC 43, 2009-10-30, iGEM team 2009, http://hdl.handle.net/1721.1/49520.

This technical protocol publication resulted from the project of team Heidelberg for the prestigious international Genetically Engineered Machine competition (iGEM). Our team won the second place among 110 competitors at the iGEM jamboree, October 31 to November 2, 2009, at Massachusetts Institute of Technology, USA. I aided the team by guiding and supporting the students as one of the bioinformatics supervisors, and co-developed the published workflow strategy.

Publication VII

Krupp M, Maass T, Marquardt JU, Staib F, **Bauer T**, König R, Biesterfeld S, Galle PR, Tresch A, Teufel A. **The functional cancer map: A systems-level synopsis of genetic deregulation in cancer.** *BMC Med Genomics.* 2011 Jun 30;4:53.

My contribution to this study was the selection and normalization of >1500 microarray gene expression profiles from Stanford Microarray database, and curation of the clinical and the genomic annotation. The revised data served as the basis for the analyses by our cooperation partners.

III) Poster presentations

Bauer T, König R, Schramm G, Durchdewald M, Nemeth J, Riehl A, Hess J, and Eils R. **Identifying key players of transcriptional changes in tumor regulatory networks.** DKFZ PhD Retreat, July 19-24 2007, Weil der Stadt, Germany

Bauer T, Schramm G, Brors B, Nemeth J, Riehl A, Durchdewald M, Hess J, Eils R, and König R. **Identifying Key Players of Transcriptional Changes in Tumor Regulatory Networks.** 6th NGFN Meeting, November 10-11 2007, Heidelberg, Germany

Bauer T, Schramm G, Brors B, Nemeth J, Riehl A, Durchdewald M, Hess J, Westermann F, Eils R, and König R. **Identifying Key Players of Transcriptional Changes in Tumor Regulatory Networks.** DKFZ Graduate Forum, October 30 2008, Heidelberg, Germany

Bauer T, Eils R, and König R. **Inferring human regulatory networks from transcription factor binding site predictions and correlation meta-analysis.** Systems Genomics 2010, September 29 - October 1, 2010, Heidelberg, Germany

IV) Patent applications

Reis Y, Richter D, Reichenzeller M, Iwamoto N, Zhu C, Uckelmann H, Eils R, Keienburg J, Hundeshagen P, Bernardo M, **Bauer T**, Kranz Al, Koenig R, Heinemann T, Mugahid D, Velten L, Haas S, Hiller C, Bartoschek M, Rademacher A, Meyer H, Kraemer S, and Zhao Bingging. **Model-guided Random Assembly PCR for the synthesis of eukaryotic promoters.**, European Patent EP 09173197.6, International Publication Number WO 2011/045374 A1 , DKFZ Erfindungsmeldung P-903

This patent application resulted from the project of team Heidelberg for the iGEM competition 2009 (see Publication VI above).

.....

(read and confirmed by PD Dr. Rainer König, primary reviewer)

Zusammenfassung

Krebszellen akkumulieren im Laufe der Karzinogenese genetische Veränderung. Diese Veränderungen können in Größenordnungen von Punktmutation bis hin zu großen chromosomalen Aberrationen entstehen. Nach unserem heutigen Verständnis werden dadurch essentielle genetische Programme dysreguliert, die im Normalzustand unkontrollierte Zellteilung und -wachstum verhindern. Transkriptionsfaktoren (TF) sind Schlüsselproteine der Genregulation und werden häufig mit genetisch bedingten Krankheiten, z.B. MYCN in Neuroblastomen (NB), in Verbindung gebracht. Der Erforschung der Genregulation im Allgemeinen wie im Speziellen kommt daher eine zentrale Rolle in der Krebsforschung zu und sie steht auch im Zentrum meiner Arbeit.

Nach einem Karzinogenesemodell von NB ohne *MYCN*-Amplifikation stehen Mutationen des Chromosomenarms 11q (11q-CNA) im Verdacht, die Tumorentwicklung maßgeblich zu beeinflussen. Unsere Genexpressionsanalysen von 11q-CNA in NB mit unterschiedlichem klinischen Verlauf ergaben ein verbessertes Karzinogenesemodell. Genexpressionsprofile von NB mit negativem klinischen Verlauf unterschieden sich drastisch zwischen Tumoren mit und ohne 11q-CNA, wohingegen die 11q-CNA bei NB mit günstigem Verlauf offensichtlich durch einen unbekannten Mechanismus kompensiert wird. Das TF-kodierende Gen *CAMTA1* befindet sich auf der chromosomalen Region 1p, die in Neuroblastomen häufig deletiert ist. *In vitro*-Experimente mit ektopischer CAMTA1-Induktion lieferten CAMTA1-regulierte Gene unterschiedlicher Genexpressionsprofile, die durch Anreicherungstests funktionell mit Zellzyklussteuerung und neuronaler Differenzierung assoziiert und anschließend experimentell bestätigt werden konnten. Die demnach für *CAMTA1* vermutete Rolle als Tumorsuppressorgen in NB wurde durch *in vivo*-Analysen in Mäusen bestätigt. Weiterhin untersuchten wir die Wirkung von MYC und MYCN in NB ohne *MYCN*-Amplifikation und fanden dabei heraus, dass diese TF auch in diesen Tumoren eine Reihe gemeinsamer Zielgene in Abhängigkeit ihrer eigenen Genexpression maßgeblich regulieren. Promoteranalysen und Chromatin-Immunopräzipitation lieferten dabei weitere Belege für die Regulation der bestimmten Zielgene durch MYC/MYCN. Die genomweite Anwendung von Promoteranalysen und Anreicherungstests in Genexpressionsdaten von Mausmodellen ermöglichte uns die Vorhersage von Ziel-TF des Rage-Signalwegs. Im Labor konnten E2f1 und E2f4 als Komponenten des Rage-abhängigen genregulatorischen Netzwerkes validiert werden.

Schließlich konnten wir die gesammelten praktischen Erfahrungen mit Genexpressionsdaten einsetzen, um eine neue Maschinenlernmethode zur präzisen Bestimmung von TF-Zielgenen im Menschen zu entwickeln. Dazu wurde eine genomweite Korrelationsmetanalyse von 4064 Genexpressionsprofilen ausgewertet und zusammen mit Promoteranalysen von TF-Bindestellen sowie bereits bekannten regulatorischen Interaktionen zwischen TF und Zielgenen verknüpft. Unsere Methode übertraf die Leistung vergleichbarer Methoden und verbesserte Nachteile herkömmlicher Algorithmen speziell für höhere Eukaryoten, insbesondere die häufig fälschlicherweise angenommene Kopplung der mRNA-Expression von TF und ihren Zielgenen. Unsere Entwicklung ist frei verfügbar als Softwarepaket mit vielfältigen Anwendungen wie z.B. die Identifikation von Schlüssel-TF in einer Vielzahl zellulärer Systeme (wie z.B. Krebszellen).

Abstract

Cancer cells accumulate genetic changes during carcinogenesis. The dimension of these changes range from point mutations to large chromosomal aberrations. It has been widely accepted that essential genetic programs are thereby dysregulated that normally would prevent uncontrolled cellular division and growth. Transcription factors (TFs) are key proteins of gene regulation and are frequently associated with genetic pathologies, e.g. MYCN in neuroblastomas (NBs). Research on gene regulation -in general or condition-specific- thus is a central aspect in cancer research, and it is also the focus of my work.

In a carcinogenesis model of NBs without *MYCN*-amplification, mutations of chromosome 11q (11q-CNA) are suspected to critically influence tumor development. We were able to refine this model by means of gene expression analysis on 11q-CNA in NBs with different clinical outcome. Gene expression profiles of NBs with unfavorable progression differed significantly between tumors with and without 11q-CNA, whereas 11q-CNA in NBs with favorable outcome is apparently compensated by a yet unknown mechanism. The TF-encoding gene *CAMTA1* is located on the chromosomal region 1p, which is frequently deleted in NBs. *In vitro* experiments with ectopic induction of CAMTA1 yielded CAMTA1-regulated genes with different gene expression profiles that were functionally associated by enrichment analyses with cell cycle regulation and neuronal differentiation. The suggested role of *CAMTA1* as a tumor suppressor gene was confirmed by additional *in vivo* experiments. Furthermore, we studied the effect of MYC and MYCN in NBs without *MYCN*-amplification and found that these TF also strongly regulate a large number of common target genes according to their own gene expression in these tumors. Promoter analyses and chromatin immunoprecipitation additionally supported the regulation of the determined target genes by MYC/MYCN. The genome-wide application of promoter and enrichment analyses on gene expression data from mouse models enabled us to predict target TFs of Rage signaling. E2f1 and E2f4 were validated experimentally as components of the Rage-dependent gene regulatory network.

Finally, we used our experience from gene expression analysis to develop a novel machine learning method to precisely predict TF target gene relationships in human. We combined results from a genome-wide correlation meta-analysis on 4064 microarray gene expression profiles and promoter analyses on TF binding sites with known regulatory interactions between TFs and target genes in our approach. Our method outperformed other comparable methods in human, as we improved shortcomings of other algorithms specifically for higher eukaryotes, in particular the frequently (erroneously) assumed correlation between the mRNA expression of TFs and their target genes. We made our method freely available as a software package with multiple applications like the identification of key TFs in a multiplicity of cellular systems (e.g. cancer cells).

Table of contents

1. MOTIVATION AND BACKGROUND 1

1.1. Motivation 1
1.2. Transcription 2
1.3. Transcription factors and their function as regulators of gene expression 2
1.4. Transcription factor binding motifs and position weighted matrices 4
1.5. Quantifying gene expression 5
1.6. Gene regulation and (challenges in) gene regulatory network reconstruction 6
1.7. Machine learning and support vector machines 7
1.8. Cross-validation 9

2. INTEGRATED GENOMIC PROFILING IDENTIFIES TWO DISTINCT MOLECULAR SUBTYPES WITH DIVERGENT OUTCOME IN NEUROBLASTOMAS WITH LOSS OF CHROMOSOME 11Q 11

2.1. Motivation 11
2.2. Main Results 12
2.2.1. Neuroblastomas with loss of 11q fall into two prognostically distinct subgroups by gene expression-based classification 12
2.2.2. In neuroblastomas with loss of 11q, global gene expression patterns differ between patients with favorable and unfavorable outcome 12
2.2.3. Differential gene expression between clinico-genetic neuroblastoma subgroups 13
2.2.4. Relating differential expression between clinco-genetic neuroblastoma subgroups to chromosomal location on 11q 14

3. GENE EXPRESSION PROFILING CONFIRMS TUMOR SUPPRESSOR EFFECTS OF TRANSCRIPTION FACTOR CAMTA1 IN NEUROBLASTOMA CELLS 16

3.1. Motivation 16
3.2. Main results 17
3.2.1. CAMTA1 suppresses growth in neuroblastoma cell lines and is associated with neuronal differentiation 17
3.2.2. Identification of genes responsive to CAMTA1 induction 17
3.2.3. CAMTA1-induced genes influence cellular processes of neuronal development, calcium ion transport, proliferation and metabolism 18
3.2.4. Characterizing CAMTA1-repressed genes 20
3.2.5. Observed CAMTA1 functionality is confirmed in an independent cell line. 20

4. PROMOTER MOTIF ANALYSES IDENTIFY COMMON MYCN/MYC BINDING SITES OF GENES THAT ARE UP-REGULATED UPON MYCN INDUCTION IN NEUROBLASTOMA CELLS AND THAT ARE ASSOCIATED WITH POOR OUTCOME OF NEUROBLASTOMAS WITHOUT MYCN AMPLIFICATION 21

4.1. Motivation 21
4.2. Main results 22
4.2.1. MYC and MYCN are inversely correlated in neuroblastoma subtypes 22
4.2.2. MYC repression upon MYCN induction in neuroblastoma cells and definition of MYC/MYCN regulated genes 22
4.2.3. Validation of potential MYC/MYCN target genes in silico and by ChIP 23
4.2.4. MYC/MYCN activity in MYC NA neuroblastoma subtypes 24
4.2.5. High MYC/MYCN target gene expression is associated with poor overall survival 24

5. RECONSTRUCTION OF THE RAGE-DEPENDENT GENE REGULATORY NETWORK IN A MOUSE MODEL OF SKIN INFLAMMATION FROM GENE EXPRESSION PROFILES AND POSITION WEIGHTED MATRIX SCANS 24

5.1. Motivation 25
5.2. Main results 25
5.2.1. Rage-dependent gene expression profiles exhibit two temporal phases in response to TPA stimulation 25
5.2.2. Rage-dependent differential expression after TPA stimulation 27
5.2.3. Predicting TFs of the Rage-dependent gene regulatory network 27
5.2.4. Expression of E2f TFs upon induction of Rage signaling by TPA stimulation 29

6. RIP: THE REGULATORY INTERACTION PREDICTOR – MACHINE LEARNING BASED APPROACH FOR PREDICTING TARGET GENES OF TFS 29

6.1. Motivation 30
6.2. Main Results 31
6.2.1. Training machine learning classifiers to predict TF target genes – the workflow 31
6.2.2. Genes with correlated gene expression share biological processes 33
6.2.3. Correlated genes are frequently regulated by common TFs 35
6.2.4. Promoters of known target genes contain TFBS enrichments of corresponding TFs 37
6.2.5. Classifier performance 37
6.2.6. Inferring new RIs 38
6.2.7. Applying the inferred regulatory interactions to a microarray gene expression study: identifying TFs responsive to interferon α 38
6.2.8. RIs predicted for a large number of TFs are supported by pathway analysis and an independent database 41
6.2.9. RIP software package 41

7. SHORT SUMMARIES OF MAIN CONCLUSIONS 42

7.1. Neuroblastomas with genomic 11q aberrations fall into two distinct subtypes depending on the clinical outcome: a revised model of for tumorigenesis (Publication I) 42
7.2. CAMTA1 TF acts as a tumor suppressor in neuroblastomas and affects cell cycle progression and neuronal differentiation (Publication II) 44
7.3. MYC and MYCN affect distinct gene expression profiles in different neuroblastoma subtypes without *MYCN* amplification (Publication III) 45
7.4. The Rage-dependent regulatory network in a tumor-promoting inflammatory context (Publication IV) 46
7.5. RIP – a powerful tool to predict regulatory interactions with multiple applications (Publication V) 46

8. BIBLIOGRAPHY 49

9. OWN PUBLICATIONS 54

9.1. Publication I: Fischer *et al.* Oncogene 2009, 29(6):865-875. 55
9.2. Publication II: Henrich *et al*. Cancer Res. 2011 Apr 15;71(8):3142-51. Epub 2011 Mar 8. 66
9.3. Publication III: Westermann *et al*. Genome Biol 2008, 9(10):R150. 76
9.4. Publication IV: Riehl *et al*. BMC Genomics 2010 Oct 5;11:537. 90
9.5. Publication V: Bauer *et al*. Bioinformatics. 2011 Aug 15;27(16):2239-47. Epub 2011 Jun 20. 103

1. Motivation and background

1.1. Motivation

Cancer is a disease that acquires its malignancy by genetic mutability and dysregulation of essential cellular genetic programs. My work throughout my thesis has been dedicated to identifying such genetic programs and their transcriptional regulators in the context of carcinogenesis.

Neuroblastomas are tumors of the early childhood with tremendously diverging clinical phenotypes. Copy number alteration (CNA) of chromosome 11q occurs frequently and has been proposed as a clinical marker for poor outcome. The objective of my first project was to implement and apply bioinformatics methods to elucidate the relationship between 11q CNA and clinical phenotype employing gene expression profiles. I continued working on Neuroblastoma cells and analyzed time-series gene expression data to examine the potential tumor suppressor functionality of the transcription factor (TF) calmodulin binding transcription activator 1 (CAMTA1). I first filtered CAMTA1 responsive genes and then determined gene clusters that are induced or repressed at different time-points upon CAMTA1 induction. I then described the biology of these genes by means of gene ontology categories. Furthermore, I established the application of transcription factor binding site (TFBS) scans to understand the involvement of transcriptional regulators v-myc myelocytomatosis viral oncogene homolog (MYC) and MYCN in Neuroblastoma cells. In the fourth study, I used this technique comprehensively to identify key TFs involved in the regulatory network of activated Rage signaling in mice models of skin inflammation. From these projects I learned about the vast potential of transcriptional analyses in cancer research and the need to better understand transcriptional gene regulation. This led me to create a new bioinformatics method to reconstruct regulatory interactions between TFs and target genes on a genome-wide scale in human. The machine learning approach I developed integrated information from three different aspects of gene regulation: 1) a meta-analysis of correlation of co-regulated genes reflected in a broad range of conditions spanning thousands of available microarray profiles, 2) putative TFBSs obtained from genome-wide position weight matrix scans, and 3) descriptors of network topology derived from a repository of known regulatory interactions between TFs and target genes.

My motivation can be summarized as employing gene expression profiling to engineer powerful tools to investigate the biology of tumors and elucidate changes in their gene regulatory networks. In particular, neuroblastomas were in the focus of most projects. All studies I contributed to have been published in well-ranking journals.

1.2. Transcription

A central aspect of cellular complexity and dynamics is transcription. Why is it central? Every somatic cell of a living organism inherits the whole genome containing the blueprint on how to reproduce and maintain itself. To carry that into effect, the encoded genes first need to be transformed into active products, and transcription is a paramount step in this process.

In human, there are approximately 22 000 genes that carry the information for the development of the organism, including cellular differentiation into several hundred distinct cell types. This high degree of complexity is achieved by thoroughly tuned genetic programs. In the post-genomic era [4], extensive research has been focused on investigating the relationships between the genome, cell physiology, development, and pathogenesis [5].

Somatic cells develop into distinct tissue types and usually remain in their differentiated state, yet they all contain the same genome. The mechanisms yielding different cellular phenotypes are implemented by proteins. While many processes are essential to most cell types and require common proteins such as enzymes involved in DNA repair or replication, RNA polymerases, ribosomal proteins, cytoskeleton proteins, enzymes of the central metabolism, or proteins building the chromosomal structure, cellular differentiation requires different sets of proteins to be synthesized in different cell types. Evidently, specific genes are expressed in some cell types but not in others. In consequence, any cell phenotype may be regarded as a result of the activity of specific gene sets. Although the vast majority of protein encoding genes guarantees the viability of many cell types and remains rather constantly expressed (housekeeping genes), a considerable number of genes are expressed tissue- or condition-specifically. These specific genes are particularly of interest in research dedicated to carcinogenesis.

1.3. Transcription factors and their function as regulators of gene expression

Human gene expression is a progress of enormous complexity. Up to date, at least six different levels of gene expression control have been identified:

1) transcriptional control (DNA transcribed into pre-mRNA)
2) RNA processing (pre-mRNA splicing and modifications)
3) RNA transport and localization (export from mRNA the nucleus into the cytosol)
4) mRNA degradation (mRNA stability, RNA silencing)
5) translation (rate of mRNA translation by ribosomes)
6) protein activity (degradation, post-translational modification, location of protein)

Several classes of proteins function as regulators of gene expression at the transcriptional level and are hence referred to as transcription factors (TFs). Transcriptional control comprises several mechanisms (to which I count epigenetic modifications), and precedes other levels of regulation. It is therefore the only level at which production of intermediate molecules that are not required can be prevented. As TFs influence gene expression via transcriptional control, they are essential components in determining cellular phenotypes.

There may be more than 2000 human genes encoding TFs [6]. TF proteins contain DNA binding domains or structures with which they attach to short specific nucleotide sequences in control regions of a gene. A further level of complexity is introduced into the variability and dynamics of TF binding by the ability of many TF proteins to form oligomers, most frequently homo- or hetero-dimers. TFs are thought to administrate gene regulation by either promoting or hindering the accessibility of the transcription start site (TSS) of a transcribed *locus* by RNA polymerases and other proteins of the transcription initiation complex, thereby either inducing or repressing transcription. This accessibility can be provided by modifying the chromatin structure, or by bending the DNA to form a loop where TFs are in physical contact with mediator proteins (co-activators) interacting with the transcriptional pre-initiation complex at the core promoter ("looping model"; e.g. [7]).

Eukaryotic protein-encoding genes are transcribed by RNA polymerase II. A model of transcriptional regulatory elements of human RNA polymerase II genes is depicted in Figure 1.1. Several TFs are required to initiate the transcription of all these genes. These general TFs are required for transcription of all RNA polymerase II genes and bind at the compact core promoter region, which spans about 60 base pairs (bp) [8] and characteristically encompasses an element called the TATA-box (because of its 5'-TATAA-3' consensus sequence). The TATA-box defines the TSS and is located 25 to 30 bp upstream of the TSS. Several transcription regulatory elements (TREs) that are bound by specific TFs are located proximal to the core promoter. Together with the core promoter, they form principal regions for transcriptional gene regulation and are often subject to epigenetic silencing via methylation. TREs are labeled enhancers or silencers when they mediate induction or repression of gene expression, respectively. However, these labels may be condition-dependent. A typical TRE is stretched over 500 bp and contains 10 TFBSs for at least three different TFs, most frequently two activators and one repressor [8]. Additionally, activators and repressors may compete for the same TFBS. A typical transcribed *locus* in the *Drosophila* genome encompasses 10 kilobases (kb) of DNA [9] around the TSS, which also includes insulators, i.e. negative-regulatory, *cis*-acting elements that limit the advance of condensed chromatin or confine the activity of enhancers to specific genes [10]. In mammalian genomes, enhancers, silencers, or insulators can be scattered over distances of more than

100 kb [8]. It is widely assumed that the combination of regulatory elements accounts for the required specificity of transcription and suffices to achieve the complexity of higher eukaryotes [5,8,11]. Evidently, the fate of a cell depends on the concerted action of TFs and in consequence TFs often play key roles in pathogenesis, e.g. as tumor suppressor genes or proto-oncogenes in cancer.

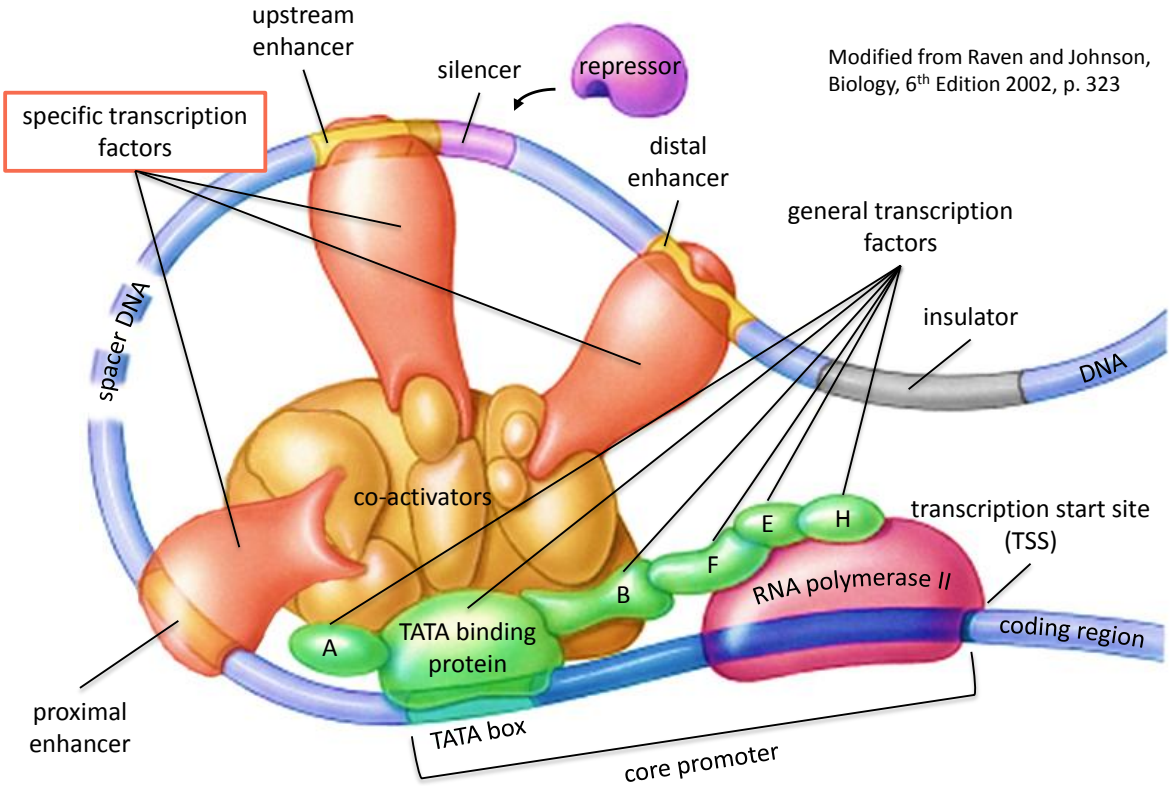


Figure 1.1 | **Gene regulatory DNA region of a human protein-encoding gene.** General TFs (green) interacting with co-activators (brown) and RNA polymerase II (dark pink) are bound to the core promoter region adjacent to the TSS. Proximal, upstream and distal enhancers (yellow) and silencers (purple) are occupied by specific TFs (red and purple). An insulator element (grey) prevents interactions with regulatory elements of other genes in proximity. Figure modified from [12].

1.4. Transcription factor binding motifs and position weighted matrices

Specific TFs bind specific sequence motifs of about 5-10 nucleotides. There are several experimental techniques to determine the sequences of transcription factor binding sites (TFBSs), such as DNase footprinting or chromatin immunoprecipitation (ChIP) assays. The combination of ChIP with high-throughput methods, like nucleotide microarrays (ChIP-chip) or deep sequencing (ChIP-seq), yields thousands of putative TFBSs for a given TF. These sequences can be aligned to identify consensus sequences reflecting TF binding specificity, and to infer position weighted matrices (PWMs, see Figure 1.2). PWMs are probabilistic

representations of each base at each position of a TFBS. So far, PWMs of hundreds of TFs have been determined, which can be used as templates by *in silico* motif scans to predict putative TFBSs. Enrichment analyses of predicted TFBSs can associate sets of genes with regulatory TFs and may yield plausible mechanistic insights on transcriptional regulation in human [13-15]. On the downside, PWM scans can not predict the actual binding of a TF to a putative TFBS and thus come along with high false positive rates [16].

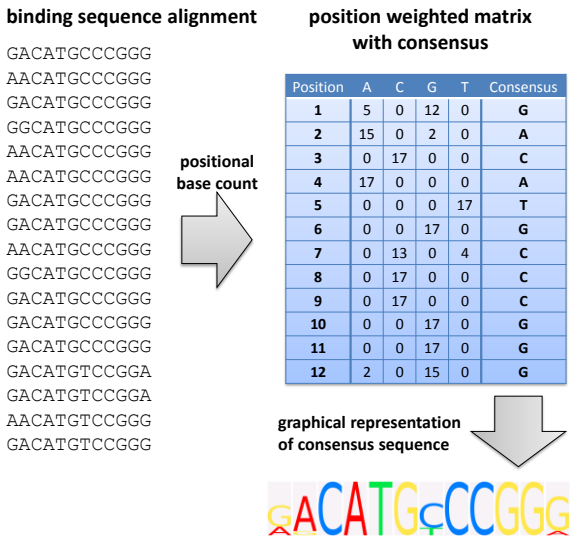


Figure 1.2 | **Position weighted matrix (PWM) generation.** Experimentally determined binding sequences are aligned and the incidence of each DNA base at every position of the alignment is counted. The counts in the PWM thus provide the relative base frequencies, which are symbolized by the corresponding letter size in a graphical representation of the PWM (high frequency = large letter). (Graphical PWM representation produced by Transfac database [2].)

1.5. Quantifying gene expression

Currently, assessing gene expression on a large scale remains difficult as it is a very dynamic process involving many levels of control. Moreover, high-throughput techniques and data are only available for DNA methylation, mRNA and regulatory RNA abundance at present, but not for protein quantity, activity and modification. Because of this and the fact that most high-throughput experiments investigate mRNA levels, scientists often generalize to *gene expression* profiles when referring to *mRNA expression* measurements of a specimen, and I will also use this term accordingly.

Modern deep sequencing approaches are developing rapidly and are being performed in large quantities as I am writing this thesis, but for the last one or two decades microarrays have dominated high-throughput gene expression analyses, producing hundreds of thousands of gene expression profiles. Advanced microarrays measure the relative abundance of thousands of different mRNAs by hybridization of fluorescence-labeled cRNA or cDNA with synthesized complementary oligonucleotides (probe) fixed on a chip or bead. The fluorescence intensity at each probe is then detected and translated into relative expression values of the corresponding mRNA. There are several different facilities and

commercial companies producing microarrays that differ in material and protocol, but they all rely on the same hybridization principle and mostly yield reliable and reproducible results.

Gene expression analyses employing microarrays have been the key to many discoveries analyzing the concerted interaction between genes in biomedical research. They have enabled scientists to track genome-wide alterations in developmental processes (e.g. [17]), knockout, knockdown, or other perturbation experiments, cancers and other diseases, and they are applied in diagnostics and prognosis or in screenings for potential drug targets. Furthermore, microarray studies have provided insight into the modular organization, structural characteristics, and temporal dynamics of biological networks [18], and they can be used to analyze functional or metabolic pathways, to learn classifiers to predict gene expression [17], and to reconstruct biological networks, such as protein-protein interactions or gene regulatory networks (GRNs). In the studies I am presenting here, my main focus has been on GRNs, so I will further describe them in the next section.

1.6. Gene regulation and (challenges in) gene regulatory network reconstruction

Identification of key regulatory elements (such as TFs) of a genetically induced disease, and clarifying how they interact and cooperate provides direct potential access points for treatment strategies. However, the reconstruction of GRNs, particularly in higher eukaryotes, remains a major task of systems biology. Approaches may be crudely separated into two categories based on the number of components considered in the model [19], which I will use in the following. Another way to differentiate the models might be by their scope and application [18], as some models are intended to describe and predict gene expression dynamics, whereas others are focused on defining the network components or edges.

Small-scale approaches have been designed to produce detailed mechanistic and quantitative models of a single or a few regulatory circuits and are based on rules derived from thermodynamics (e.g. Hill functions), or kinetic models. They require accurate measurements of many parameters (e.g. TF activity, DNA binding dynamics and affinity, TF oligomerization, TF cooperativeness and interactivity), and a detailed knowledge of the regulatory DNA regions of the network components (usually genes). Such approaches have been successfully applied to prokaryotes (e.g. [20,21]), where the system tractability and accumulated knowledge are sufficient [19]. In higher eukaryotes, the demands of the methods and the much higher complexity of gene regulation have rendered these approaches impractical in the past.

Large-scale approaches aim to reconstruct GRNs on a genome-wide scale. The methods range from Boolean networks or probabilistic Bayesian approaches [11,14,19,22]

over linear models [23,24] to differential equations [25]. Unlike small-scale models, where the components and edges of the network are known *a priori*, large-scale models infer the components and edges *de novo*. The abundance of quantitative mRNA expression data, as well as the availability of the genome sequence and an increasing number of experiments measuring interactions between proteins and DNA are being exploited to infer eukaryotic GRNs [19]. Most techniques are focused on defining regulatory modules or components of GRNs that fit a given input of mRNA data for a specific cell-type, condition, or a TF, but they are usually applied to more simple model organisms, e.g. for *Escherichia coli* [26], or *Saccharomyces cerevisiae* [27-29]. Large-scale approaches have been rarely applied in mammals and have been limited to individual or few regulatory TFs (e.g. [30]). Besides the lack of high-throughput protein data of all kinds, the major challenges of accurate GRN reconstruction in human are the inability to accurately identify the gene regulatory region [19], and (in my opinion) the need for simplification, leading to invalid generalization of a biological concept used to infer regulatory interactions (RIs). For example, some methods assume a direct relationship between the gradient of TF mRNA levels and their target genes, which has worked efficiently e.g. for the reconstruction of MYC target genes in human B-cells [31,32], but the concept is hardly generalizable (even though it has been shown for MYC TFs as I will describe in the presented studies), as the activity of many human TFs is controlled in post-translational events [33].

Essentially, there is a great need for improved approaches to reconstruct human GRNs, in particular for methods that can comprehensively identify RIs between TFs and target genes on a genome-wide scale with high precision and recall.

1.7. Machine learning and support vector machines

Computational analyses using machine learning algorithms have become a valuable and indispensable aid in biomedical research. Their applications in combination with gene expression data include diagnosis and prognosis in disease, eliciting clusters, functional pathways or regulatory modules, reconstruction of GRNs, or dissecting and predicting the behavior of systematic cellular transcriptional changes, to name only a few.

Many questions in research can be defined as classification problems and therefore can be addressed by machine learning classifiers. In general, these classifiers can be separated into unsupervised and supervised methods. Unsupervised classifiers can be applied to observations (samples) without *a priori* knowledge of group memberships (classes). The standard unsupervised method (in respect to gene expression analysis) is hierarchical clustering. In contrast, a supervised classifier is trained to recognize patterns in empirical data that distinguish observations (=samples) of known classes and subsequently applies the “learned” patterns to predict the class affiliation of unseen observations. Several

machine learning classifiers have been developed that are kernel-based, and among them, Support Vector Machines (SVMs) [34,35] have become popular because of their wide-range applicability (to data with linear and non-linear class boundaries), computational feasibility, excellent performance, and generalization capability [36]. A kernel can be thought of as a function for measuring similarity [37]. SVMs utilize kernels based on dot products of vectors. As I have made extensive use of SVMs in my main project (chapter 6, page 29) I will highlight some of their principles in the following.

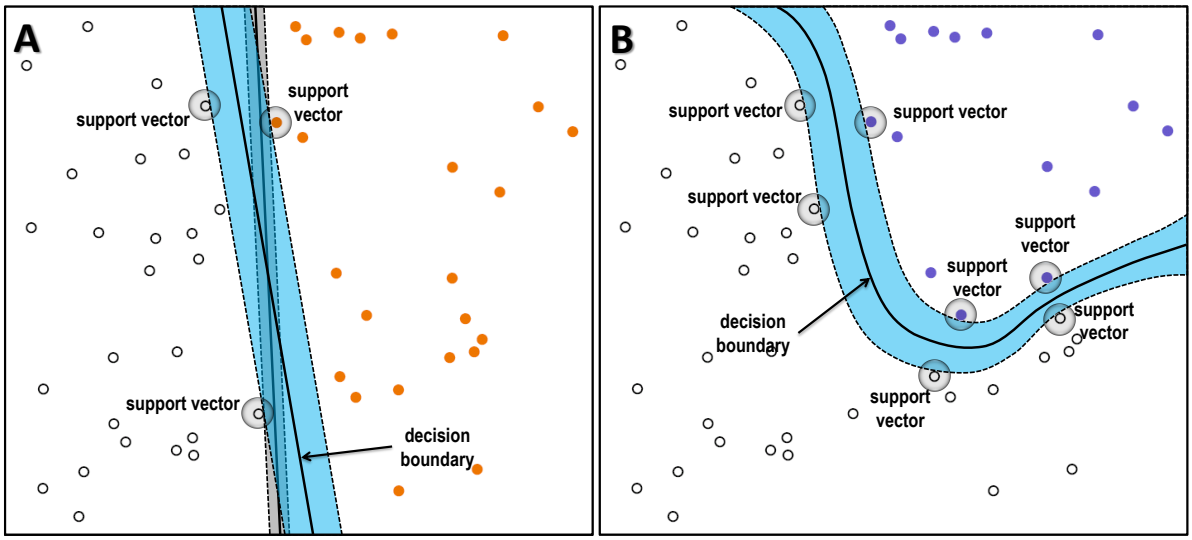


Figure 1.3 | **Decision boundaries of SVMs.** Vectors from different classes are represented by black circles *versus* orange or blue dots. In the linearly separable case (**A**), an SVM finds the discriminating hyperplane (=decision boundary) with the largest distance (=margin; colored light blue) to the closest vectors (=support vectors, highlighted in grey). Another decision boundary with a thinner margin (and therefore not optimal) is marked in grey. In the linearly non-separable case (**B**), the SVM can employ a non-linear kernel that maps the data into a higher dimensional feature space within which the data becomes linearly separable. The light blue margin illustrates how the margin around the decision boundary could be represented in the original vector space.

The idea behind SVMs can be described rather simply: For a training set of vectors (i.e. observations or samples), the algorithm searches for an optimal hyperplane (i.e. a decision boundary) in the vector space that separates the vectors according to their class. The hyperplane is optimal in that the width of the margin between the closest vectors of different classes defining the hyperplane (called support vectors) is maximized, as illustrated in Figure 1.3A. Vapnik showed that maximizing the margin also maximizes the generalizability of the yielded separation [38]. In most cases however, such a *linear* decision boundary does not exist. A *kernel trick* [39] is then applied, which is also known as kernel substitution, because the dot product used by linear kernel SVMs is replaced by a different

adequate kernel function (e.g. a non-linear Gaussian radial basis function (RBF)) that maps the training vector space into a Euclidean space of higher dimensionality, in which the vectors become separable (Figure 1.4B).

When transformed into the original vector space, the hyperplane can be regarded as a (non-linear) hypersurface (Figure 1.3B and Figure 1.4C). The scaling can be computationally expensive, so the *trick* is to use a mapping function that does not need explicit calculation, but can be computed for all high-dimensional vectors within the original space. Because most natural data sets are noisy and non-separable (leading to poor generalization ability of an optimally fit decision boundary), additional slack variables are introduced that relax the constraints and further facilitate the computation of the optimized solution for the SVM kernel. Allowing (penalized) misclassification errors during the training process (soft-margin classifiers) also helps to avoid overfitting of the SVM.

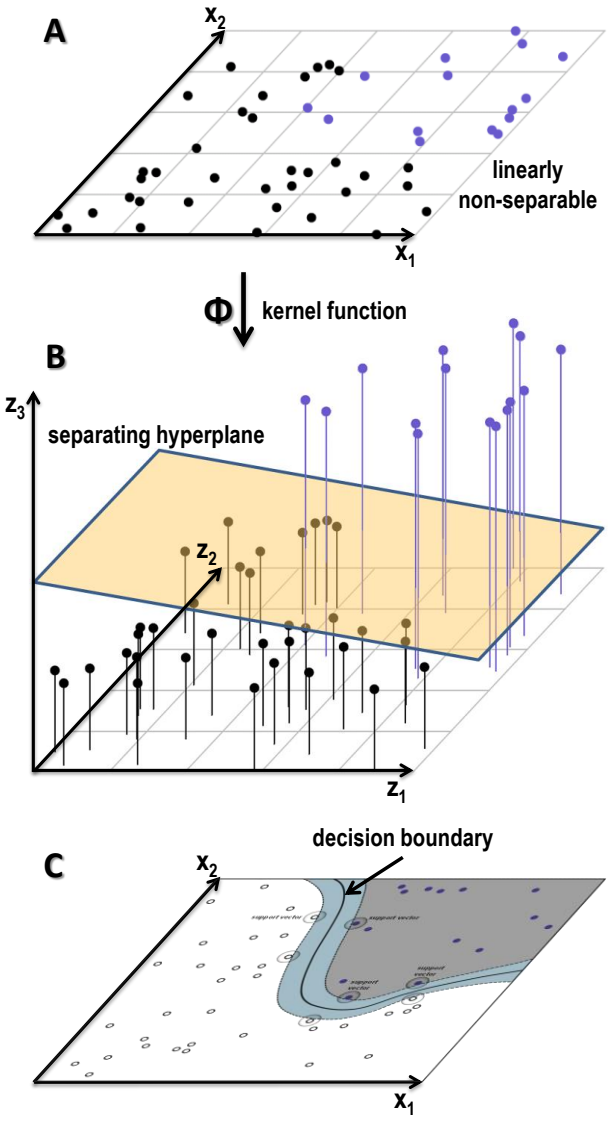


Figure 1.4 | **Kernel trick.** As the data set (two classes, black versus blue points) is not linearly separable in the vector space (**A**), the classifier kernel function is substituted with a (non-linear) function Φ that maps the vectors into a feature space of higher dimensionally (**B**), where they may become distinguishable by a hyperplane (marked light brown). This hyperplane may be represented in the original vector space as shown in **C**.

1.8. Cross-validation

Several cross-validation procedures have been established to estimate generalization capacity and potential overfitting of a classifier. Cross-validation is integrated into the classifier building process and -if applied correctly (!)- enables finding optimal parameters for a classifier, e.g. C (misclassification penalty) and γ (corresponding to the variance of a Gaussian RBF kernel) in SVMs. Additionally, it may provide an accurate estimate of a classifier's performance that is useful to judge its utility, in particular if

the amount of data is a limiting factor. In a basic k-fold cross-validation, the training data S is partitioned into k sets ($S_{1,\ldots,k}$), to train k classifiers where each time a different set is left apart from the training for testing, so that each training set of S serves as the test set exactly once. The parameters of the classifier with the best performance (accuracy) on its corresponding test set are then chosen to build a classifier with the whole data S. A problem of this strategy is that the classifier performance is likely to be over-optimistic because it is estimated from the same data that was used for parameter optimization.

A better estimate of the classifier performance is obtained by using a test set that is completely independent of the classifier building. A good approach thereto is nested cross-validation (illustrated in Figure 1.5), where an m-fold inner cross-validation loop is nested into a k-fold outer cross-validation loop. In the outer loop, the data S is partitioned randomly into k sets ($S_{1,\ldots,k}$), where in iteration i (i=1, …, k) set S_i is left apart as a test set and the remaining sets are passed to the inner loop as T_i (T_i=$S \setminus S_i$). In the inner loop, T_i is again divided randomly into m subsets T_{ij} (j=1, …, m), where in each inner iteration j subset T_{ij} is defined as validation set V_{ij} and the remaining data becomes training set U_{ij} (U_{ij}=$T_i \setminus V_{ij}$). A classifier is then trained on each U_{ij} and its parameters are optimized to the best accuracy on validation set V_{ij}, while the actual accuracy of the classifier is estimated with the independent test set S_i. The nested cross-validation is computationally demanding because of the high number of trained classifiers (k x m), but it provides a more conservative estimate of the average performance of a classifier.

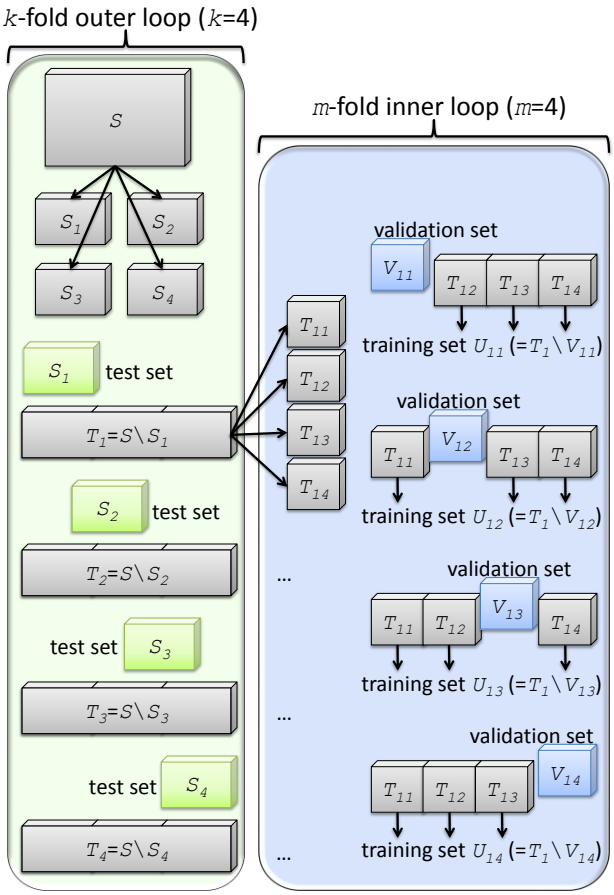


Figure 1.5 | **Nested cross-validation.** A four-fold inner cross-validation loop is executed within a four-fold outer cross-validation loop. For details on the procedure and symbols, please refer to the main text (section 1.8).

Further cross-validation approaches make use of bootstrapping. The sampling can be stratified in case of skewed class distributions, and it is possible to integrate these sampling methods into nested cross-validation.

2. Integrated genomic profiling identifies two distinct molecular subtypes with divergent outcome in neuroblastomas with loss of chromosome 11q

This study has been published [40]:

- Publication I

  Fischer M, **Bauer T**, Oberthuer A, Hero B, Theissen J, Ehrich M, Spitz R, Eils R, Westermann F, Brors B, König R, and Berthold F: **Integrated genomic profiling identifies two distinct molecular subtypes with divergent outcome in neuroblastoma with loss of chromosome 11q.** *Oncogene* 2009, 29(6):865-875.

2.1. Motivation

Neuroblastomas cover a spectrum of heterogeneous phenotypes to an extent that is rarely observed in other cancer types. This is reflected on the genome level by numerous non-random chromosomal copy number aberrations (CNAs) [41]. Neuroblastoma may therefore constitute a good tumor model to study the impact of CNAs on the transcriptome and to relate them to different clinico-genetic phenotypes.

Some frequently observed CNAs are thought to be critical events in tumorigenesis. The effects of *MYCN* oncogene amplification have been extensively studied and the results dictate that there is a direct impact of elevated gene dosage on gene expression levels [42]. In contrast, it is less clear how low-level copy number gains (<fivefold) or hemizygous losses of large chromosomal regions affect gene expression in neuroblastoma. Loss of 11q has a prevalence of nearly 30% and correlates highly with an unfavorable outcome [43,44]. Therefore, it has been proposed to be included in clinical trials as a stratified prognostic marker [45,46].

MYCN amplification, which is observed in ~20% of neuroblastomas, and 11q CNAs are almost mutually exclusive events, suggesting that they constitute genetically distinct subgroups [1,47,48]. This indicates that tumorigenesis of these two phenotypes is characterized by different cellular mechanisms. Whereas the effect of *MYCN* amplification on the transcriptome has been well investigated, the influence of 11q CNAs on global gene expression is poorly understood.

In this project, we worked in cooperation with the clinician PD Dr. Matthias Fischer and the team of Prof. Dr. Frank Berthold from the Cologne university children's hospital, department of pediatric oncology. We followed a strategy that incorporated results from various bioinformatics approaches, clinical information, cytogenetic characteristics and promoter methylation analyses to elucidate the relationship between neuroblastoma

phenotypes with and without 11q CNAs and their transcriptome. In particular, I performed several analyses to elucidate how 11q CNA is linked to neuroblastoma tumorigenesis.

2.2. Main Results

2.2.1. Neuroblastomas with loss of 11q fall into two prognostically distinct subgroups by gene expression-based classification

Previously, cooperation partners from our groups published a prediction analysis of microarrays (PAM)-based classifier that uses the gene expression patterns of 144 genes to accurately predict the outcome of neuroblastomas [49]. When applied to a subset of 61 specimens with 11q CNAs that were not included in the training set, the classifier separated the patients into two distinct groups: 20 patients were predicted to be favorable and 41 patients unfavorable. Event-free survival at five years was significantly different between these two predicted groups (P=0.001), even after exclusion of six patients with *MYCN* amplification (P=0.005). These results strongly indicated that neuroblastomas with loss of 11q fall into two distinct groups with divergent clinical course based upon the gene expression patterns of 144 selected genes.

2.2.2. In neuroblastomas with loss of 11q, global gene expression patterns differ between patients with favorable and unfavorable outcome

The next step was to see how the clinical outcome of neuroblastomas with 11q CNAs relates to their *global* expression patterns. To avoid bias on gene expression by other CNAs, we excluded neuroblastomas with *MYCN* amplification and/or loss of 1p, leaving a selection of 110 specimens for the analysis. These were sorted into four defined clinico-genetic subgroups according to 11q status ("normal" *versus* "del11q" - deletion/imbalance) and clinical outcome ("fav" - at least two years event-free survival without cytotoxic treatment *versus* "unfav" - metastatic or multiple loco-regional progression/relapse or death of disease). Figure 2.1 illustrates this division into subgroups. In unsupervised analyses (principal component analysis and hierarchical clustering, data shown in the publication only), the tumors clustered together primarily according to their clinical phenotype. Notably, favorable neuroblastomas were inseparable by 11q status (normal/fav and del11q/fav), whereas unfavorable neuroblastomas with loss of 11q (del11q/unfav) and unfavorable neuroblastomas without loss of 11q (normal/unfav) formed individual clusters each.

To objectify these observations, I implemented a centroid distance analysis (adapted from [50]). Overall differences in gene expression can be judged by this approach. A centroid is defined as the vector of mean gene expression values over all patients of a subgroup. Euclidean distances were calculated pairwise between the centroids of different clinico-

genetic neuroblastoma subgroups. The significance of the centroid distances between the subgroups was assessed by a permutation analysis.

We obtained highly significant differences (P<0.001) between all pairs of subgroups except between del11q/fav and normal/fav (P=0.19, see Figure 2.2). Taken together, these results provided evidence that favorable neuroblastomas with and without 11q CNAs do not differ in their overall gene expression.

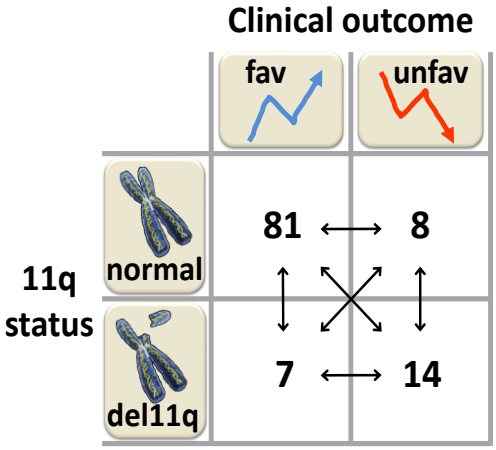


Figure 2.1 | **Neuroblastoma subgroups defined by clinical outcome and status of chromosome 11q.** A favorable outcome ("fav") was defined by event-free survival of >2 years in absence of cytotoxic treatment whereas unfavorable outcome ("unfav") meant malignant progression or death of disease. CNA of 11q was abbreviated "del11q" as opposed to the "normal" karyotype. The numbers give the count of specimens in each subgroup. The arrows indicate pairwise comparisons of subgroups that were conducted.

Table 2.1 | **Differentially expressed genes between selected Neuroblastoma subgroups and subsets located on chromosome 11q.**

	normal/fav *vs.* normal/unfav	normal/fav *vs.* del11q/fav	normal/unfav *vs.* del11q/unfav	del11q/fav *vs.* del11q/unfav
SAM	1187 genes	2 genes	64 genes	282 genes
Genes on 11q	38 (3%)	0	27 (42%)	10 (4%)

2.2.3. Differential gene expression between clinico-genetic neuroblastoma subgroups

I conducted significance analysis of microarrays (SAM; Figure 2.3 and Table 2.1) to determine differentially expressed genes between the four clinico-genetic subgroups. There were only two differentially expressed genes between favorable tumors with and without 11q loss (P<0.05), which underlined our previous observations. In contrast, much larger numbers ranging from 64 to 2470 differentially expressed genes resulted from comparing all other subgroup pairs (Figure 2.3A).

We then took a closer look at the lists of differentially expressed genes. The question was if there was an overlap of differentially expressed genes between favorable and unfavorable tumors with respect to their 11q status that might account for a malignant

phenotype. In fact, we counted 100 genes that were common in the comparison of subgroups with normal 11q and the comparison of subgroups with 11q CNAs, making up 35% of the differing genes in the latter case (Figure 2.3B and Table 2.1). With only one exception, all genes from the overlap were down-regulated in the unfavorable subgroups. This emphasizes their potential biological relevance for malignant progression of neuroblastomas independent of the 11q status. Several of these genes have been previously suggested to correlate with adverse outcome of neuroblastomas, such as FYN oncogene related to SRC, FGR, YES (*FYN)* [51], microtubule-associated protein 7 *(MAP7)* [49,52], and *CAMTA1* [53]. These results point to a common mechanism that promotes malignant progression of unfavorable neuroblastomas with and without 11q CNAs.

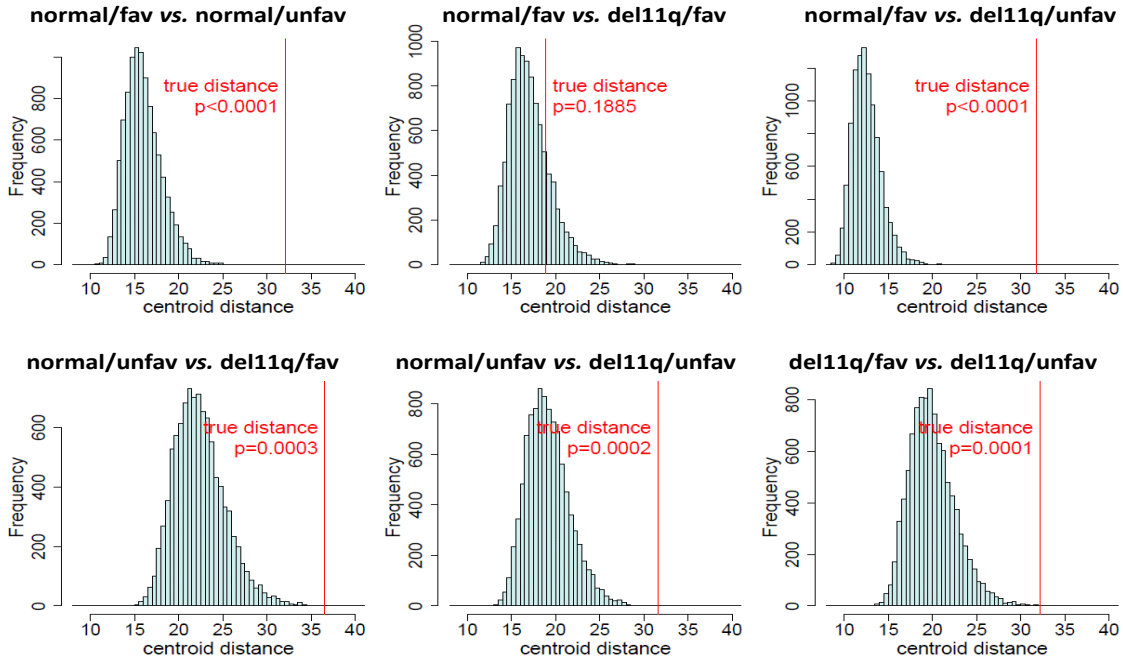


Figure 2.2 | **Analysis of pairwise centroid distances between favorable and unfavorable tumors with and without 11q aberration.** The histograms indicate the distributions of centroid distances from permutation analyses. Original centroid distances are marked by red lines and assigned p-values. In contrast to the pair normal/fav *vs.* del11q/fav, all pairs exhibited highly significant differences.

2.2.4. Relating differential expression between clinco-genetic neuroblastoma subgroups to chromosomal location on 11q

We analyzed the number of differentially expressed genes located on chromosome 11q (see Table 2.1). Neither of the two genes differentially expressed between del11q/fav and normal/fav tumors were located on 11q. In contrast, 42% (27/64) of genes identified by comparing unfavorable neuroblastomas with and without loss of 11q were annotated to

11q. This is a highly significant enrichment (P<1e-18), and as expected, most of these transcripts (24 of27) had lower transcription levels in the subgroup with loss of 11q. We did not find such enrichment when comparing the two subgroups with loss of 11q: only 10 of 282 differentially expressed genes were encoded on 11q (P=0.40), whereas comparison of the two subgroup with normal 11q status yielded a slight under-representation of 11q genes (38 of 1187 genes, P=0.01). These findings indicate that the expression of genes located on chromosome 11q is affected by 11q CNAs when a malignant phenotype has been developed, whereas in favorable neuroblastomas, loss of 11q appears to have no pronounced impact on the expression of these genes. An enrichment of differentially expressed genes located on 11q was also found between del11q/unfav and normal/fav tumors (176 of 2470 genes, 7%, P<1E-05).

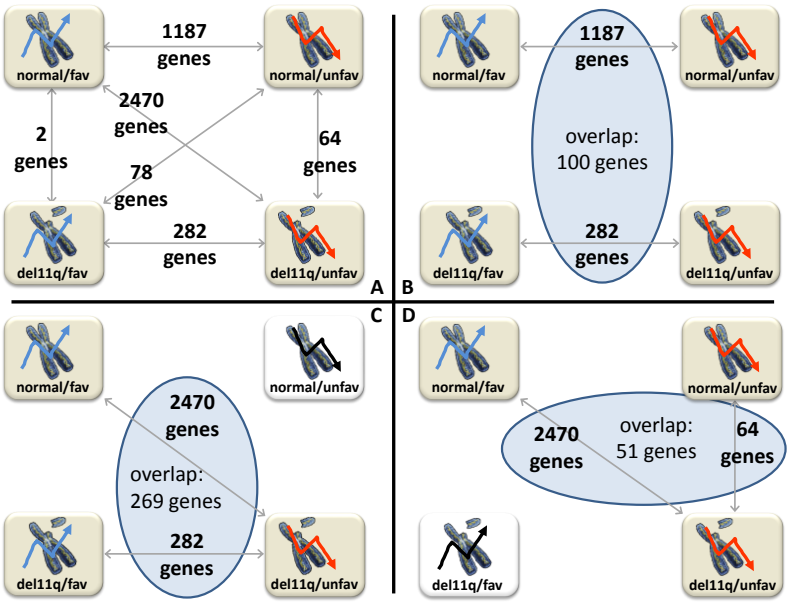


Figure 2.3 | **Results obtained by significance analysis of microarrays (SAM). A** provides an overview of the number of differentially expressed genes in all pairwise comparisons. **B**, **C**, and **D** provide further details on the overlap of differentially expressed genes in selected subgroup comparisons (see main text).

When comparing the two favorable groups (del11q/fav and normal/fav) against del11q/unfav (Figure 2.3C), the large overlap of 269 differentially expressed genes with similar up- or down-regulation supported our hypothesis that the favorable tumors actually formed a homogenous subgroup together. Moreover, 51 differentially expressed genes were common among comparisons of normal/fav and normal/unfav to del11q/unfav tumors, corresponding to 80% of differentially expressed genes between normal/unfav and del11q/unfav (Figure 2.3D). All 11q genes from the intersection were consistently downregulated in del11q/unfav neuroblastomas. With respect to the large number of differentially expressed genes between the two subgroups of normal 11q status, these findings further indicated that loss of 11q affects genes encoded on chromosome 11q particularly in unfavorable neuroblastomas.

Array-based comparative genomic hybridization (aCGH) analysis was conducted and excluded a potential bias in our findings as a result of different sizes of 11q deletions (or additional 17q CNAs) in the favorable and unfavorable subgroups. Similarly, the clinico-genetic subgroups did not differ in the CpG-methylation status of promoter regions of 10 differentially expressed genes located on 11q. Although the status of only a sample of genes on 11q was determined, these findings argue against a general involvement of differential CpG-methylation on the down-regulation of genes located on 11q in del11q/unfav neuroblastoma.

3. Gene expression profiling confirms tumor suppressor effects of transcription factor CAMTA1 in neuroblastoma cells

This study has been published [54]:

- Publication II

  Henrich KO, **Bauer T**, Schulte J, Ehemann V, Deubzer H, Gogolin S, Muth D, Fischer M, Benner A, König R, Schwab M, and Westermann F: **CAMTA1, a 1p36 Tumor Suppressor Candidate, Inhibits Growth and Activates Differentiation Programs in Neuroblastoma Cells.** *Cancer Res.* 2011 Apr 15;71(8):3142-51. Epub 2011 Mar 8.

3.1. Motivation

Similar to 11q CNA (see previous chapter), deletion of distal 1p is another recurrent chromosomal aberration in neuroblastoma. About 30% of neuroblastomas carry 1p mutations and, in contrast to tumors with 11q CNAs, they often coincide with genomic amplification of MYCN, an oncogenic transcription factor. Besides neuroblastoma, loss of heterozygosity (LOH) of distal 1p has been observed in other malignant tumors including breast cancer [55], colon cancer [56], glioma [57], and melanoma [58]. Consequently, genomic regions that are prone to frequent genomic mutation associated with cancers are in focus when searching for potential tumor suppressor genes and oncogenes. This can be done by finding the smallest region of consistent deletion. For example, analysis of frequently deleted regions on 11q yielded *CADM1*, a gene encoding a cell adhesion molecule that is associated with several clinical markers of poor outcome in neuroblastoma [59]. CADM1 is involved in signaling and may therefore control many downstream target genes. Moreover, low expression of CADM1 is associated with poor outcome in tumors with and without 11q CNAs [59].

Mapping studies of 1p deletions in neuroblastomas pointed to a 261 kb genomic region encompassing the TF CAMTA1 [60,61]. Similar to CADM1, low expression of CAMTA1

correlates with adverse outcome of neuroblastomas independent from 1p deletions and other clinical markers [53]. Besides neuroblastoma, the prognostic value of CAMTA1 has been described for colorectal cancer [31].

In this study, we employed mouse models and inducible cell line models to explore the biological impact of CAMTA1 induction in neuroblastoma cells. Specifically, I applied and partially extended established gene expression analyses and conducted enrichment analyses to determine gene sets induced or repressed upon CAMTA1 induction and evaluated the contexts in which they take action in the cell.

3.2. Main results

3.2.1. CAMTA1 suppresses growth in neuroblastoma cell lines and is associated with neuronal differentiation

In vitro experiments were conducted on cell lines (SH-EP and IMR5-75) that exhibit low endogenous *CAMTA1* expression levels. The cell lines were transfected with an inducible *CAMTA1* expression vector. After *CAMTA1* induction, colony formation ability and growth rate were significantly reduced (SH-EP) and anchorage-independent growth inhibited (IMR5-75, please refer to Publication II for details). In mice, *in vivo* induction of *CAMTA1* via inoculation of inducible IMR5-75 cells into established tumors resulted in decreased tumor volume. These findings strongly supported the idea of CAMTA1 taking influence on cell cycle, proliferation and growth control.

CAMTA1-induced SH-EP cells exhibited morphological attributes of differentiation, such as neurite-like processes. In a reverse approach, i.e. employing *in vitro* models of neuroblastoma differentiation, morphological differentiation and induction of neuronal marker genes was correlated with increase of *CAMTA1* expression. Taken together, these results imply an impact of differentiation signals and neuronal differentiation onto CAMTA1 regulation.

3.2.2. Identification of genes responsive to CAMTA1 induction

In this project, my task was the bioinformatics gene expression analyses of data from *CAMTA1*-induced neuroblastoma cell lines to identify and categorize the transcriptional changes. I received the raw microarray time series data measuring the gene expression of induced SH-EP cells (and IMR5-75 cells, data not published yet). I modified established approaches [62] to filter out genes that fulfilled three different criteria:

1) Sufficient minimal expression in a defined minimum number of time points.
2) Sufficient variance (*via* interquartile range).
3) Adequate correlation between biological replicates.

The first two filters are established procedures, whereas the third filter was my extension. The filter guarantees similar gene expression profiles within biological groups. Genes that may be regulated in an experiment-independent way are regarded as noise and are removed as they are nonrelevant for the clustering. Applying this extended gene filtering approach yielded 683 transcripts (referred to as genes in the following). I conducted hierarchical clustering employing Pearson correlation distances to further differentiate distinct expression profiles within these genes, resulting in five clusters (see Figure 3.1).

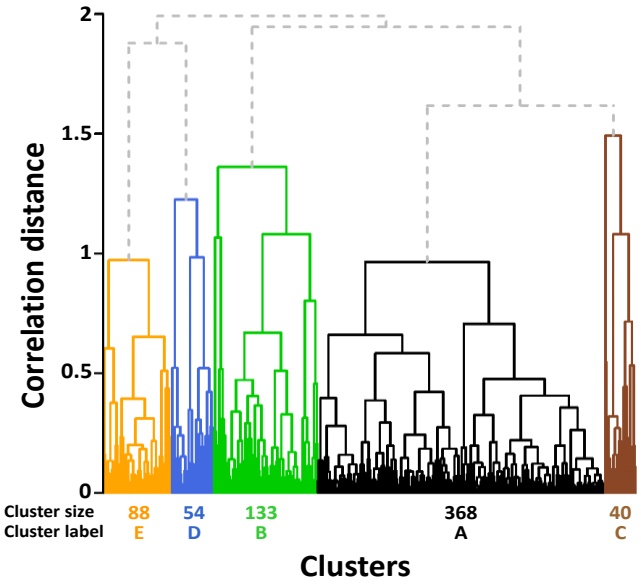
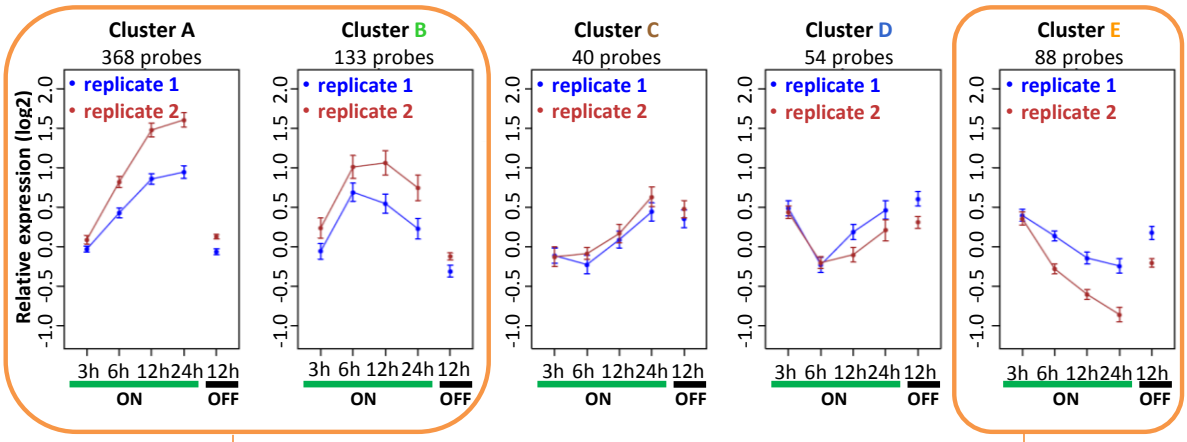


Figure 3.1 | **Clustering dendrogram of selected genes in SH-EP cells upon CAMTA1 induction.** Five clusters with specific gene expression patterns were defined. The number of genes in each cluster and a designated letter for reference are plotted below the graph.

Two clusters (A and B) contained genes that were up-regulated upon *CAMTA1* induction in a time-dependent manner, but unchanged in the 12 hour non-induced control (Figure 3.2). Clusters C and D also exhibited time-dependent distinct expression profiles, but with respect to expression levels of the non-induced controls, the relevance of these observations was less clear. In contrast, cluster E presented a well-defined profile of down-regulated genes upon CAMTA1 induction.

3.2.3. CAMTA1-induced genes influence cellular processes of neuronal development, calcium ion transport, proliferation and metabolism

In the original publication, we used Gene Ontology Tree Machine to functionally annotate two gene sets obtained from clustering (combined clusters A + B, and cluster E). Here, I am presenting an update of the analysis of enriched Gene Ontology (GO) terms that I have conducted just prior to writing this thesis. I decided to do so for two reasons: First, GO database is constantly updated and the annotation accurateness has improved substantially in the meanwhile, and second, I have added a multiple testing correction step (Benjamini-Hochberg). Terms with a corrected $P<0.1$ were defined as significant. The analysis was conducted with the enhanced original tool, now integrated into the WebGestalt v2 toolbox [3,63].



BIOLOGICAL PROCESS
cellular component organization
→ cytoskeleton organization
→ intermediate filament bundle assembly
cellular process
→ muscle cell proliferation
→ microtubule-based movement
→ hormone metabolic process
→ regulation of cell communication
→ negative regulation of signal transduction
→ cell motility
→ negative regulation of smooth muscle cell migration
developmental process
→ system development
→ nervous system development
→ cell differentiation
→ neuron differentiation
→ forebrain neuron fate commitment
metabolic process
→ steroid metabolic process
→ protein repair
transport
→ metal ion transport
→ release of sequestered calcium ion into cytosol
→ regulation of calcium ion transport into cytosol

MOLECULAR FUNCTION
binding
→ protein binding
→ calmodulin binding
→ ion binding
→ sodium ion binding
catalytic activity
→ transferase activity
→ glycine N-acyltransferase activity
→ glycogen phosphorylase activity
transcription regulatory activity
→ basal transcription repressor activity

CELLULAR COMPONENT
cell part
→ cell projection
→ neuron projection
→ cell fraction
→ synaptosome

BIOLOGICAL PROCESS
cell adhesion
→ cell-matrix adhesion
→ regulation of focal adhesion assembly
cellular process
→ cell cycle
→ M phase
→ mitosis
→ cell cycle arrest
metabolic process
→ lipid metabolic process
→ steroid metabolic process
→ vitamin D metabolic process
→ DNA metabolic process
→ DNA replication
→ DNA repair
response to stimulus
→ cellular to UV
transport
→ negative regulation of transport
→ nuclear transport
→ positive regulation of transcription factor import into nucleus

MOLECULAR FUNCTION
binding
→ nucleic acid binding
→ double-stranded DNA binding
→ translation initiation factor activity
→ nucleoside binding
→ adenyl nucleotide binding
→ ATP binding
→ protein binding
→ collagen binding
→ kinase binding
→ histone binding
catalytic activity
→ pyrophosphatase activity
→ DNA helicase activity
→ oxidoreductase activity
→ vitamin D 24-hydroxylase activity
→ oxidoreductase activity, acting on paired donors, ...
→ transferase activity
→ protein kinase activity
molecular transducer activity
→ receptor signaling protein activity

CELLULAR COMPONENT
intracellular organelle
→ cytoskeleton
→ spindle microtubule
→ chromosome
→ chromatin
→ condensed chromosome kinetochore
→ nucleus
→ nucleoplasm

Figure 3.2 | **Gene expression profiles in defined clusters of genes responding to CAMTA1 induction in SH-EP cells.** Each line in the graphs corresponds to a biological replicate. Gene expression was measured between three and 24 hours upon induction and at 12 hours without induction (control). Gene ontology term enrichments associated with combined clusters A and B, and with cluster E are shown below the graphs (algorithm from WebGestalt tool [3], Benjamini-Hochberg corrected $P<0.1$). Many of the terms describe processes or functions that are characteristic of tumor suppressor genes.

GO terms that were significantly (corrected P<0.1) associated with CAMTA1 induced genes are listed in Figure 3.2 (box on the lower left side). The results described in Publication II have been mostly confirmed but are rendered more precisely. Several GO terms describe processes of neuronal development and function, e.g. “nervous system development”, “forebrain neuron fate commitment”, “neuron projection”, or “synaptosome”, which is in line with the suggested role of CAMTA1 in differentiation and neuron development. In addition, shaping of cytoskeleton structures may be regarded as a feature of differentiating cells (e.g. “microtubule-based movement”, “intermediate filament bundle assembly”). Inhibitory effects on signaling by CAMTA1-induced genes are reflected by “negative regulation of signal transduction” and “basal transcription repressor activity”, which are expected for a tumor suppressor gene. GO terms linked to calcium and calmodulin (“calmodulin binding”, “regulation of calcium ion transport into cytosol”, and “release of sequestered calcium into cytosol”) were of particular interest with respect to the Ca^{2+}/calmodulin-dependent activity of CAMTA1. Moreover, these associations suggest that CAMTA1 acts both as an integrator and effector of Ca^{2+} signaling.

3.2.4. Characterizing CAMTA1-repressed genes

GO terms associated with down-regulated genes after *CAMTA1* induction (Figure 3.2, box on the lower right side) were in agreement with these results. Several significant terms were related to cell cycle progression, most prominently reflected by “mitosis”. Interestingly, four genes were also linked to “cell cycle arrest”. It is plausible that because of feedforward loops, inverse feedback loops in cell cycle progression, and cross-talk between pathways, some genes may be associated with this seemingly opposing term. Nevertheless, several checkpoints in the cell cycle require affirmative checks on DNA integrity and positive signaling input. The inhibition of genes associated with “DNA repair”, “DNA replication”, “response to UV”, “spindle microtubule”, and “condensed chromosome kinetochore” implies that CAMTA1 inhibits passage of these checkpoints. Other down-regulated genes are involved in “positive regulation of transcription factor import into nucleus”, “translation initiation factor activity”, “receptor signaling protein activity”, and “protein kinase activity”, which corroborates the hypothesis of an inhibitory effect of CAMTA1 on signaling (and signaling-induced cell cycle progression).

3.2.5. Observed CAMTA1 functionality is confirmed in an independent cell line.

To validate the finding that CAMTA1-regulated genes are involved in neuronal differentiation and cell cycle arrest, we selected five genes from the extrapolated CAMTA1-responsive clusters that were annotated to GO terms “neuronal differentiation” and “cell

cycle arrest": cyclin-dependent kinase inhibitor 1C (*CDKN1C*), tropomodulin 2 (*TMOD2*), sodium channel, voltage gated, type VIII, alpha subunit (*SCN8A*), S100 calcium binding protein B (*S100B*), and stathmin-like 3 (*STMN3*). The expression of these genes was then evaluated by quantitative RT-PCR in an independent CAMTA1-induced SH-EP cell line clone. We found that all five genes were consistently regulated in dependency of CAMTA1 induction. These observations supported the robustness and generalizability of our previous conclusions that CAMTA1 promotes cell differentiation and arrests cell cycle progression.

4. Promoter motif analyses identify common MYCN/MYC binding sites of genes that are up-regulated upon MYCN induction in neuroblastoma cells and that are associated with poor outcome of neuroblastomas without MYCN amplification

This study has been published [64]:

- Publication III

  Westermann F, Muth D, Benner A, **Bauer T**, Henrich KO, Oberthuer A, Brors B, Beissbarth T, Vandesompele J, Pattyn F, Hero B, König R, Fischer M, and Schwab M: **Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas.** *Genome Biol* 2008, 9(10):R150.

4.1. Motivation

In the previous two chapters, I described how gene expression profiling can be applied to elucidate properties of clinico-genetic subgroups, mechanisms of tumorigenesis and the role of a TF therein as a tumor suppressor gene. TFs can also function as oncogenes. Some TFs regulate a small distinct set of genes in a very specific manner, whereas others influence large genetic programs comprising hundreds of genes. TFs have been associated with all hallmarks of cancer development [65]. Moreover, TFs are thought to drive critical cellular genetic programs from development to apoptosis and can therefore become promising therapeutic targets in a wide range of pathological conditions.

MYC gene family members encode TFs and are potential proto-oncogenes. MYC TFs are involved in all aspects of tumorigenesis [66,67]: unlimited proliferation, loss of differentiation, cell growth, neo-angiogenesis, cell motility, and genomic instability.

As stated earlier, genomic amplification of *MYCN* occurs in ~20% of neuroblastoma and identifies a subtype with poor prognosis. *MYCN* amplification is reflected by elevated protein levels and increased activity, and has been implicated in both tumor initiation and

progression of neuroblastoma [68,69]. Seemingly contradictory to its association with malignancy, *MYCN* levels in neuroblastomas without chromosomal *MYCN* amplification (non-amplified; NA) are higher in more favorable localized tumors (stages 1-3; localized-NA) and in stage 4S tumors (4S-NA). 4S-NA tumors have good prognosis in contrast to stage 4 (4-NA) tumors, which are predominantly malignant cancers [70-72]. Furthermore, MYC TFs (including MYCN) can also render cells more sensitive to apoptosis [73], and cell lines from *MYCN* amplified neuroblastomas are still capable of differentiation [74].

All MYC family members require MYC associated factor X (MAX) to form heterodimers to bind DNA via a helix-loop-helix leucine zipper domain to regulate transcription of their targets. The MYC:MAX binding motif has been well studied and the consensus sequence, 5'-CACGTG-3', is known as "E-box".

In this project, we applied gene expression profiling and cluster analyses to determine a core set of MYC and MYCN target genes in *MYCN*-inducible neuroblastoma cells. To validate the target gene set, I conducted promoter analyses employing both canonical motif searches and PWM scans and quantified enrichments of potential MYC binding sites in the target gene set. Additionally, ChIP experiments were conducted in the wet lab (by cooperation partner PD Dr. Frank Westermann and his group) to validate MYC/MYCN binding to the target gene promoters. After this validation, expression levels of the target gene set was used to assess transcriptional activity of MYC/MYCN in different clinical subtypes of neuroblastomas, and the inferred activities were compared to gene expression levels of *MYC* and *MYCN*. Furthermore, we evaluated MYC/MYCN target gene expression in regard to overall survival of neuroblastoma patients.

4.2. Main results

4.2.1. MYC and MYCN are inversely correlated in neuroblastoma subtypes

MYC expression is suppressed in neuroblastoma cells with high levels of *MYCN* RNA [75]. When analyzing 251 neuroblastomas, we observed a similar inverse correlation of *MYC* and *MYCN* expression levels over different tumor subtypes. In *MYCN* NA neuroblastomas, the *MYCN* expression level was highest in stage 4S-NA tumors, followed by localized-NA tumors, and was lowest in 4-NA tumors. Conversely, the gradient of *MYC* RNA decreased from stage 4-NA tumors over localized-NA, stage 4S-NA to *MYCN* amplified tumors.

4.2.2. MYC repression upon MYCN induction in neuroblastoma cells and definition of MYC/MYCN regulated genes

We analyzed gene expression profiles of neuroblastoma cell lines with ectopically inducible MYCN (SH-EP^{MYCN} cells). These cells express endogenous MYC, but not MYCN.

Upon induction of the tetracycline-dependent *MYCN* expression system by tetracycline removal, MYC mRNA and protein levels decreased already prior to elevation of MYCN levels. This trend was reflected by the known MYC target genes prothymosin alpha (*PTMA*), dyskeratosis congenita 1 (*DKC1*), Mdm2 p53 binding protein homolog (*MDM2*), and minichromosome maintenance complex component 7 (*MCM7*). Expression levels of these targets decreased shortly after induction of the tetracycline system just as MYC levels were reduced and before MYCN levels increased. As MYCN levels were elevated later on, so were target expression levels. In general, these genes followed the trend of combined MYC/MYCN expression levels.

We clustered the gene expression profiles taken in a time-series upon induction of the *MYCN* expression system using self-organizing maps (SOMs). All 504 clusters were analyzed for gene expression patterns similar to MYC/MYCN expression maxima, and for enrichments of known MYC/MYCN targets. This yielded two subgroups formed by 167 genes from six clusters: subgroup I contained genes expressed equally high at both maximum MYC levels and MYCN levels. MYCN maximum levels were higher than levels of endogenously expressed MYC, so these genes appeared to be less responsive to MYCN compared to MYC. Gene expression profiles of subgroup II genes matched the maximum protein expression of MYC and MYCN and were highest at time-points with maximum MYCN expression. We found enrichments of genes from the MYC target gene database in these two subgroups ($P < 0.05$). No significant enrichments were found in clusters containing MYC or MYCN repressed genes, so we focused on the two defined subgroups for further validation.

4.2.3. Validation of potential MYC/MYCN target genes in silico and by ChIP

To predict the direct regulation of the identified subgroups by MYC/MYCN *in silico*, I conducted PWM scans and canonical consensus binding motif searches on the gene promoters. I downloaded the available sequences 2kb up- and downstream of the annotated TSS of all genes and scanned for MYCN TFBSs. I then assessed the enrichment of genes with MYCN TFBSs in each cluster from the SOM clustering by Fisher's exact tests and ranked them according to P-values. All six clusters from subgroups I and II were ranked among the top 15 clusters, underlining their potential MYC/MYCN responsiveness. ChIP-on-chip experiments conducted on cell lines expressing MYCN and/or MYC further confirmed these findings. Almost all 140 genes with probes on the tiling arrays corresponding to their promoters were bound by MYC or MYCN in the six cell lines.

Together, these results suggest that genes from the two defined subgroups form a core set of targets for MYC or MYCN, depending on which of these TFs is expressed.

4.2.4. MYC/MYCN activity in MYC NA neuroblastoma subtypes

A portion of 92% (154 of 167) of the defined MYC/MYCN target genes were expressed highest in *MYCN*-amplified neuroblastoma, suggesting similar regulation in the tumors as in the cell lines. In NA tumors, we observed distinctly lower but increasing portions of induced target genes starting from localized-NA tumors (lowest) *via* tumors of subtype 4S-NA towards 4-NA tumors. This indicates that (apart from *MYCN*-amplified neuroblastomas) MYC/MYCN activity is increased in stage 4-NA and to a lesser extent in stage 4S-NA. In agreement with this are elevated levels of either MYC or MYCN in stage 4-NA and 4S-NA respectively compared to localized-NA tumors. Regarding the smaller number of induced genes in 4S-NA, we concluded that MYCN regulates a smaller set of genes in these tumors, while MYC appears to induce a larger set of MYC/MYCN target genes in 4-NA tumors.

4.2.5. High MYC/MYCN target gene expression is associated with poor overall survival

Global tests were applied to all 504 clusters. The six clusters defining MYC/MYCN targets were significantly associated with overall survival of neuroblastoma patients, even after adjustment for co-variables *MYCN* amplification, staging (stage 4 *versus* stages 1-3 and 4s), and age at diagnosis (≥1.5 years). Two of the six clusters, both from subgroup I, ranked on top of all clusters. Finally, there was a substantial overlap of MYC/MYCN targets to genes with previously published neuroblastoma classifiers based on gene expression. These findings support that the expression levels of defined MYCN/MYC targets can be applied to determine MYC/MYCN activity and that high activity serves as an independent robust marker of poor outcome.

5. Reconstruction of the Rage-dependent gene regulatory network in a mouse model of skin inflammation from gene expression profiles and position weighted matrix scans

This study has been published [76]:

- Publication IV

  Riehl A, **Bauer T**, Brors B, Busch H, Mark R, Németh J, Gebhardt C, Bierhaus A, Nawroth P, Eils R, König R, Angel P, Hess J: **Identification of the Rage-dependent gene regulatory network in a mouse model of skin inflammation.** *BMC Genomics* 2010 Oct 5;11:537.

5.1. Motivation

Throughout the previous studies I showed that analyzing gene expression and transcriptional regulation by TFs grants insight into biological mechanisms of tumorigenesis and provides a better understanding of the nature of cancer entities, specifically of neuroblastomas. In the project described in this chapter, we worked together with Dr. Astrid Riehl and colleagues to address the questions how signaling triggered by the receptor for advanced glycation end products (Rage) contributes to the establishment and maintenance of a pro-inflammatory microenvironment that supports neoplastic transformation and malignant progression of skin cancer.

Unresolved inflammation is thought to foster multiple hallmarks of cancer [65] by providing a tumor-promoting environment. A better insight into molecular mechanisms underlying tumorigenesis in the context of inflammation, such as signaling and gene regulatory networks, is implicitly required. Rage is a signaling protein that mediates and maintains the strength of inflammatory responses in a mouse model of skin carcinogenesis upon inflammation [77]. Rage functions as a pattern recognition receptor with multiple ligands and is expressed at increased levels at sites of inflammation. Several downstream target genes have been identified that are expressed context-specifically [78] and have been implicated in neoplastic cell transformation and tumor progression [79-81]. However, in-between mediators of Rage signaling and involved TFs remain mostly unknown. In this project, my cooperation partners conducted time-resolved gene expression profiling of skin samples taken from *Rage*$^{-/-}$ and wild-type (wt) mice treated with tetradecanoyl phorbol acetate (TPA), a potent inducer of inflammation and tumor promoter, to identify genes that are affected by Rage signaling. Subsequently, I applied comprehensive TFBS scans to predict associations of TFs with gene sets exhibiting Rage-dependent gene expression patterns. I further divided the extracted Rage-responsive genes into clusters and predicted associations of TFs with these more specific gene sets. My analysis provided candidate TFs that were investigated on the protein level in skin samples and found to be involved in the gene regulatory network downstream of Rage signaling.

5.2. Main results

5.2.1. Rage-dependent gene expression profiles exhibit two temporal phases in response to TPA stimulation

Rage$^{-/-}$ and wt mice (three biological replicates each) were treated with TPA. Gene expression profiles of skin samples were measured at 6, 12, 24, and 48 hours after treatment and compared to TPA-untreated controls. Transcripts were ranked by average and peak expression relative to controls of respective genotypes, yielding 341 common transcripts

among the three replicate series (Figure 5.1). These transcripts were divided into six clusters by k-means clustering. Expression profiles, particularly of the two largest clusters (cluster 3, n=125; cluster 6, n=84), were in line with the previously described function of Rage in sustaining inflammatory stimuli: cluster 3 contained transcripts that were repressed six hours after TPA induction in both genotypes. The repression was persistent throughout later time-points in wt, but not in *Rage*-deficient mice. *Vice versa*, cluster 6 transcripts were induced transiently in *Rage*$^{-/-}$ mice, but maintained at induced levels in wt animals. These results indicate that gene expression dynamics of genes targeted by Rage can be divided into two phases. The first phase is characterized by an early Rage-independent response to the inflammatory stimulus, whereas Rage is essential for sustaining the changed transcription levels in the subsequent second phase.

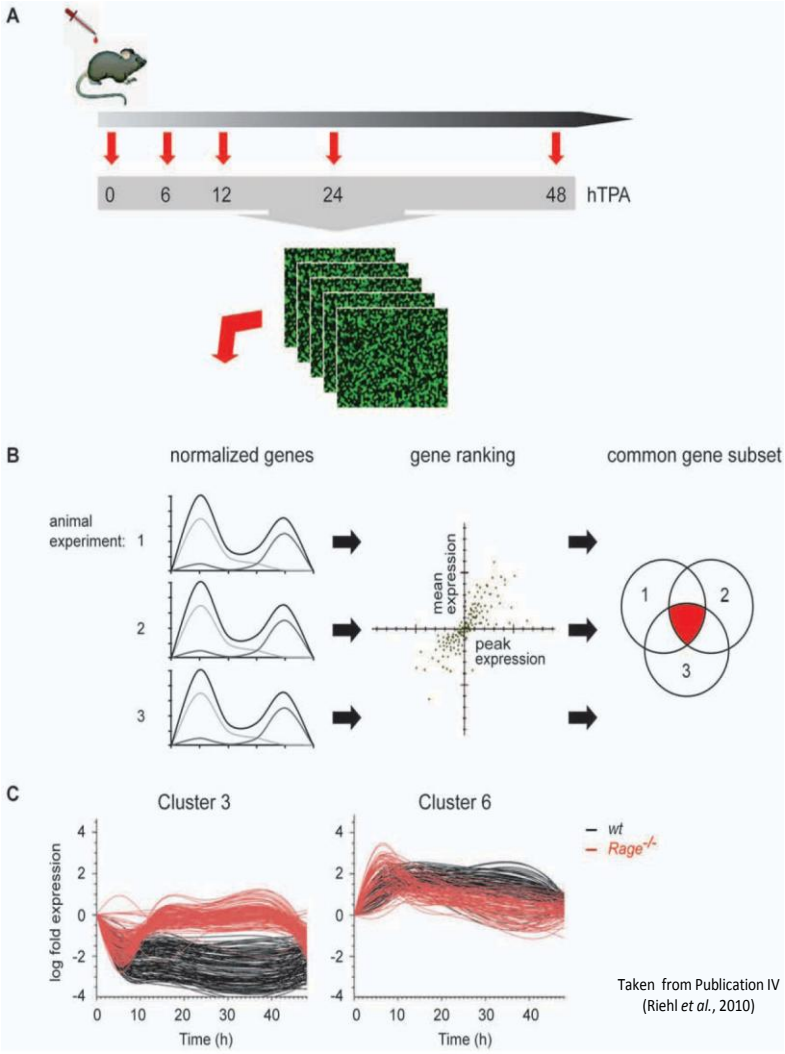


Figure 5.1 | **Gene expression dynamics of wt and *Rage*$^{-/-}$ mice upon TPA treatment.** **A** Back skin was isolated 6, 12, 24, or 48 hours after TPA stimulation. Non-treated and acetone-treated mice served as controls (0). Microarray global gene expression analysis of RNA samples was performed (three replicates for each genotype and time point). **B** Following quantile normalization, wt genes were ranked according to high mean and peak expression separately to filter for TPA-responsive genes. A common subset of 341 genes was identified among the 1000 top-ranked genes of three experiment series. **C** K-means clustering produced six clusters of which cluster 3 and 6 were the largest. The kinetic of these clusters showed a response independent of the *Rage* genotype at t=6h, but the stimulus response of both repressed (cluster 3) and induced (cluster 6) transcripts was only sustained in wt mice.

5.2.2. Rage-dependent differential expression after TPA stimulation

We aimed at identifying genes that were differentially expressed between *Rage*$^{-/-}$ and wt mice in the second phase after the TPA stimulus. (This analysis was conducted independently of the clustering described in the previous passage.) A linear model was applied that revealed differential expression at a significant level (corrected P<0.05) only at time point t=24h upon TPA administration. A total of 122 transcripts (representing 114 different genes) were differentially expressed at this time point, including induced and repressed genes.

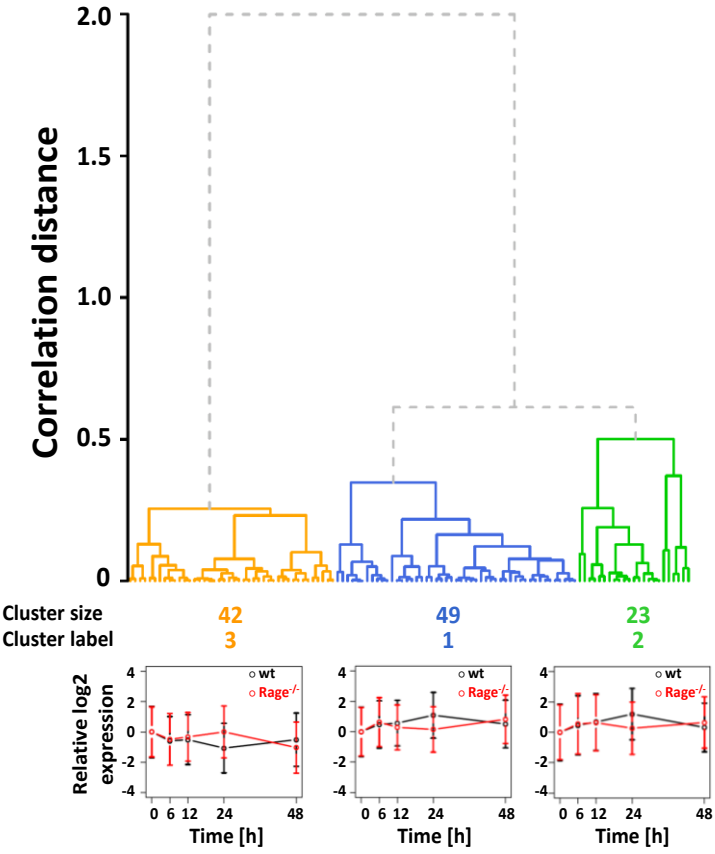


Figure 5.2 | **Hierarchical clustering of differentially expressed transcripts between wt and *Rage*$^{-/-}$ mice 24 h after TPA treatment.** Three distinct clusters were derived from the dendrogram. Clusters 1 and 2 contained significantly up-regulated transcripts at 24h in wt mice, whereas cluster 3 transcripts were significantly down-regulated at that time point.

5.2.3. Predicting TFs of the Rage-dependent gene regulatory network

I applied hierarchical clustering employing Pearson correlation distances to the gene expression values, thereby revealing three clusters with distinct expression profiles (Figure 5.2). Then I assessed which TFs were likely to regulate the individual clusters as well as the whole set of differentially expressed genes. I conducted PWM scans on promoter regions (±2 kb of the TSS) with all PWMs available for TFs in Transfac database [2]. After mapping TFBSs from PWMs to TFs, I calculated over-representation of TFBSs in the gene sets by Fisher's exact tests. Within all differentially expressed genes, putative binding sites of 17 TFs and

isoforms were significantly enriched (corrected P<0.05, Table 5.1). Some TFs could thereby be associated with individual clusters: trans-acting transcription factor 1 (Sp1) and 4 (Sp4), hepatic nuclear factor 4 (Hnf4), and CAC-binding protein (CAC-bp) with cluster 1, Wilms tumor 1 homolog (Wt1) with cluster 2, and E2F transcription factor (E2f) with cluster 3. TFBSs for MAZ related factor (Mazr) were enriched in both clusters 1 and 2.

Table 5.1 | **Enriched TFBS in differentially expressed genes 24 hours after TPA stimulation.**

Clusters	TF	Fischer test P-value	Corrected P-value	Cluster genes	
				with PWM	without PWM
all	Sp1	5.33E-07	1.06E-04	94	3
	Sp1 isoform 1	5.33E-07	1.06E-04	94	3
	Sp4	1.72E-06	2.27E-04	83	14
	AP-2beta	1.24E-06	1.22E-03	77	20
	AP-2alpha	1.60E-05	1.27E-03	79	18
	AP-2gamma	2.13E-05	1.41E-03	79	18
	MAZR	6.03E-05	3.41E-03	74	23
	CAC-binding protein	1.32E-04	6.52E-03	81	16
	Egr-1	3.56E-04	1.56E-02	85	12
	Egr-3	4.38E-04	1.73E-02	78	19
	E2F	4.85E-04	1.75E-02	58	39
	c-Myc	7.31E-04	2.41E-02	67	30
	Egr-2	9.69E-04	2.94E-02	80	17
	COUP-TF1	1.06E-03	2.94E-02	89	8
	WT1	1.19E-03	2.94E-02	67	30
	WT1-isoform1	1.19E-03	2.94E-02	67	30
	COUP-TF2	1.45E-03	3.38E-02	48	49
cluster 1	Sp4	2.01E-06	7.93E-04	40	2
	Sp1	6.13E-05	6.86E-03	42	0
	Sp1 isoform 1	6.13E-05	6.86E-03	42	0
	MAZR	6.93E-05	6.86E-03	36	6
	HNF-4alpha7	1.84E-04	1.46E-02	29	13
	CAC-binding protein	3.12E-04	2.06E-02	38	4
cluster 2	MAZR	2.14E-04	7.64E-02	19	1
	WT1	5.79E-04	7.64E-02	18	2
	WT1-isoform1	5.79E-04	7.64E-02	18	2
cluster 3	E2F	1.88E-05	7.45E-03	28	8
	E2F-1	8.50E-05	1.68E-02	28	8

Taken together, the enrichment analyses predicted several TFs that had not been reported previously in connection to Rage signaling. These TFs were therefore promising

candidates for further investigation. In particular, TFs associated with cell cycle and tumor pathology, such as E2f and Wt1 proteins, are of interest in the context of carcinogenesis.

5.2.4. Expression of E2f TFs upon induction of Rage signaling by TPA stimulation

We wanted to test the hypothesized involvement of members from the E2f family of TFs in Rage signaling. No significant changes in gene expression levels were detected in the expression data, so we considered post-transcriptional alterations that would affect the activity of these TFs. Protein levels of E2f1, a transcriptional activator, and E2f4, a transcriptional repressor, were quantified on Western blots and further visualized by immunohistochemical staining.

Indeed, protein levels of the transcriptional activator E2f1 were induced in keratinocytes of both *Rage*$^{-/-}$ and wt samples at time points t=6h and t=12h after TPA stimulation. While the expression level was still kept up high in wt keratinocytes 24 hours after TPA stimulation, it was not increased in *Rage*-deficient skin lysate. E2f4 protein was induced and a constant increase was observed until 24 hours after TPA stimulation in wt samples. In contrast, E2f4 levels were not induced and remained constant in *Rage*$^{-/-}$ samples. The findings are consistent with the two phase model of Rage signaling described earlier and corroborate a potential impact of Rage signaling on E2f activity. Whether this impact is direct or indirect could not be concluded at this time, however. It is worth mentioning that the E2f-Rb pathway is critical for strict regulation of cell cycle progression and often directly targeted in carcinogenesis. It is therefore plausible that downstream targeting of this pathway by Rage signaling may provide a molecular link between inflammation keratinocyte hyperproliferation supporting skin cancer development. The results from the immunohistochemical staining were in line with the protein dynamics observed on the Western blots and provided an additional highly informative visualization.

6. RIP: The regulatory interaction predictor – machine learning based approach for predicting target genes of TFs

This study has been published [82]:

- Publication V

  Bauer T, Eils R, and König R: **RIP: The regulatory interaction predictor - a machine learning based approach for predicting target genes of transcription factors.** *Bioinformatics.* 2011 Aug 15;27(16):2239-47. Epub 2011 Jun 20

  .

6.1. Motivation

When I started my PhD, I was exited by the idea of reconstructing a genome-scale human regulatory network that would elucidate the means by which cell signaling drives the dynamics of gene expression. The applications of such a network would be enormous. It would be possible to trace observed changes in mRNA levels back to the source, i.e. the controlling elements (TFs and upstream signaling pathways). In pathogenesis, causative molecular mechanisms could be extrapolated and their elements targeted in therapy to name only one exciting application. All through my cooperation projects I worked on gene expression profiling and follow-up analyses with the aim to understand the molecular alterations behind carcinogenesis. I learned how large-scale promoter analyses can identify potential TFs in control of transcriptional changes. However, the PWM scan technique, even though applicable with good results as demonstrated in previous chapters, tends to produce large numbers of false positives which reduces the precision of the predictions considerably.

The core elements of regulatory networks are TFs, target genes, and regulatory interactions (RIs) between them. Several approaches have been developed to reconstruct regulatory networks on different scales and model organisms (reviewed in [18,19]), but essentially they have not achieved satisfactory results in the attempt to realize the idea I described above. Major issues of most present methods are:

a) Statistics for inferring RIs based on questionable assumptions of gene regulation and/or missing validation of the assumptions in the used data (proof of principle).
b) Improper transfer of gene regulatory principles from prokaryotes to eukaryotes.
c) High computational demands of the models that drastically limit the number of included network components.
d) Over-simplification, i.e. the use of gene expression data only instead of including data representing different aspects of gene regulation.
e) Lack of an objective true positive set (True RIs) to estimate the performance or to validate the findings.
f) Insufficient precision (high false positive rate) or insufficient recall (low re-discovery of known RIs).

Up to date, a lot of attention and effort are still focused on providing solutions to this unresolved major task of systems biology. In this chapter, I will describe the method I have developed to contribute to the realization of this idea. I will illustrate the achieved improvements and provide examples of successful application of my method.

Most current algorithms for large scale RI inference are based solely on gene expression data and assume a direct relationship between the gradients of TF mRNA and its target genes. While this assumption may be true for a large number of TFs in prokaryotes, it

is not met by many TFs in human, where post-translational modifications affect TF activity or degradation kinetics [33]. Unfortunately, techniques to measure protein quantity and kinetics in high-throughput are at the best in a developmental stage and data are available in insufficient numbers only. I therefore developed an *in silico* method to compensate this limitation.

A biological concept that has been around for long is the principle that genes that share biological functionality are co-expressed, and this co-expression is achieved by co-regulation. So instead of considering statistics between TF mRNA gradients and potential target genes, I analyzed statistics of co-regulated target gene sets and subsequently deduced their regulatory TFs from known RIs, thereby overcoming shortcomings of conventional methods and lack of protein data. Human gene expression data covering a large spectrum of biological conditions are available in abundance, and thus I conducted a correlation meta-analysis of thousands of gene expression profiles to identify co-expressed genes in a large number of primary human tissues. Additionally, I analyzed gene promoters employing comprehensive PWM scans to acquire putative TF binding data that are unbiased by experimental conditions, as in case of e.g. ChIP analysis. Finally, I extracted a considerable amount of RIs identified in published experiments that were assembled in Transfac database [2]. Our concept was to have a machine learning classifier learn the trends of correlation and TFBS enrichments within RIs known to be co-regulated and then predict RIs on a genome-wide scale to discover new RIs. For this purpose, we defined 10 elaborate features (quantifiable characteristics) that combined the results of correlation and PWM analyses of known RIs. I trained numerous SVMs with the features of defined training sets and performed cross-validations to estimate the quality of the predictions. I eventually combined all SVMs into one master classifier termed “regulatory interaction predictor” (RIP) that achieved considerably good recall and precision. RIP was then used to predict RIs between 303 TFs and 13 069 genes. The predictions were validated by pathway analysis, with an independent RI database, and further applied to a (published) *in vivo* study on interferon α (IFNα) signaling in monocytes to identify key TFs affected by IFNα induction.

6.2. Main Results

6.2.1. Training machine learning classifiers to predict TF target genes – the workflow

The algorithm we developed for our supervised machine learning approach to predict RIs between TFs and target genes is depicted in Figure 6.1. Defining sets of true positives (TP) and true negatives (TN) of sufficient sizes was an essential prerequisite for training the SVMs. I extracted 2896 RIs between 303 TFs and 949 target genes from Transfac database,

which were defined as the TP set (True RIs). *Vice versa*, all other possible combinations of the 303 TFs and 949 genes (=284 641 unknown RIs) were defined as the TN set (True non-RIs). There may be a number of True RIs within the set of unknown RIs, but even if one assumed that at present only 10% of RIs had been discovered in total, the defined True non-RIs would only contain ~26 000 wrongly labeled RIs. Compared to the much larger amount of remaining ~258 000 True non-RIs, this would still be acceptable.

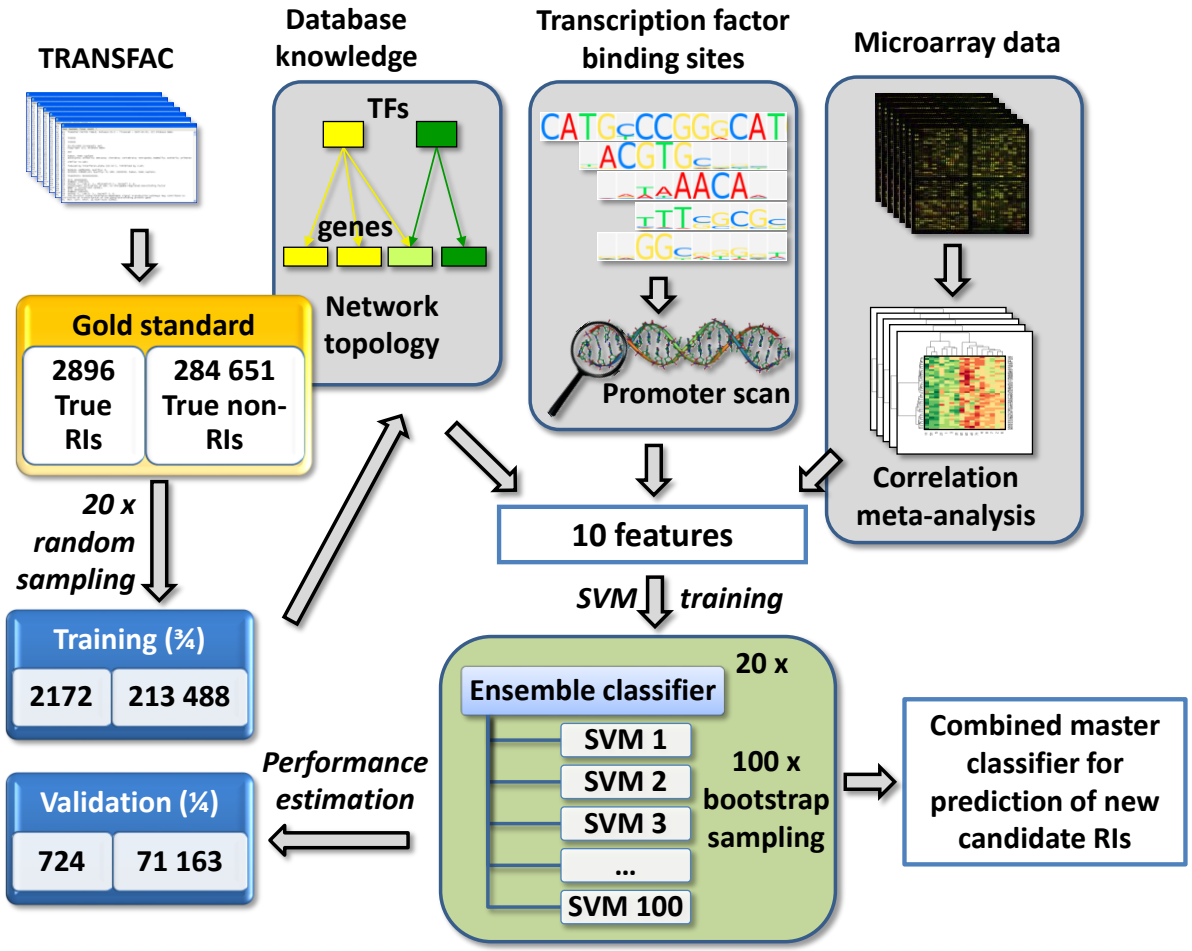


Figure 6.1 | **General workflow of RIP.** Features for inferring regulatory interactions (RIs) between TFs and genes were derived from three different aspects: tightly correlated genes identified by meta-analysis of gene expression profiles, TF binding site predictions, and database content of co-regulated genes from the training set (gold standard). The information of the gold standard was also used to define True RIs and True non-RIs.. For training of Support Vector Machines (SVMs), True RIs and True non-RIs were divided into a training set and a validation set. An equal number of True RIs and True non-RIs were randomly drawn (by bootstrapping) 100 times and used to train 100 different SVMs yielding one ensemble classifier. Each ensemble classifier was evaluated with its validation set. This procedure was repeated 20 times yielding an averaged estimate about their performances. The classifiers were combined to one master classifier (RIP) containing 2000 SVMs, and applied to predict new RIs.

We then needed to describe RIs by quantifiable features that reflect characteristics of regulatory influence of TFs on target genes. We based these features on two assumptions:

1) Gene sets that are involved in a common biological process are co-regulated. Common TFs should thus control these gene sets under specific conditions, and these genes should frequently show correlation on the mRNA level.
2) Gene sets directly regulated by a common TF (TF-modules) ought to possess (enrichments of) corresponding TFBSs in their promoter sequences.

We deduced 10 features from these assumptions by a) analyzing correlation of gene pairs in 4064 human gene expression profiles from 76 biological conditions (e.g. tumor type, tissue type, etc.), b) conducting genome-wide PWM scans, and c) using statistical descriptors of network structure arising from the True RIs. Before training the SVMs, I tested if the assumed principles underlying our features were reflected by the data.

6.2.2. Genes with correlated gene expression share biological processes

I conducted a correlation meta-analysis by calculating Pearson correlation coefficients for all possible gene pairs within 13 069 genes (all genes represented on the microarray platform Affymetrix HGU133A) in 76 biological conditions. The correlation coefficients were used to select gene pairs at different stringency by applying two filters CC and FoC. CC was the minimum (absolute) correlation coefficient that was required in a minimum fraction of conditions FoC. Therefore, CC controlled correlation intensity, and FoC controlled correlation frequency, and they were both applied at different stringency levels. The functional relation of the filtered gene-pairs was estimated using selected Gene Ontology annotations (GO, [83]) and a method adapted from [84]. In brief, I selected 81 GO terms that represented a broad range of biological functions, and that were still sufficiently specific. The functional relatedness of gene pairs was quantified by the Functional Similarity score (FS-score), which is the percentage of gene pairs sharing at least one selected GO term. Figure 6.2 shows the results for gene pairs filtered at various stringency levels. FS-scores between 14.8% and 58.3% were achieved (stringency parameters CC=0.6 to 0.9, FoC=0.25 to 0.5). For a wide range of cutoffs (selecting ≤5000 genes, see Figure 6.2), the FS-scores increased with higher stringency (up to CC=0.8, FoC=0.35) from 14.8% to 57.3%., which was what we expected assuming that genes sharing biological functions tend to correlate (assumption 1). Interestingly, the FS-score of filtered gene pairs fluctuated to some extent towards the highest stringency levels (<300 selected gene pairs) before recovering and reaching its summit. This behavior resulted from an increased proportion of constitutively expressed gene families (e.g. hemoglobins, histones, immunoglobulins) that

showed high correlation of expression between each other without sharing any common biological processes.

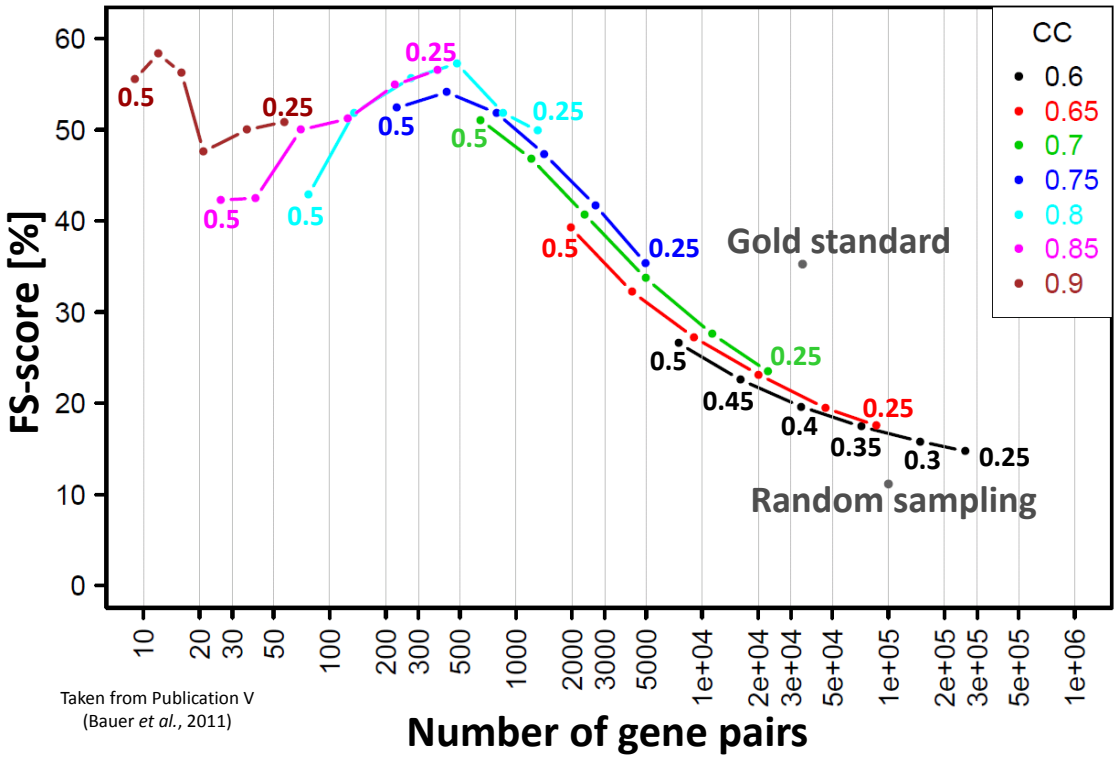


Figure 6.2 | **Functionally relatedness of gene pairs with high correlation of expression.** The graph shows the FS-score which is the percentage of gene pairs sharing at least one functional category for a variety of different stringency criteria, i.e. Pearson correlation (CC) and fraction of classes (FoC). For example, setting the threshold for CC to 0.85 in>25% (FoC=0.25) of the datasets yielded 380 annotated gene pairs of which 56.6% (=215) shared the same functional GO category. In comparison, the gold standard comprised 35.3% (12 176 out of 34 538) pairs having at least one functional GO category in common and merely 11.2% of 100 000 randomly selected gene pairs shared functional GO categories.

To investigate the relationship between biological cooperation (sharing biological functions) and co-regulation, I calculated the FS-score of all gene-pairs derived from TF-modules of the True RIs (each of these gene pairs is regulated by at least one common TF) and compared it to 100 000 randomly selected gene pairs. The gene pairs of True RIs (the gold standard) had a FS-score of 35.3%, whereas the FS-score of randomly selected gene pairs was distinctively lower (11.2%). Notably, the maximum FS-score of the filtered gene pairs was even higher than that of gene pairs of the True RIs, and both were substantially increased in comparison to randomly selected gene pairs.

In summary, these results demonstrate that functional relatedness of genes is associated with increasing transcriptional correlation levels. Both correlated gene pairs and

co-regulated genes showed substantially higher functional relatedness than randomly selected gene pairs. This showed that assumption 1 (described above) was generally fulfilled within the gene expression data.

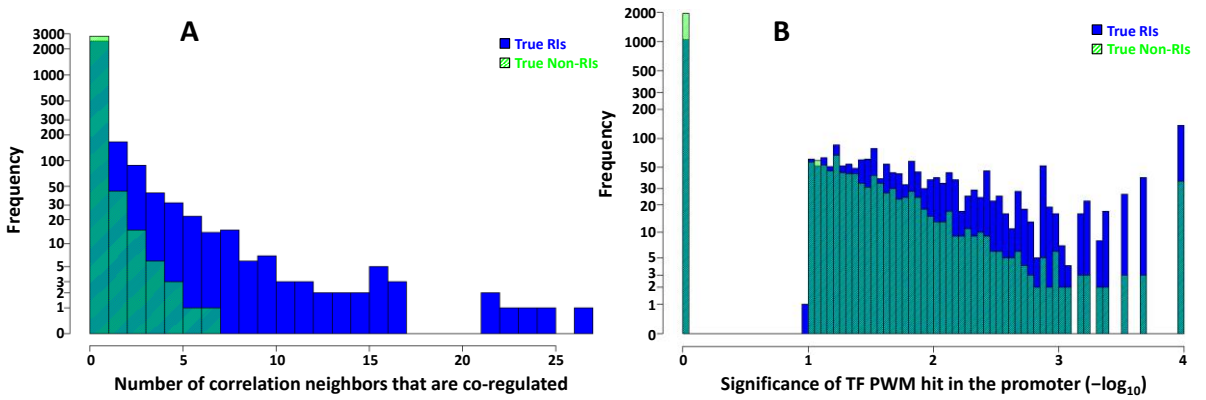


Figure 6.3 | **Distributions for True RIs and True non-RIs of two selected features. A** Frequency distribution of feature 2 True RIs (blue bars) and True non-RIs (green bars): the number of correlation neighbors known to be regulated by the corresponding TF (True RIs from Transfac). Genes of True RIs had more correlation neighbors regulated by the same TF than genes of True non-RIs. **B** Frequency distribution of feature 7: binding motif significance obtained from PWM scans of the gene promoters (1kb upstream of the TSS). A $-\log_{10}$ transformation was applied to linearize the significance (P-values). Genes with a True RI to a TF contained a significant binding motif of the regulating TF more frequently than genes not known to be regulated by the TF (True non-RIs). For comparability, counts for True non-RIs were stratified to the total number of True RIs in this figure.

6.2.3. Correlated genes are frequently regulated by common TFs

The next question was if the consequence of assumption 1, i.e. correlated gene pairs are co-regulated, could be generalized to describe features of RIs that are useful for machine learning. I therefore assigned *correlation links* between genes with sufficiently high correlation. I chose CC=0.6 and FoC=25% as a robust cutoff yielding the most correlation links for the highest number of genes while having an FS-score that was still sufficiently higher than the FS-score of random gene pairs. For a gene of interest, all genes with a correlation link to that gene were defined as its *correlation neighbors*. I then compared the correlation neighbors of genes from True RIs with those of genes from True non-RIs. I found that correlation neighbors were generally much more likely to have a True RI with a TF if the considered gene also was known to be regulated by that TF (feature 3; see Figure 6.3A). To quantify the relevance of this observation, we employed Fisher's exact test to calculate enrichment significances of True RIs among correlation neighbors. In total, we defined six features for RIs based on correlation neighbors (features 1-6) and two additional features containing the averaged correlation over all conditions (features 9 and 10). All features

based on correlation neighbors significantly discriminated True RIs and True non-RIs. Table 6.1 lists descriptions of the 10 features and their discriminatory power calculated by a Wilcoxon-test.

Table 6.1 | **Feature descriptions, included data resources, and discriminatory significance between True RIs and True non-RIs.**

Index	Feature Description	Feature includes				Wilcoxon test (P)
		PWM scans	correlation neighbors	median correlation	True RI structure	
1	The number of correlation neighbors of the corresponding gene.		X			5.1E-03
2	The number of correlation neighbors (including the corresponding gene) which were known to be regulated by the corresponding TF (True RIs).		X		X	<4.6E-86
3	The -log10(P) in which P was the enrichment significance (Fisher test) of known regulated genes (True RIs) in the correlation neighbors (including the corresponding gene) compared to all other genes.		X		X	<4.6E-86
4	The number of correlation neighbors with a significant PWM hit of the corresponding TF.	X	X			1.5E-50
5	The -log10(P) in which P was the enrichment significance (Fisher test) of PWM-hits of the TF within the correlation neighbors including the corresponding gene.	X	X			4.6E-86
6	The number of correlation neighbors that were known to be regulated (True RIs in the training sets, taken from the gold standard) and which had a significant PWM hit of the TF.	X	X		X	<4.6E-86
7	The -log10(P) in which P was the significance of the (best) PWM hit of the corresponding TF.	X				<4.6E-86
8	The number of genes regulated by the TF (True RIs). This feature was added to enable differentiation between common and specific TFs.				X	<4.6E-86
9	Select all genes regulated by the corresponding TF (True RIs). For these genes, calculate the average median correlation to the corresponding gene of the RI over all 76 conditions.			X	X	7.2E-55
10	Select all genes which were putatively not regulated by the corresponding TF (True non-RIs). Feature 10 was then calculated like feature nine.			X	X	1.1E-02

6.2.4. Promoters of known target genes contain TFBS enrichments of corresponding TFs

To find putative TFBSs, the promoter of each gene was scanned for known binding motifs (using PWMs). I found that True RIs had more often a putative TFBS of the particular TF than True non-RIs (feature 7, P<4.6E-86; Table 6.1; Figure 6.3B). This implicated the general validity of assumption 2 for the 13 069 investigated genes. In total, we derived four features from TFBS predictions (features 4-7). All four features showed highly discriminative power distinguishing True RIs from True non-RIs (Table 6.1).

6.2.5. Classifier performance

A total of 2000 SVMs were trained employing a 20x 100-fold stratified cross-validation to compensate for the imbalance in numbers between True RIs and True non-RIs. The sampling scheme is illustrated in Figure 6.1. Each of the 20 SVM sets from the outer cross-validation encompassed 100 SVMs and is denoted "ensemble classifier" in the following. The performance of the 20 ensemble classifiers were computed and the average used as a performance estimate of the combined classifier, which I designated "regulatory interaction predictor" (RIP). I compared the performance of RIP to other methods for inferring RIs, including the algorithm for reconstruction of accurate cellular networks (ARACNE; [32]), context-likelihood of relatedness (CLR; [26]), two approaches of RI inference by means of correlation between TF-coding genes and potential target genes, and RI inference by conventional PWM scans (see Figure 6.4 and Publication V, Supplement S3 for details).

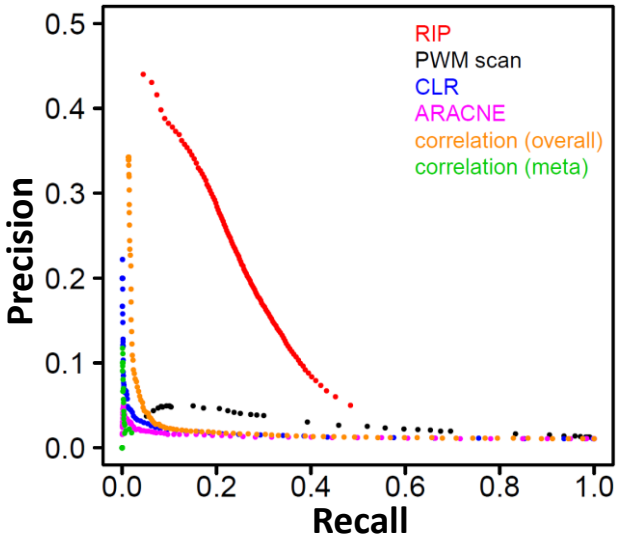


Figure 6.4 | **Precision *versus* recall curve of RI inferring methods.** The performance of the RIP classifier was compared to that of PWM scans, CLR, ARACNE, and correlation-based inference (meta: using filtered gene pairs with CC and FoC; overall: average correlation over all 4064 samples). The RIP classifier achieved considerable precision and recall, and clearly outperformed all compared methods in the human data.

Because of the sparsity of links in regulatory networks (many non-RIs), precision (rate of true positives out of all positively classified) and recall (true positive discovery rate) are

critical for judging the classifier utility. Notably, RIP obtained good precision levels. At the most stringent cutoff, RIP reached maximum precision (44.0%) and a recall of 4.5%. For the lowest stringency, the precision decreased to 5% at the highest observed recall (48.4%). The recall-precision curve is shown in Figure 6.4. The RIP classifier clearly outperformed the other compared methods that did not yield convenient precision levels at any stringency. This probably resulted from the assumptions behind those algorithms, which deduce RIs from statistics relating TF mRNA expression gradients to that of the target genes. While this worked well in lower model organisms used originally by CLR and ARACNE, within the human datasets TF mRNA correlated with target gene expression in only 2.2% of True RIs (average correlation r≥0.6). Unlike RIP, neither CLR nor ARACNE succeeded in recovering True RIs specifically in this wide range of eukaryotic cellular contexts.

I subsequently analyzed the features of True RIs that were never classified positively by RIP and compared them to True RIs that were always classified correctly. More than 50% of these misclassified True RIs did not have any significant PWM hits within the promoter regions in contrast to >99% of the correctly classified True RIs. This indicates that RIP tends to favor regulation mediated by direct binding of TFs to the promoter over indirect effects. It is to note that the performance of RIP was estimated rather conservatively. The actual number of True RIs is likely to be higher than our estimates based on Transfac entries because many RIs have not been discovered so far, and RIP was designed to discover such new RIs.

6.2.6. Inferring new RIs

The training of RIP encompassed a total of 303 TFs and 949 genes. To predict new RIs, I calculated the 10 features for all 3 959 907 possible RIs between 303 TFs and an extended target gene pool (n=13 069). The RIP master classifier was applied and provided 2000 votes (one vote from each SVM) for each candidate RI. Confidence values were assessed from the averaged precisions of the 20 ensemble classifiers (see last section). RIP predicted 6073 RIs with 44.0% confidence at the most stringent cutoff. At ≥31.5% confidence (≥1600 SVM votes; 17.7% recall), it predicted 73 923 RIs for 301 TFs and 11 263 genes, which was a sufficient number of RIs with an acceptable portion of true positives, and was thus used for further analyses.

6.2.7. Applying the inferred regulatory interactions to a microarray gene expression study: identifying TFs responsive to interferon α

An important application of RIP predictions is the identification of TFs associated with observed gene expression profiles. For example, a TF can be associated to a set of

differentially expressed genes if predicted TF targets are significantly enriched therein. In a study on the effect of the cytokine IFNα on monocytes [85], a gene set of 241 genes showed significantly altered expression between samples with and without IFNα induction. We used this data as a case study for testing the utility of RIP predictions. Fisher's exact tests revealed over-representation of predicted targets (≥1600 votes) of 13 TFs in the gene set (Table 6.2). All these TFs have been described previously in IFNα-induced signaling [86-90]. Predicted targets of the heterotrimeric TF-complex interferon stimulated gene factor 3 (ISGF3) were most significantly enriched, with 20 out of 28 predicted ISGF3 targets differentially expressed (71.4%). ISGF3 is composed of signal transducer and activator of transcription 1 and 2 (STAT1 and STAT2) and IFN regulatory factor 9 (IRF9). ISGF3 is activated by cytokines and inflammatory factors [85] and functions as an IFNα treatment-dependent transactivator of IFN-inducible genes [91]. These findings clearly demonstrated the utility of RIP in reconstructing regulatory effectors from gene expression profiles. To compare this result to a standard method, I used PWM scans at a similar stringency level and predicted RIs with 29 genes for ISGF3, of which only four were differentially expressed, corroborating the superior precision of RIP.

Table 6.2 | **Association of predicted TF-modules with differentially expressed genes upon IFNα induction in monocytes.**

Transcription factor	Differentially expressed	Module size	%	Corrected P-value
STAT1:STAT2:IRF9	20	28	71.4	6.95e-23
IRF1	58	1187	4.9	5.72e-03
IRF2	15	169	8.9	1.07e-02
STAT1	67	1513	4.4	1.15e-02
GAF	3	5	60	1.15e-02
NFKB1	36	681	5.3	1.59e-02
STAT3	23	384	6	3.21e-02
IRF7	4	17	23.5	3.53e-02
ETS1	48	1065	4.5	3.53e-02
RELA	37	762	4.9	3.53e-02
IRF3	4	18	22.2	3.53e-02
ELF2	3	9	33.3	3.70e-02
SPI1	24	439	5.5	4.63e-02

Among the genes encoding the 13 TFs associated by RIP predictions, many exhibited consistently altered transcription levels, but only one (IRF7) was significantly up-regulated upon IFNα induction. This may indicate post-translational regulation of TF activity, which has been previously reported for most (if not all) of the 13 TFs (e.g. [87,89,90,92]). Nevertheless,

RIP successfully predicted RIs for the affected genes and key TFs that are likely to be missed by other methods.

Table 6.3| **Associations of predicted TF-modules with KEGG pathways.**

Transcription factors	Pathway
IRF1, IRF2, IRF3, IRF5, **IRF7**, STAT1, STAT3, **NFATC2**, **NF-GMa**, **NFκB**, **NFKB1**, **NFKB1:RELA**, **RELA**, CD28RC, HMGA1, JUN, CEBPA, CEBPB	Cytokine-cytokine receptor interaction
IRF7, **STAT4**, STAT1:STAT2:IRF9, CD28RC, NFATC2, **NF-GMa**, POU1F1	Jak-STAT signaling pathway
IRF1, **IRF3**, **IRF7**, **NFκB**, **RBPJ**	Toll-like receptor signaling
NFATC2, **NF-GMa**	Fc epsilon RI signaling pathway
NF-AT, **NFATC2**, **NF-Gma**, SPI1	Hematopoietic cell lineage
IRF2, **NF-AT**, **NF-AT1**	T cell receptor signaling
ELF1	Natural killer cell mediated cytotoxicity
IRF1, IRF2, LEF1, XBP1, **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK**	Antigen processing and presentation
IRF1, XBP1, **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK**	Cell adhesion molecules (CAMs)
ETS1, STAT1	MAPK signaling pathway
IRF2, **NFKB1:RELA**	Apoptosis
SP4	Calcium signaling pathway
TCF7L2	Wnt signaling pathway
p53, p73	p53 signaling pathway
E2F:DP, E2F4, NFYA	Cell cycle
E2F:DP, E2F1:TFDP1/TFDP2, E2F4	DNA replication
E2F:DP	Purine metabolism
E2F:DP, E2F4	Pyrimidine metabolism
E2F:DP, E2F1:TFDP1/TFDP2, E2F4	Nucleotide excision repair
E2F1:TFDP1/TFDP2, E2F4	Mismatch repair
GATA4, **NR5A1**	C21-Steroid hormone metabolism
NR5A1	Androgen and estrogen metabolism
NR1H4, **PPARA:RXRA**, **RXRA**	PPAR signaling pathway
NR1I2, **RXRA:NR1I2**, **RXRA:NR1I3**	Retinol metabolism
NR1I2, **RXRA:NR1I2**	Linoleic acid metabolism
HNF1A, **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3**	Drug metabolism - cytochrome P450
HNF1A, NFE2, **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3**	Xenobiotics of cytochrome P450
FLI1, HNF1B, SMAD3	ECM-receptor interaction
HNF1B	Focal adhesion
	Cell junctions
NFE2L2	Glutathione metabolism
NR1H3, SP4	Neuroactive ligand-receptor
RARB	Non-homologous end-joining
	Proteasome
	Protein export
	Oxidative phosphorylation

6.2.8. RIs predicted for a large number of TFs are supported by pathway analysis and an independent database

In the previous section, I described the relevance of RIs predicted by RIP in a specific cellular context. To evaluate the validity of predictions on a broader basis, I followed two approaches. The first was a pathway analysis. I associated TFs to pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) by means of over-representation of predicted target genes (≥1600 votes) among the pathway genes. The results cover a variety of signaling and metabolic pathways and are presented in Table 6.3. Most of the associations have been described previously in numerous publications and reflect the biology of the pathways.

A similar analysis conducted on the original set of True RIs from Transfac further confirmed many of the associations, and several potentially novel associations were found (bold TF names in Table 6.3). An elaborate description of the results can be found in Publication V, in the main text, section 3.8, and in Supplement S5. In summary, key TFs involved in the following pathways were found (highlighted in Table 6.3 in colors indicated in brackets):

a) Cytokine response and immune system (red).
b) Cell cycle and proliferation-related signaling pathways (yellow).
c) Steroid (grey), retinol and drug metabolism (green).

In a second validation approach I compared predicted RIs (≥1600 votes) to known RIs of 74 TFs from 25 TF families of the Transcriptional Regulatory Element Database (TRED). The procedure and all results are described in Publication V, Supplement S6. In brief, I calculated over-representation of TF-modules from TRED (TRED TF-modules) within TF-modules from the predicted RIs. In 85.4% of the tested TFs, the correct TF family was assigned (among the top three hits), and the actual TF was assigned correctly (top three hits) in 73.5%. Taken together, these results strongly support the broad functional relevance of RIs predicted by RIP, particularly for biological contexts involving the associated pathways and TFs tested.

6.2.9. RIP software package

To provide RIP to the public I implemented an easy-to-apply software package with a manual containing application examples for the statistical software R. The package is published under GNU general public license ≥2 and freely available for download at http://www.ichip.de/software/RIP.html. The RIP-package allows the application of the RIP

classifier to predict RIs specific to a single condition as well as to predict RIs common to sets of distinct experiments on a genome-wide scale and for 303 TFs.

7. Short summaries of main conclusions

The following paragraphs contain major conclusions from the studies presented above and describe how the studies are linked together with respect to the central concept of my thesis, i.e. understanding transcriptional gene regulation and development of a new method to infer gene regulatory interactions (RIs). For additional discussion, please refer to the attached original publications in section 9.

7.1. Neuroblastomas with genomic 11q aberrations fall into two distinct subtypes depending on the clinical outcome: a revised model of for tumorigenesis (Publication I)

In a model outlined by Brodeur [1], neuroblastomas fall into two different subtypes: Type 1 tumors exhibit numerous CNAs and have the ability to regress spontaneously or to differentiate into benign ganglioneuroma, whereas type 2 tumors follow an aggressive clinical course and are mainly characterized by either *MYCN* amplification (type 2A) or loss of 11q (type 2B). In this work, we combined clinical data with gene expression profiling, cytogenetic and epigenetic analyses to investigate if tumors with 11q CNAs formed a distinct clinico-genetic subgroup.

We found that neuroblastomas with 11q CNAs are actually divided into two subgroups with different phenotypes and global gene expression patterns. When compared to tumors without 11q CNAs, surprisingly, there was no contrast between the gene expression profiles of favorable neuroblastomas with and without 11q CNAs. On the contrary, global gene expression of neuroblastomas with unfavorable phenotype and 11q CNAs deviated much from unfavorable neuroblastomas with normal 11q karyotype. We realized that loss of 11q affects the expression of genes located on 11q only in neuroblastomas with adverse outcome. In favorable tumors, the events leading to such changes in gene expression are apparently compensated by a yet unknown molecular mechanism. This led us to conclude that loss of 11q is insufficient to change gene expression in neuroblastomas that are characteristic of an aggressive phenotype. Instead, the events causing malignant transformation of neuroblastomas are likely to occur previously to the acquisition of loss of 11q and possibly other CNAs (see Figure 7.1). Further in line with our modified model is the fact that favorable phenotypes with low primary risk hardly ever turn unfavorable [1]. In unfavorable neuroblastomas, 11q CNAs contribute specific properties to

the phenotype that stick out on the gene expression level. Yet we observed a large overlap of differentially expressed genes between unfavorable and favorable phenotypes regardless of the 11q status which indicates a common mechanism of tumorigenesis in these distinct phenotypes.

In general, our study demonstrated that the occurrence of a prognostically important genomic CNA does not necessarily imply a homogeneous clinico-genetic subgroup, and illustrates how integration of data comprising several different aspects can lead to a revised tumorigenesis model.

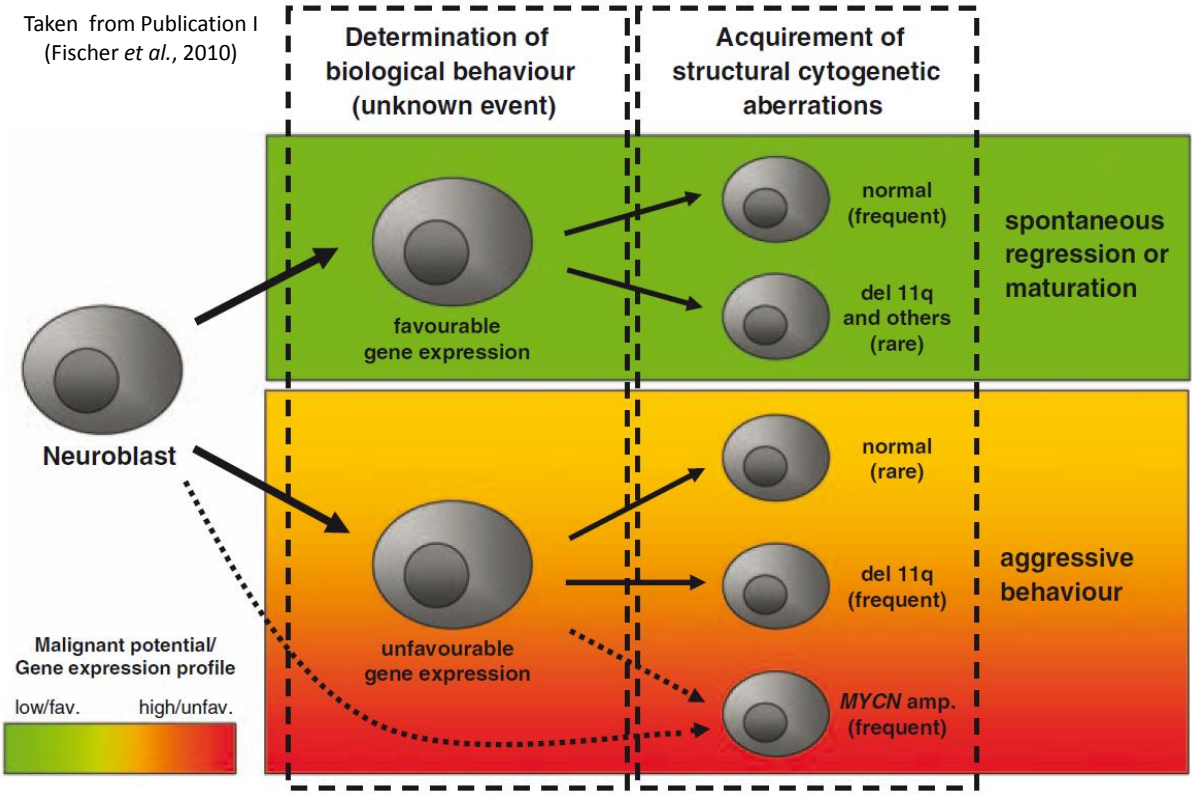


Figure 7.1 | **Revised model of neuroblastoma tumorigenesis modified according to Brodeur [1].** In the early steps of tumorigenesis, clinical and biological phenotypes are formed with distinct gene expression patterns. Structural genomic alterations occur during later course and are more frequent in unfavorable neuroblastoma, where they may contribute to specific properties of the tumors. In the favorable subgroup, genomic alterations occur occasionally only and remain silent events. Amplification of *MYCN* may represent an early event of tumorigenesis as an exception.

As gene expression levels of genes located on 11q are apparently not only affected by corresponding 11q CNAs, other transcriptional regulatory effects must be responsible for the transcriptional differences between neuroblastoma subtypes with different outcome. We found evidence that these changes are not due to epigenetic modifications of genomic DNA, therefore a significant role of TFs as key regulators that mediate the distinct gene expression

profiles is a likely scenario that needs further exploration. We were able to show the impact of CAMTA1, as well as MYC and MYCN (even in *MYCN* non-amplified neuroblastomas) in follow-up studies (see next paragraphs), but it is likely that there exist even more components in neuroblastoma regulatory networks that need to be considered when deciphering neuroblastoma tumorigenesis.

Essentially, malignant transformation of neuroblastomas in combination with 11q CNA has a strong impact that is reflected by differential gene expression levels, so the gene expression data must conceal the required information to a considerable extent. With regard to understanding gene RIs, I concluded that a method delivering TF candidates with high stringency is urgently required, but for this additional experiments covering other aspects of gene regulation were mandatory.

7.2. CAMTA1 TF acts as a tumor suppressor in neuroblastomas and affects cell cycle progression and neuronal differentiation (Publication II)

CAMTA1 is constantly expressed at lower levels in aggressive neuroblastomas compared to more benign phenotypes and constitutes a prognostic marker of poor outcome. From this we concluded that down-regulation of *CAMTA1* grants malignant neuroblastoma cells a selective advantage which may be reversed by re-expression of *CAMTA1*. We found experimentally that induction of *CAMTA1* in human neuroblastoma cell lines with low endogenous *CAMTA1* expression impairs colony formation and reduces growth rates, and leads to reduced tumor size in *in vivo* mouse models.

The considerably high expression levels of *CAMTA1* in favorable neuroblastomas of stage 1, 2, and 4S indicated that CAMTA1 acts as a tumor suppressor gene. Time-series gene expression profiles of neuroblastoma cell lines were employed to test this hypothesis. I extrapolated genes affected by induction of *CAMTA1* with a modified gene filtering approach prior to clustering. This approach identified subsets of genes with different gene expression profiles in response to CAMTA1 induction. Two of five identified clusters were up-regulated either constantly or transiently whereas another cluster exhibited constantly decreasing gene expression levels upon CAMTA1 induction (in comparison to the non-induced control). Furthermore, a GO term enrichment analysis revealed several functional attributes of genes from individual clusters that fit well with the proposed role of CAMTA1, particularly in neuron differentiation and cell cycle control. In turn, these findings were confirmed in a subsequent quantitative gene expression analysis of five CAMTA1-regulated genes with specific GO annotations in an independent neuroblastoma cell line.

Intracellular Ca^{2+} levels show pleiotropic effects in neuron differentiation and Ca^{2+} influx can induce neuritic outgrowth of neuroblastoma cell lines [93]. CAMTA1 responds to

Ca^{2+} signaling by binding of calmodulin. The GO term analyses further suggested that CAMTA1 affects neuronal differentiation both by integrating and mediating Ca^{2+} signaling.

As a tumor suppressor candidate, CAMTA1 may provide a good therapeutic target and deserves more attention in research of strategies to treat malignant neuroblastomas.

Finally, gene expression profiling proved to be a valuable technique to deduct properties of a TF and correlated (and co-regulated) target genes. In my main project, I used these principles to reconstruct RIs on a large-scale. Nevertheless, in order to distinguish between background noise contained in the gene expression data, and generally valid RIs, I needed to incorporate both numerous sets of gene expression profiles from different conditions and an independent source for evidence of putative RIs into my approach to RI reconstruction.

7.3. MYC and MYCN affect distinct gene expression profiles in different neuroblastoma subtypes without *MYCN* amplification (Publication III)

MYCN amplification is an established marker of poor outcome in neuroblastoma: It is associated with malignant progression and tumorigenesis. Furthermore, MYCN expression frequently correlates inversely to MYC expression. Our findings implicate that MYC and MYCN also play an active part in neuroblastomas without *MYCN* amplification, particularly in stage 4-NA, but also in stage 4S-NA. We were able to define a core set of putative MYC/MYCN-regulated target genes by gene expression profiling. Our findings suggest both mutual and distinct roles of MYC and MYCN as putative key transcriptional regulators in several malignant neuroblastoma subtypes without *MYCN* amplification. To this end, exclusive or cumulative mechanisms of action come into consideration, depending on the observed expression levels of MYC and MYCN and their target genes with respect to the corresponding subtype.

In pursue of the working hypothesis, I applied bioinformatics methods for predicting TFBSs in DNA sequences and established an approach for identifying over-represented TFBSs in promoters of a gene set. This approach is applicable to a large number of TFs with known binding motifs and holds a great potential to identify key TFs in any condition where gene expression profiling can be conducted, e .g. in cancer research.

In this study [64], the results from our analyses were further validated by ChIP experiments. The defined MYCN/MYC target gene set served as an indicator of MYC/MYCN activity in neuroblastoma subtypes. We observed a considerable induction of the target gene set in stage 4-NA and linked it to MYC activity. To a lesser extent, the targets were induced by MYCN in stage 4S-NA tumors. We concluded therefore that MYC is a stronger transactivator in stage 4-NA tumors than MYCN in stage 4S-NA, which may be related to the more favorable nature of the ladder subtype and is in line with the antagonistic roles

described for MYCN in tumorigenesis and apoptosis. Additionally, high expression levels of MYC/MYCN target genes were associated with poor outcome of neuroblastoma patients with and without adjustment for co-variables.

7.4. The Rage-dependent regulatory network in a tumor-promoting inflammatory context (Publication IV)

In essence, we were able to further dissect how Rage signaling affects long-term dynamics of gene regulation involved in inflammation and potentially in carcinogenesis. We were able to identify target genes of Rage signaling and predicted potent mediators of the signal on the TF level. Furthermore, we proposed different transcriptional regulators for Rage-responsive clusters with distinct dynamic gene expression profiles, and thereby provided key components for a core model of Rage signaling.

This project exemplified the benefits of iterative application of wet lab experiments and bioinformatics analyses. My *in silico* analyses provided candidate TFs (E2f, Sp1, Sp4, Hnf4, Mazr, CAC-bp, and Wt1) controlling Rage-dependent transcriptional changes in murine keratinocytes, and these candidates were subsequently validated *in vivo* as components of the regulatory network affected by Rage signaling.

The successful application of genome-wide PWM scans was substantial for this work and provided a solid basis for prediction of transcriptional regulation. We improved these predictions for general application by in-depth analyses and reconstruction of regulatory networks. In the context of Rage signaling, we were able to confirm the principle that genes that are affected transcriptionally by certain TFs contain enrichments of regulatory sequences in a region adjacent to the TSS that can be detected by binding motif scans. In combination with gene expression profiling, such results provide independent and unbiased evidence that can be used to deduce RIs between TFs and potential target genes.

We further demonstrated that the identification of key TFs provides meaningful hypotheses for the understanding of cellular systems that can be tested directly in follow-up experiments.

7.5. RIP – a powerful tool to predict regulatory interactions with multiple applications (Publication V)

We developed the regulatory interaction predictor (RIP), a novel machine learning approach to predict gene regulation on a genome-wide scale. The predictions resulted from experiments spanning a wide range of biological conditions and can be applied to more specific conditions as well. The RIP classifier reached considerably high precision and recall and outperformed other comparable methods.

RIP predicted 6073 RIs at 44.0% confidence and 73 923 RIs at ≥31.5% confidence. These figures appear reasonable considering the dimensions of the input (949 genes with 2896 known RIs) in relation the number of candidate genes (13 069), and a large number of predicted RIs proved to be valuable in various applications.

I established a comprehensive promoter analysis that supports the prediction of TFBS and TFs potentially regulating a set of genes. I critically assessed the validity of several assumptions that need to be considered for regulatory network reconstruction. I confirmed that cooperation in specific biological functions is reflected by (frequent yet condition-specific) correlation of co-regulated genes. Furthermore, I showed that correlation meta-analysis can be used efficiently to improve the prediction of general and condition-specific co-regulation of genes by applying correlation filters. Additionally, I found that TF mRNA gradients do not correlate to mRNA gradients of their target genes in a wide range of experimental conditions.

We succeeded in transferring the described relationships of gene regulation into quantifiable features to infer gene regulation. I found that these features distinguished True RIs from True non-RIs. Subsequently, I trained SVMs to predict novel RIs by combining descriptors of 1) a correlation meta-analysis of 4064 gene expression profiles from 76 different experiments and conditions, 2) TFBS predictions from PWM scans, and 3) association of co-regulation, correlation and TFBS predictions employing a set of known RIs from Transfac database.

RIs predicted by RIP effectively identified key regulators of IFNα signaling, TFs that are associated with pathways, and a considerable portion of RIs from an independent database (TRED).

The algorithm behind RIP is generic and can easily be extended by other TFs, for which a PWM, and (favorably but not necessarily) a set of target genes is available. The method can be transferred and applied to predict RIs for other species than human. Furthermore, the features may even be modified and improved in future versions of the classifier.

The presented RIP classifier is available to the public as a software package and offers a wide range of applications for gene expression analyses, such as identification of key TFs and pathways involved in the pathology and changed function of the investigated cells.

RIP has the potential to fill gaps in the understanding of regulatory networks of cancer entities like neuroblastomas.

My aim in the future is to apply RIP to define key TFs of various tumor types and gain a better mechanistic insight into their pathogenesis. Additionally, I want to further improve the algorithm and evaluate its potential in reconstructing condition-specific RIs, and to tune it to deliver stringent predictions with such high precision that individual RIs can be

considered in models of carcinogenesis and other experiments. Finally, genomic deep sequencing approaches, high-throughput protein data, and high content DNA methylation analyses are promising technical developments that I want to employ to extend our method.

8. Bibliography

1. Brodeur GM (2003) Neuroblastoma: biological insights into a clinical enigma. Nat Rev Cancer 3: 203-216.
2. Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24: 238-241.
3. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33: W741-748.
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
5. Weinberg RA (2007) The biology of cancer. New York: Garland Science. 1 v. (various pagings) p.
6. Alberts B, Wilson JH, Hunt T (2008) Molecular biology of the cell. New York: Garland Science. xxxiii, 1601, [1690] p. p.
7. Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. Mol Cell 10: 1467-1477.
8. Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424: 147-151.
9. Davidson EH (2001) Genomic regulatory systems : development and evolution. San Diego: Academic Press. xii, 261 p. p.
10. Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, et al. (2002) The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci U S A 99 Suppl 4: 16433-16437.
11. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117: 185-198.
12. Raven PH, Johnson GB (2002) Biology. Boston: McGraw-Hill.
13. Meng G, Mosig A, Vingron M A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes. BMC Bioinformatics 11: 267.
14. Segal E, Yelensky R, Koller D (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19 Suppl 1: i273-282.
15. Sinha S (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics 22: e454-463.
16. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.
17. Samsonova AA, Niranjan M, Russell S, Brazma A (2007) Prediction of gene expression in embryonic structures of Drosophila melanogaster. PLoS Comput Biol 3: e144.
18. Bonneau R (2008) Learning biological networks: from modules to dynamics. Nat Chem Biol 4: 658-664.
19. Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. Science 325: 429-432.
20. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. Science 307: 1962-1965.
21. Kuhlman T, Zhang Z, Saier MH, Jr., Hwa T (2007) Combinatorial transcriptional control of the lactose operon of Escherichia coli. Proc Natl Acad Sci U S A 104: 6043-6048.
22. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7: 601-620.

23. Das D, Nahle Z, Zhang MQ (2006) Adaptively inferring human transcriptional subnetworks. Mol Syst Biol 2: 2006 0029.
24. Nguyen DH, D'Haeseleer P (2006) Deciphering principles of transcription regulation in eukaryotic genomes. Mol Syst Biol 2: 2006 0012.
25. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, et al. (2007) A predictive model for transcriptional control of physiology in a free living cell. Cell 131: 1354-1365.
26. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8.
27. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, et al. (2003) Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21: 1337-1342.
28. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T (2009) Module networks revisited: computational assessment and prioritization of model predictions. Bioinformatics 25: 490-496.
29. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166-176.
30. Taylor RC, Acquaah-Mensah G, Singhal M, Malhotra D, Biswal S (2008) Network inference algorithms elucidate Nrf2 regulation of mouse lung oxidative stress. PLoS Comput Biol 4: e1000166.
31. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, et al. (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37: 382-390.
32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1: S7.
33. Everett LJ, Jensen ST, Hannenhalli S (2011) Transcriptional regulation via TF-modifying enzymes: an integrative model-based analysis. Nucleic Acids Res 39: e78.
34. Vapnik VN (1995) The nature of statistical learning theory. New York: Springer. xv, 188 p. p.
35. Vapnik VN (1998) Statistical learning theory. New York: Wiley. xxiv, 736 p. p.
36. Witten IH, Frank E, Holmes G, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques: Morgan Kaufmann. 629 p.
37. Schölkopf B, Smola AJ (2002) Learning with kernels : support vector machines, regularization, optimization, and beyond. Cambridge, Mass.: MIT Press. xviii, 626 p. p.
38. Vapnik VN (1982) Estimation of dependences based on empirical data. New York: Springer-Verlag. xvi, 399 p. p.
39. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers; 1992; Pittsburgh, Pennsylvania, United States. ACM. pp. 144-152.
40. Fischer M, Bauer T, Oberthuer A, Hero B, Theissen J, et al. (2009) Integrated genomic profiling identifies two distinct molecular subtypes with divergent outcome in neuroblastoma with loss of chromosome 11q. Oncogene 29: 865-875.
41. Fischer M, Spitz R, Oberthur A, Westermann F, Berthold F (2008) Risk estimation of neuroblastoma patients using molecular markers. Klin Padiatr 220: 137-146.
42. Savelyeva L, Schwab M (2001) Amplification of oncogenes revisited: from expression profiling to clinical application. Cancer Lett 167: 115-123.
43. Attiyeh EF, London WB, Mosse YP, Wang Q, Winter C, et al. (2005) Chromosome 1p and 11q deletions and outcome in neuroblastoma. N Engl J Med 353: 2243-2253.

44. Spitz R, Hero B, Simon T, Berthold F (2006) Loss in chromosome 11q identifies tumors with increased risk for metastatic relapses in localized and 4S neuroblastoma. Clin Cancer Res 12: 3368-3373.
45. Cohn SL, Pearson AD, London WB, Monclair T, Ambros PF, et al. (2009) The International Neuroblastoma Risk Group (INRG) classification system: an INRG Task Force report. J Clin Oncol 27: 289-297.
46. Maris JM, Hogarty MD, Bagatell R, Cohn SL (2007) Neuroblastoma. Lancet 369: 2106-2120.
47. Picard JL, Data LC, Riker WT, LaForge J, Troi D, et al. (1987) To boldly go where no one has gone before. United Fed Plan 1: 1-10.
48. Bilke S, Chen QR, Westerman F, Schwab M, Catchpoole D, et al. (2005) Inferring a tumor progression model for neuroblastoma from genomic data. J Clin Oncol 23: 7322-7331.
49. Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, et al. (2006) Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. J Clin Oncol 24: 5070-5078.
50. Classen S, Zander T, Eggle D, Chemnitz JM, Brors B, et al. (2007) Human resting CD4+ T cells are constitutively inhibited by TGF beta under steady-state conditions. J Immunol 178: 6931-6940.
51. Berwanger B, Hartmann O, Bergmann E, Bernard S, Nielsen D, et al. (2002) Loss of a FYN-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. Cancer Cell 2: 377-386.
52. Fischer M, Oberthuer A, Brors B, Kahlert Y, Skowron M, et al. (2006) Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome. Clin Cancer Res 12: 5118-5128.
53. Henrich KO, Fischer M, Mertens D, Benner A, Wiedemeyer R, et al. (2006) Reduced expression of CAMTA1 correlates with adverse outcome in neuroblastoma patients. Clin Cancer Res 12: 131-138.
54. Henrich KO, Bauer T, Schulte J, Ehemann V, Deubzer H, et al. (2011) CAMTA1, a 1p36 tumor suppressor candidate, inhibits growth and activates differentiation programs in neuroblastoma cells. Cancer Res 71: 3142-3151.
55. Wang ZC, Lin M, Wei LJ, Li C, Miron A, et al. (2004) Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. Cancer Res 64: 64-71.
56. Kim MY, Yim SH, Kwon MS, Kim TM, Shin SH, et al. (2006) Recurrent genomic alterations with impact on survival in colorectal cancer identified by genome-wide array comparative genomic hybridization. Gastroenterology 131: 1913-1924.
57. Barbashina V, Salazar P, Holland EC, Rosenblum MK, Ladanyi M (2005) Allelic losses at 1p36 and 19q13 in gliomas: correlation with histologic classification, definition of a 150-kb minimal deleted region on 1p36, and evaluation of CAMTA1 as a candidate tumor suppressor gene. Clin Cancer Res 11: 1119-1128.
58. Smedley D, Sidhar S, Birdsall S, Bennett D, Herlyn M, et al. (2000) Characterization of chromosome 1 abnormalities in malignant melanomas. Genes Chromosomes Cancer 28: 121-125.
59. Nowacki S, Skowron M, Oberthuer A, Fagin A, Voth H, et al. (2008) Expression of the tumour suppressor gene CADM1 is associated with favourable outcome and inhibits cell survival in neuroblastoma. Oncogene 27: 3329-3338.

60. Bauer A, Savelyeva L, Claas A, Praml C, Berthold F, et al. (2001) Smallest region of overlapping deletion in 1p36 in human neuroblastoma: a 1 Mbp cosmid and PAC contig. Genes Chromosomes Cancer 31: 228-239.
61. White PS, Thompson PM, Gotoh T, Okawa ER, Igarashi J, et al. (2005) Definition and characterization of a region of 1p36.3 consistently deleted in neuroblastoma. Oncogene 24: 2684-2694.
62. von Heydebreck A, Huber W, Gentleman R (2004) Differential Expression with the Bioconductor Project. Bioconductor Project Working Papers.
63. Duncan D, Prodduturi N, Zhang B (2010) WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. BMC Bioinformatics 11(Suppl. 4): P10.
64. Westermann F, Muth D, Benner A, Bauer T, Henrich KO, et al. (2008) Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. Genome Biol 9: R150.
65. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646-674.
66. Adhikary S, Eilers M (2005) Transcriptional regulation and transformation by Myc proteins. Nat Rev Mol Cell Biol 6: 635-645.
67. Prochownik EV, Li Y (2007) The ever expanding role for c-Myc in promoting genomic instability. Cell Cycle 6: 1024-1029.
68. Schwab M, Varmus HE, Bishop JM (1985) Human N-myc gene contributes to neoplastic transformation of mammalian cells in culture. Nature 316: 160-162.
69. Weiss WA, Aldape K, Mohapatra G, Feuerstein BG, Bishop JM (1997) Targeted expression of MYCN causes neuroblastoma in transgenic mice. EMBO J 16: 2985-2995.
70. Cohn SL, London WB, Huang D, Katzenstein HM, Salwen HR, et al. (2000) MYCN expression is not prognostic of adverse outcome in advanced-stage neuroblastoma with nonamplified MYCN. J Clin Oncol 18: 3604-3613.
71. Tang XX, Zhao H, Kung B, Kim DY, Hicks SL, et al. (2006) The MYCN enigma: significance of MYCN expression in neuroblastoma. Cancer Res 66: 2826-2833.
72. Westermann F, Henrich KO, Wei JS, Lutz W, Fischer M, et al. (2007) High Skp2 expression characterizes high-risk neuroblastomas independent of MYCN status. Clin Cancer Res 13: 4695-4703.
73. Fulda S, Lutz W, Schwab M, Debatin KM (1999) MycN sensitizes neuroblastoma cells for drug-induced apoptosis. Oncogene 18: 1479-1486.
74. Edsjo A, Nilsson H, Vandesompele J, Karlsson J, Pattyn F, et al. (2004) Neuroblastoma cells with overexpressed MYCN retain their capacity to undergo neuronal differentiation. Lab Invest 84: 406-417.
75. Breit S, Schwab M (1989) Suppression of MYC by high expression of NMYC in human neuroblastoma cells. J Neurosci Res 24: 21-28.
76. Riehl A, Bauer T, Brors B, Busch H, Mark R, et al. (2010) Identification of the Rage-dependent gene regulatory network in a mouse model of skin inflammation. BMC Genomics 11: 537.
77. Gebhardt C, Riehl A, Durchdewald M, Nemeth J, Furstenberger G, et al. (2008) RAGE signaling sustains inflammation and promotes tumor development. J Exp Med 205: 275-285.
78. Clynes R, Moser B, Yan SF, Ramasamy R, Herold K, et al. (2007) Receptor for AGE (RAGE): weaving tangled webs within the inflammatory response. Curr Mol Med 7: 743-751.

79. Kang R, Tang D, Schapiro NE, Livesey KM, Farkas A, et al. (2010) The receptor for advanced glycation end products (RAGE) sustains autophagy and limits apoptosis, promoting pancreatic tumor cell survival. Cell Death Differ 17: 666-676.
80. Logsdon CD, Fuentes MK, Huang EH, Arumugam T (2007) RAGE and RAGE ligands in cancer. Curr Mol Med 7: 777-789.
81. Taguchi A, Blood DC, del Toro G, Canet A, Lee DC, et al. (2000) Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases. Nature 405: 354-360.
82. Bauer T, Eils R, Konig R (2011) RIP: the regulatory interaction predictor--a machine learning-based approach for predicting target genes of transcription factors. Bioinformatics 27: 2239-2247.
83. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
84. Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, et al. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat Biotechnol 23: 238-243.
85. Tassiulas I, Hu X, Ho H, Kashyap Y, Paik P, et al. (2004) Amplification of IFN-alpha-induced STAT1 activation and inflammatory function by Syk and ITAM-containing adaptors. Nat Immunol 5: 1181-1189.
86. Brach MA, Arnold C, Kiehntopf M, Gruss HJ, Herrmann F (1993) Transcriptional activation of the macrophage colony-stimulating factor gene by IL-2 is associated with secretion of bioactive macrophage colony-stimulating factor protein by monocytes and involves activation of the transcription factor NF-kappa B. J Immunol 150: 5535-5543.
87. Friedman AD (2007) Transcriptional control of granulocyte and monocyte development. Oncogene 26: 6816-6828.
88. Grenningloh R, Kang BY, Ho IC (2005) Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. J Exp Med 201: 615-626.
89. Honda K, Taniguchi T (2006) IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat Rev Immunol 6: 644-658.
90. Yu H, Pardoll D, Jove R (2009) STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer 9: 798-809.
91. Fu XY, Kessler DS, Veals SA, Levy DE, Darnell JE, Jr. (1990) ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains. Proc Natl Acad Sci U S A 87: 8555-8559.
92. Lallemand C, Blanchard B, Palmieri M, Lebon P, May E, et al. (2007) Single-stranded RNA viruses inactivate the transcriptional activity of p53 but induce NOXA-dependent apoptosis via post-translational modifications of IRF-1, IRF-3 and CREB. Oncogene 26: 328-338.
93. Wu G, Fang Y, Lu ZH, Ledeen RW (1998) Induction of axon-like and dendrite-like processes in neuroblastoma cells. J Neurocytol 27: 1-14.

9. Own publications

Oncogene (2010) 29, 865–875

www.nature.com/onc

ORIGINAL ARTICLE

Integrated genomic profiling identifies two distinct molecular subtypes with divergent outcome in neuroblastoma with loss of chromosome 11q

M Fischer[1], T Bauer[2,3], A Oberthür[1], B Hero[1], J Theissen[1], M Ehrich[4], R Spitz[1], R Eils[2,3], F Westermann[5], B Brors[3], R König[2,3] and F Berthold[1]

[1]Department of Paediatric Oncology, University Children's Hospital, and Center for Molecular Medicine Cologne (CMMC), Cologne, Germany; [2]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Bioquant, Heidelberg, Germany; [3]Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany; [4]SEQUENOM, Inc., San Diego, CA, USA and [5]Department of Tumourbiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Imbalances in chromosome 11q occur in approximately 30% of primary neuroblastoma and are associated with poor outcome. It has been suggested that 11q loss constitutes a distinct clinico-genetic neuroblastoma subgroup by affecting expression levels of corresponding genes. This study analysed the relationship of 11q loss, clinical phenotype and global transcriptomic profiles in four clinico-genetic subgroups (11q alteration/favourable outcome, $n=7$; 11q alteration/unfavourable outcome, $n=14$; no 11q alteration/favourable outcome, $n=81$; no 11q alteration/unfavourable outcome, $n=8$; tumours with *MYCN* amplification and/or 1p loss were excluded). Unsupervised and supervised comparisons of gene expression profiles consistently showed significantly different mRNA patterns between favourable and unfavourable neuroblastomas, both in the subgroups with and without 11q loss. In contrast, favourable tumours with and without 11q loss showed highly similar transcriptomic profiles. Disproportionate downregulation of 11q genes was observed only in unfavourable tumours with 11q loss. The diverging molecular profiles were neither caused by considerable differences in the size of the deleted regions nor by differential methylation patterns of 11q genes. Together, this study shows that neuroblastoma with 11q loss comprises two biological subgroups that differ both in their clinical phenotype and gene expression patterns, indicating that 11q loss is not a primary determinant of neuroblastoma tumour behaviour.
Oncogene (2010) **29,** 865–875; doi:10.1038/onc.2009.390; published online 9 November 2009

Keywords: neuroblastoma; integrative genomics; loss of 11q; gene expression; outcome; cancer

Correspondence: Dr M Fischer, Department of Paediatric Oncology, University Children's Hospital, Kerpener Street 62, Cologne, NRW 50924, Germany.
E-mail: matthias.fischer@uk-koeln.de
Received 13 May 2009; revised 30 September 2009; accepted 8 October 2009; published online 9 November 2009

Introduction

Human cancer genomes are characterized by multiple genetic aberrations and epigenetic modifications. Recurrent DNA copy number alterations (CNA) in malignant cells are thought to be critical events in human tumourigenesis, and have been suggested to determine the biological phenotype of the tumour by changing the expression of cancer genes located at the affected sites. The effects of oncogene amplification on mRNA expression levels have been well characterized in various entities and serve as a paradigm for a direct relationship of gene dosage and expression levels in cancer (Savelyeva and Schwab, 2001). However, the influence of low-level copy number gains (<fivefold change) and hemizygous losses of large genomic regions on expression levels of the corresponding genes and the global transcriptome is less clear, and its delineation has remained a challenge of cancer research.

In neuroblastoma, several non-random genomic alterations have been described to be closely associated with distinct phenotypes of the disease (reviewed in Fischer *et al*., 2008). This paediatric tumour may therefore represent a valuable model to analyse the interactions of CNA and transcriptomic aberrations. A hallmark of neuroblastoma is its biological and clinical heterogeneity, ranging from spontaneous regression of the tumour to relentless progression with fatal outcome of the patients. At diagnosis, these two contrasting subtypes can be largely distinguished by specific chromosomal alterations. Amplification of the oncogene *MYCN* occurs in ~20% of neuroblastomas and is strongly associated with a poor prognosis. More recently, loss of 11q has been reported to be highly correlated with an adverse patients' outcome (Attiyeh *et al*., 2005; Spitz *et al*., 2006a), and has thus been proposed as a stratifying prognostic marker in the International Neuroblastoma Risk Group classification system (Cohn *et al*., 2009) as well as in the upcoming clinical trial of the Children's Oncology Group (Maris *et al*., 2007). As 11q CNA and *MYCN* amplification are inversely correlated in neuroblastoma, these two genomic alterations have been suggested to delineate two

molecularly distinct subgroups (Brodeur, 2003; Bilke *et al*., 2005). In contrast, the majority of favourable neuroblastomas lack structural genomic aberrations but show numerical variations of whole chromosomes (Spitz *et al*., 2006b; Janoueix-Lerosey *et al*., 2009). In addition to genomic alterations, the divergent biological and clinical neuroblastoma subtypes show distinct gene expression profiles (reviewed in Fischer *et al*., 2008). In a previous study, we have described a prognostic 144-gene expression classifier for neuroblastoma patients using the prediction analysis for microarrays algorithm (PAM) that predicts outcome of neuroblastoma patients with high accuracy (Oberthuer *et al*., 2006).

Together, there is convincing evidence today that both genomic and transcriptomic alterations are associated with the diverging clinical phenotypes of neuroblastoma. However, whereas the effect of *MYCN* amplification on the transcriptome has been well characterized (Boon *et al*., 2001; Alaminos *et al*., 2003; Westermann *et al*., 2008), little is known about the interactions of gains or deletions of large genomic regions and gene expression patterns in neuroblastoma. In this study, we aimed to determine the effect of 11q loss on global gene expression patterns and clinical phenotypes in neuroblastoma. For this purpose, we used various bioinformatics strategies in an integrative genomics approach, for which we considered clinical information of neuroblastoma patients as well as whole genome expression profiles, cytogenetic characteristics and promoter methylation data of the tumours.

Results

Gene expression-based classification identifies two prognostically distinct subgroups of neuroblastoma patients with loss of 11q

To determine the predictive power of our previously defined 144-gene expression-based PAM classifier (Oberthuer *et al*., 2006) in the subset of neuroblastoma with loss of 11q, 61 tumours that had not been included in the original training set and that showed a CNA at 11q were analysed. The PAM classifier predicted 20 patients to be favourable and 41 patients to be unfavourable. Event-free and overall survival of patients with favourable PAM prediction were significantly better than those with an unfavourable prediction (event-free survival at 5 years, 0.79 ± 0.09 vs 0.27 ± 0.09, $P = 0.001$; overall survival at 5 years, 0.95 ± 0.05 vs 0.64 ± 0.09, $P = 0.013$; Figures 1a and b). After exclusion of patients with *MYCN* amplification ($n = 6$) from the analysis, the event-free survival of PAM favourable and unfavourable tumours was still significantly different (0.79 ± 0.09 vs 0.31 ± 0.10 at 5 years, $P = 0.005$, Figure 1c), whereas there was a trend towards a significantly differing overall survival (0.95 ± 0.05 vs 0.75 ± 0.09 at 5 years, $P = 0.061$, Figure 1d). These data strongly suggest that neuroblastoma with loss of 11q comprises two distinct subsets with different clinical courses, which is mirrored by distinct gene expression patterns.

Global gene expression patterns differ in favourable and unfavourable neuroblastomas with loss of 11q

We next analysed the relationship of clinical phenotypes, 11q aberrations and global gene expression patterns of the tumours. For this purpose, four neuroblastoma subgroups were defined according to 11q status (normal vs deletion/imbalance as determined using fluorescence *in situ* hybridization (FISH)) and the outcome of patients (event-free survivors for at least 2 years without any cytotoxic treatment vs metastatic or multiple loco-regional progression/relapse or death of disease). These clinical parameters were chosen to avoid biased results elicited by possible treatment influences on the natural courses of the disease. Tumours showing *MYCN* amplification and/or loss of 1p were excluded from the study because of their known or potential effect on gene expression profiles and the outcome of patients. According to these criteria, a total of 110 samples were selected from the cohort of our previous study (Oberthuer *et al*., 2006).

At first, we applied unsupervised algorithm methods (principal component analysis and hierarchical clustering of unfiltered gene expression data) to determine the effects of the clinical phenotype and the 11q status on gene expression profiles independently of the clinico-genetic classification. Both principal component analysis and hierarchical clustering revealed that tumours group primarily according to their clinical phenotype (Figures 2a and b). Most notably, favourable tumours with 11q loss (del11q_fav) did not associate with unfavourable tumours with 11q loss (del11q_unfav) but with favourable tumours without 11q loss (normal_fav). A discriminatory effect of 11q CNA was observed only within the subgroup of unfavourable tumours, but not within the subgroup of favourable tumours. Unfavourable tumours without 11q loss (normal_unfav) were more associated with del11q_unfav than with normal_fav samples but did not build a delimited cluster on their own.

We next aimed to objectify these observations in a supervised analysis approach. Every subgroup was compared with each other by applying a method termed analysis of centroid distances, which is suitable to measure differences in overall gene expression patterns (Classen *et al*., 2007). We calculated the pairwise Euclidean distances between all possible pairs of subgroups and determined their significance using permutation analysis. Highly significant differences ($P < 0.001$ for each comparison, Figure 3) were observed between all groups except for the pair del11q_fav versus normal_fav ($P = 0.19$). This finding corroborates the results of the unsupervised analyses, suggesting that favourable neuroblastomas with and without 11q loss do not differ in their overall gene expression profiles.

Identification of genes that are differentially expressed between clinico-genetic neuroblastoma subgroups

To identify genes that are differentially expressed between the clinico-genetic subgroups, we performed significance analysis of microarrays (SAM; Table 1 and

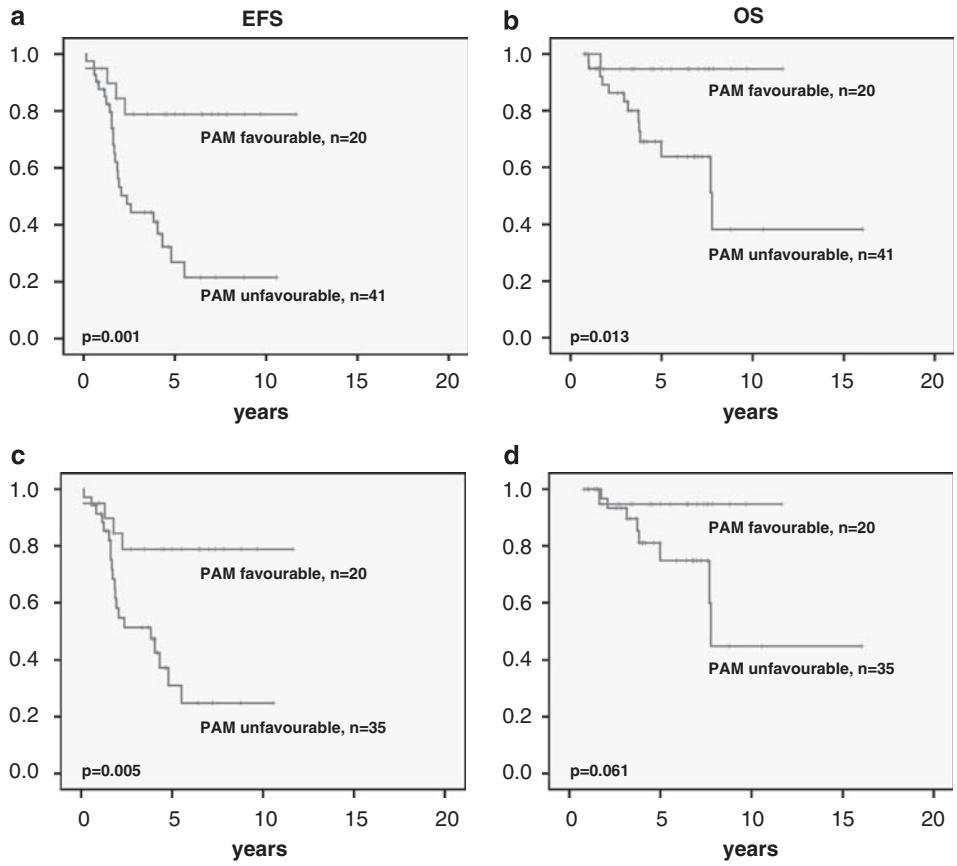


Figure 1 Kaplan–Meier estimates for event-free survival (EFS) and overall survival (OS) of neuroblastoma patients with 11q aberrations according to the prognostic prediction analysis for microarrays (PAM) gene signature in the whole cohort ($n=61$, **a** and **b**, respectively) and after exclusion of *MYCN* amplified tumours ($n=55$, **c** and **d**, respectively). Blue, favourable PAM prediction; red, unfavourable PAM prediction. A full colour version of this figure is available at the *Oncogene* journal online.

Supplementary Table 3) and analysis of variance (ANOVA). In line with our above mentioned observations, only two and three genes were differentially expressed between favourable tumours with and without 11q loss by SAM and ANOVA, respectively ($P<0.05$ each). In contrast, expression profiles of favourable and unfavourable tumours with loss of 11q differed in 282 and 227 transcripts using SAM and ANOVA, respectively. Comparison of favourable versus unfavourable tumours without 11q aberrations revealed the largest number of differentially expressed genes (1187 and 322 transcripts as determined using SAM and ANOVA, respectively). In the cohort of tumours from patients with unfavourable outcome, subgroups with and without loss of 11q differed by 64 and 69 genes using SAM and ANOVA, respectively.

On the basis of the SAM results, we next determined whether unfavourable tumours with and without 11q aberrations shared common features of gene expression in comparison with their favourable counterparts that may define their malignant phenotype. We observed 100 transcripts with a common differential expression between the subgroups normal_unfav versus normal_fav and del11q_unfav versus del11q_fav, corresponding to 35% of the mRNAs differing between the latter subtypes (Supplementary Figure 1a and Supplementary Table 4A). It is noteworthy that all of these genes were regulated in the same direction in both comparisons, indicating their biological relevance in malignant progression of neuroblastoma. With one exception (*HIST1H1C*), all transcripts had lower expression levels in the unfavourable tumours. Among these mRNAs, genes were detected that have been described to be differentially regulated in benign and adverse neuroblastoma previously, such as *FYN* (Berwanger *et al*., 2002), *MAP7* (Fischer *et al*., 2006; Oberthuer *et al*., 2006) and *CAMTA1* (Henrich *et al*., 2006). These data suggest that common mechanisms drive the malignant phenotype in both unfavourable neuroblastomas with and without 11q CNA.

We then asked whether expression levels of genes located on 11q were preferentially affected by loss of this chromosomal region using the SAM data (Table 1 and Supplementary Table 3). The two genes that were differentially expressed between del11q_fav and normal_fav tumours were not located on chromosome 11q. In contrast, 27 of the 64 mRNAs (42%) that were differentially expressed between unfavourable neuroblastomas with and without loss of 11q were located on 11q, which is a significant enrichment of genes at this

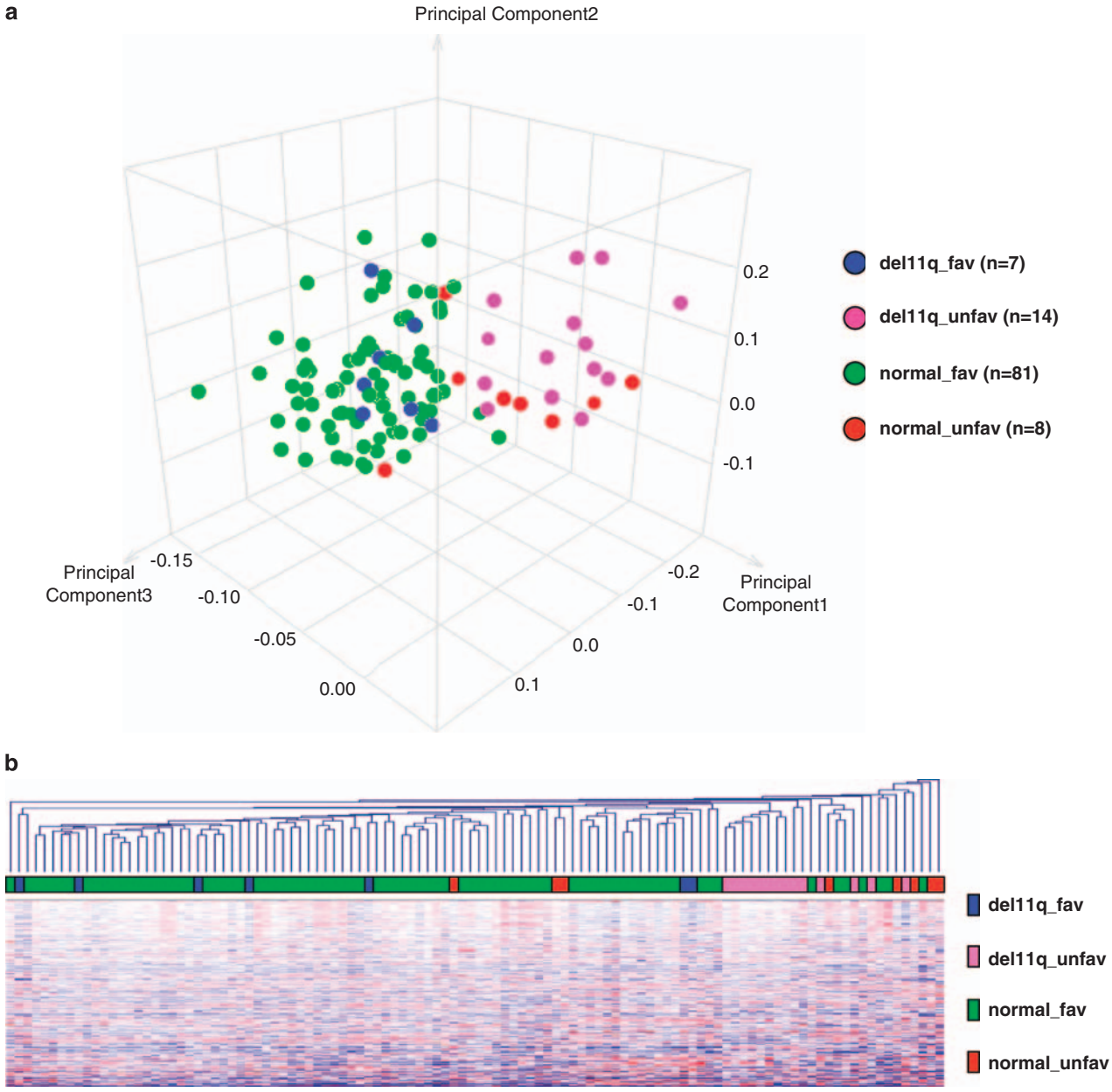


Figure 2 Principal component analysis plot (**a**), and hierarchical clustering of unfiltered gene expression data (**b**) of favourable and unfavourable neuroblastomas with and without 11q aberrations. For hierarchical clustering, Pearson's correlation and average linkage were used.

chromosomal region ($P<10^{-18}$). As expected, most of these transcripts were downregulated in tumours of the del11q_unfav subgroup (24/27 genes). Comparison of del11q_fav and del11q_unfav tumours revealed 10/282 genes that were located on 11q ($P=0.40$), whereas comparison of normal_fav and normal_unfav tumours revealed an under-representation of genes located on 11q (38/1187 genes, $P=0.01$). These results indicate that loss of 11q affects the expression of genes located at this region, but only in neuroblastoma with an unfavourable phenotype.

We finally examined the number of genes that were differentially expressed between del11q_unfav and normal_fav tumours. The number of transcripts that were differentially expressed between these diametrically opposed subgroups was high (2470 genes), and significantly enriched for genes located on 11q (176 genes, 7%; $P<10^{-5}$). We then determined the overlap of mRNAs that were differentially expressed between these subgroups and between the del11q_unfav versus del11q_fav tumours. If the latter tumours formed a homogeneous subgroup with normal_fav tumours, one would expect a large overlap of genes in these two comparisons. Indeed, 95% (269/282) of the genes of the comparison of del11q_unfav versus del11q_fav tumours were found in both analyses, all of which were regulated

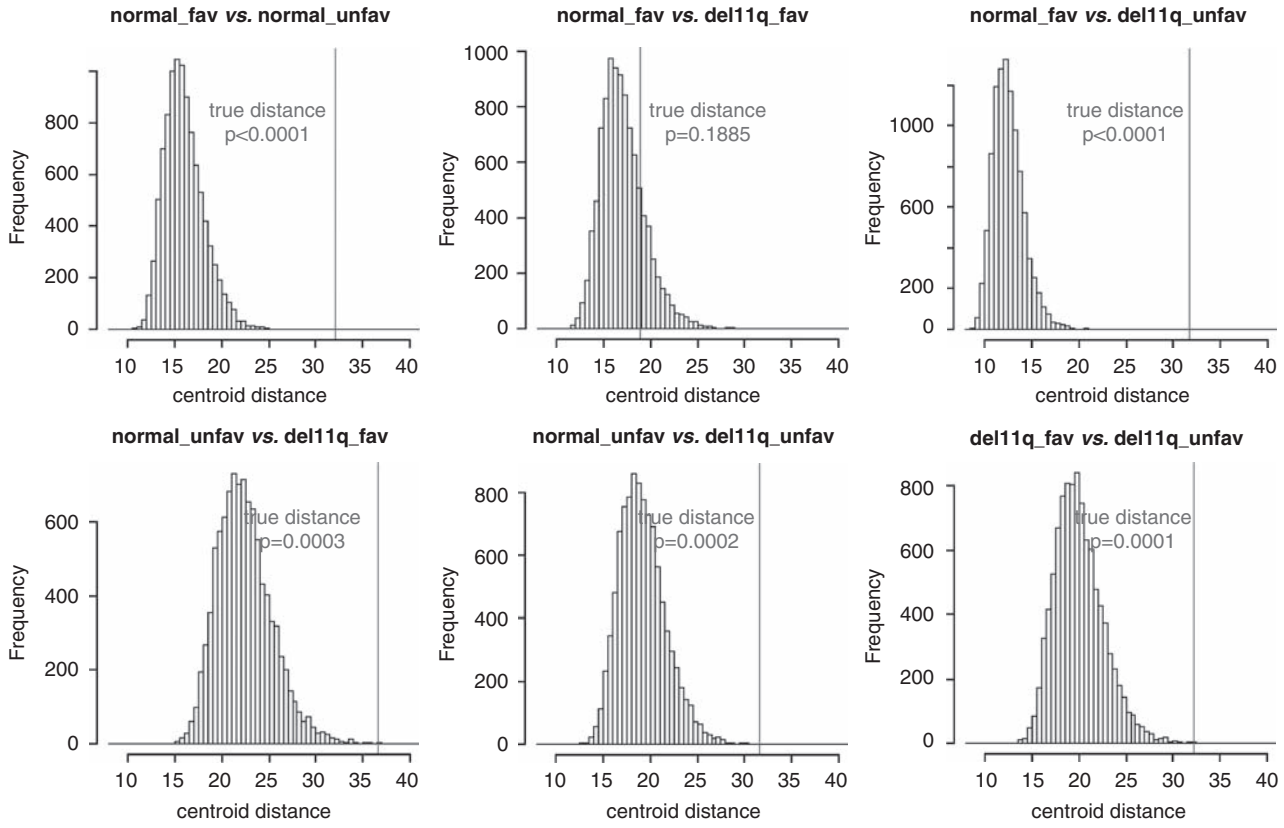


Figure 3 Pairwise analysis of centroid distances between favourable and unfavourable tumours with and without 11q aberration. Indicated are the centroid distances of the permutation analyses and the true centroid distance (red) for each comparison. A full colour version of this figure is available at the *Oncogene* journal online.

Table 1 Differentially expressed genes as determined by pairwise comparison of neuroblastoma subgroups using significance analysis of microarrays (SAM)

	Normal_fav vs normal_unfav	*Normal_fav vs del11q_fav*	*Normal_unfav vs del11q_unfav*	*del11q_fav vs del11q_unfav*
SAM	1187 genes	2 genes	64 genes	282 genes
Genes on 11q	38 (3%)	0	27 (42%)	10 (4%)

in the same direction (Supplementary Figure 1b and Supplementary Table 4B). Similarly, a large number of commonly regulated genes (51/64 genes, 80%) was found when the overlap of the comparisons normal_fav versus del11q_unfav and normal_unfav versus del11q_unfav was determined (Supplementary Figure 1c and Supplementary Table 4C). Notably, genes located on 11q were consistently downregulated in del11q_unfav in comparison with both normal_unfav and normal_fav tumours, supporting the hypothesis that 11q loss does affect the expression of corresponding genes in unfavourable neuroblastoma.

Genomic aberrations at 11q comprise large chromosomal regions in both favourable and unfavourable neuroblastoma

To rule out the possibility that our findings were influenced by substantial differences in the size of 11q alterations, we performed array-based comparative genomic hybridization (aCGH) of those tumours with loss of 11q for which DNA was available (del11q_unfav, $n=10$; del11q_fav, $n=6$). Apart from an interstitial 53.3-Mb deletion of one tumour of the del11q_unfav subgroup (NB327), all aberrations represented large terminal deletions. Breakpoints clustered in two regions at 70–72 and 77–84 Mb, as reported previously (Stallings *et al*., 2006; Spitz *et al*., 2006b). In the del11q_unfav subgroup, loss of terminal genomic material ranged from 54.8 to 64.0 Mb (62.2 Mb on average), whereas it ranged from 49.9 to 64.1 Mb in del11q_fav tumours (55.7 Mb on average, Figure 4a and Supplementary Table 5). In one case of the latter subgroup (NB062), only a weak decrease in the signal strength at 11q could be detected by aCGH, which was probably due to a mosaicism of approximately 50% diploid cells without 11q aberrations and 50% triploid cells with one deleted and two intact copies of 11q in this sample. It is to be

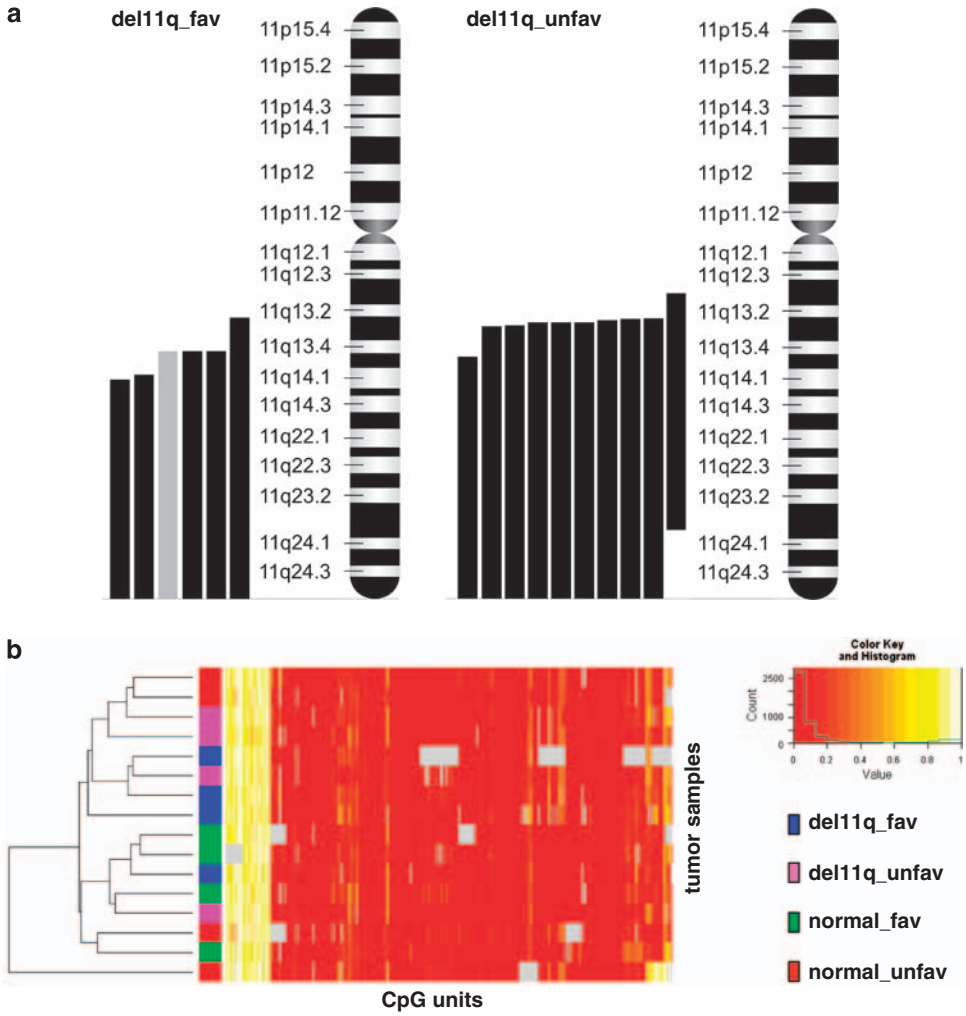


Figure 4 Schematic representation of deleted genomic regions at 11q in the del11q_fav and the del11q_unfav subgroups as determined using array-based comparative genomic hybridization (aCGH) (**a**). Case NB062, in which only a weak decrease in the signal strength at 11q was detected, is indicated by a grey bar. Hierarchical clustering of methylation ratios (**b**). A total of 425 CpG units of 10 genes located on 11q were analysed in 16 tumour samples (four of each subgroup). DNA-methylation values are indicated by colours ranging from dark red (non-methylated) to bright yellow (100% methylated). Poor-quality data are indicated in grey. A histogram is given in the inset that indicates the frequency of each colour in the hierarchical cluster analysis of this figure. For hierarchical clustering, Euclidian distance and complete linkage were used.

noted that although deletions at 11q in the subgroup of del11q_unfav tumours were slightly larger on average than those of the del11q_fav subgroup, the largest terminal deletion of the whole cohort was observed in a del11q_fav tumour. These results argue against a significant influence of the size of the deleted region in determining the clinical phenotype of the tumour.

In addition to 11q aberrations, almost all tumours of both the del11q_fav and the del11q_unfav subgroup were characterized by large 17q gains (Supplementary Table 5). These genomic alterations ranged from 41.2 to 56.1 Mb in the favourable subgroup (49.9 Mb on average), and from 39.6 to 55 Mb in the unfavourable subgroup (46.7 Mb on average). At a lower frequency, a gain of 17q was also found in tumours of both subgroups without 11q CNA (3/19 of normal_fav tumours and 2/5 of normal_unfav tumours). These data may indicate that a gain of 17q does not have an effect in addition to 11q loss on the determination of the clinical phenotype or on the gene expression profile of the tumour.

Promoter regions of differentially expressed genes located on 11q do not differ in their methylation status in clinico-genetic subgroups

To determine whether epigenetic regulation by promoter hypermethylation might contribute to the specific gene expression profile of unfavourable neuroblastomas with loss of 11q, we finally analysed the methylation status of promoter regions from genes that were found to be downregulated in this subgroup. For this purpose, we selected 10 genes located at 11q, six of which had diminished expression levels in del11q_unfav in comparison with normal_unfav tumours, and four of which had lower mRNA levels in del11q_unfav in comparison with

del11q_fav tumours (Supplementary Table 6). In all, 22 genomic regions, ranging from 300 to 500 bp (median length, 432 bp), were analysed in four tumours of each subgroup (Supplementary Table 1), and the methylation status of 425 CpG units was determined. In total, a strikingly homogeneous methylation pattern was observed in tumours of all subgroups, as indicated by an unsupervised hierarchical cluster analysis (Figure 4b). The largest methylation differences were found in a region approximately 600 bp upstream of the *DLG2* transcription start site, in which methylation ratios of four CpG units varied between 20 and 60%. In the remaining regions, methylation ratios did not differ considerably apart from outlying values of single CpG units. Although these results do not exclude the possibility that methylation of specific genes at 11q may contribute to the determination of the neuroblastoma phenotype, they indicate that the downregulation of multiple genes at 11q in del11q_unfav neuroblastoma does not result from global differences in CpG methylation patterns.

Discussion

Cancer genomes are characterized by numerous alterations including low-level copy number gains and losses of large chromosomal regions. Recurrent genomic CNA are thought to define distinct tumour subsets and contribute to tumourigenesis by affecting the expression of cancer genes. To identify such genes, transcriptomic patterns of tumours with and without the respective alterations have been compared in several studies. In most of these, the CNA under investigation were found to correlate with de-regulated gene expression of both the corresponding genes and global gene expression profiles (Pollack *et al*., 2002; Nigro *et al*., 2005; Chen *et al*., 2007; Yoshimoto *et al*., 2007; Gallegos Ruiz *et al*., 2008; Potter *et al*., 2008). However, other studies did not observe such correlations between CNA and gene expression patterns (Platzer *et al*., 2002; Huang *et al*., 2006). These heterogeneous results may be explained not only by the divergent biological behaviour of different malignancies in these studies, but also by the methodical difficulty to integrate CNA spanning hundreds of genes and the corresponding gene expression patterns in small tumour cohorts.

In neuroblastoma, loss of large genomic regions at 11q occurs in roughly 30% of the tumours and is associated with an unfavourable clinical outcome. To evaluate the hypothesis that tumours with 11q CNA form a distinct clinico-genetic subgroup and to examine the effect of this genomic alteration on the transcriptome *in vivo*, we integrated clinical information, cytogenetic characteristics, gene expression profiles and promoter methylation data of neuroblastomas with and without loss of 11q. Classification of 61 neuroblastomas with 11q loss by our previously defined gene expression-based classifier (Oberthuer *et al*., 2006) reliably distinguished tumours from patients with favourable and unfavourable outcome, suggesting that tumours with 11q CNA do not represent a homogeneous clinico-genetic subgroup of neuroblastoma. We next performed a more global assessment of the interactions of 11q CNA, gene expression patterns and the clinical phenotype by analysing four clinico-genetic subgroups defined according to the presence or absence of 11q CNA and the clinical outcome of the patients. Using both unsupervised and supervised analyses, it was shown that neuroblastoma with loss of 11q comprises two distinct subgroups that differ both in their clinical phenotype and their gene expression profile. Surprisingly, the gene expression profiles of tumours with 11q CNA from patients with a favourable disease did not deviate from benign tumours without 11q CNA. In contrast, unfavourable tumours with loss of 11q seem to constitute a specific subgroup and show downregulation of genes located at 11q in comparison with neuroblastoma without 11q CNA. Promoter methylation patterns of 10 genes showing downregulated expression in unfavourable tumours with 11q loss did not differ significantly among the four clinico-genetic subgroups. Together, these findings strongly suggest that 11q loss affects the expression levels of multiple corresponding genes *in vivo* in adverse neuroblastoma. In favourable neuroblastoma, however, the molecular effects of 11q loss are obviously compensated by yet unknown mechanisms.

As imbalances in 11q (that is, at least two intact copies of chromosome 11 but additional deleted copies) were present only in the subgroup of favourable neuroblastoma, one might suggest that these tumours may represent a more benign clinico-genetic subgroup in comparison with tumours with 11q deletions, and that the observed differences might be attributed to a less pronounced gene dosage effect. It has, however, been shown that patients with imbalances and patients with deletions of 11q do not differ in their clinical courses (Spitz *et al*., 2006a). In addition, 3/7 tumours within the favourable subgroup with 11q loss showed hemizygous deletions at this genomic site, which clearly shows that our observations do not merely reflect different transcriptomic effects of either imbalances or deletions.

Current models of neuroblastoma pathogenesis propose two biologically different subtypes of the disease, based on the presence of recurrent CNA (Brodeur, 2003): type 1 describes neuroblastoma with the capacity to spontaneously regress or to differentiate into benign ganglioneuroma and is characterized by numerical chromosomal alterations, whereas type 2 tumours follow an aggressive clinical course, and are further separated into two subtypes, 2A and 2B. Whereas type 2B is defined by an amplification of the oncogene *MYCN*, type 2A is mainly characterized by loss of chromosome 11q. According to this model, the phenotype of type 2A is conferred by 11q CNA, which has been suggested to downregulate the expression levels of one or multiple tumour suppressor genes in this region by haploinsufficiency or complex multigene repression mechanisms (Bilke *et al*., 2005; Maris *et al*., 2007; Stallings, 2007). This hypothesis has been substantiated

by several studies that concurrently reported that loss of 11q affects the expression levels of multiple corresponding genes (McArdle *et al*., 2004; Wang *et al*., 2006; Lastowska *et al*., 2007; Mosse *et al*., 2007). However, the sample numbers of these analyses were medium sized at most ($n=22-n=101$), and in some of these, the data sets were dichotomized according to the cytogenetic aberrations under investigation, which renders the identification of subsets within a CNA-defined subgroup impossible. Notably, McArdle *et al*. (McArdle *et al*., 2004) observed that two low-stage tumours with loss of 11q grouped together with hyperdiploid low-stage tumours in a hierarchical cluster analysis. The researchers concluded from their data that loss of 11q (and gain of 17q) may be insufficient events to lead to a global gene expression profile indicative of aggressive stage 4 tumours, which is well in line with the results of our study.

In this study, we provide strong evidence that loss of 11q does not determine the phenotype of neuroblastoma by its own. It seems more likely that the decision between favourable and unfavourable neuroblastoma is made by yet undefined transforming events (for example, activating mutations in the tyrosine kinase *ALK* in a small subgroup; Mosse *et al*., 2008) previously to the acquirement of 11q loss and possibly also previously to the acquirement of other chromosomal alterations (Figure 5). The biology of each subtype seems to be reflected accurately by the gene expression profile of the tumour cells. After the tumour phenotype has been specified, numerical aberrations occur primarily in the favourable subset, which might result from a mitotic defect in these tumours. Structural alterations (for example, 11q loss) may rarely arise in this subtype but do not contribute to the development of aggressive tumours. In contrast, a defect in genomic stability may represent an inherent property of unfavourable neuroblastoma, leading to multiple structural aberrations at fragile genomic sites. Selective pressure on the cancer cells may then promote the development of a tumour subset with 11q loss, whereas the presence of *MYCN* amplification seems to be almost incompatible with 11q CNA. The effect of 11q loss can be recognized in the transcriptome of unfavourable neuroblastoma and may contribute to specific properties of this subtype. In favourable tumours, however, the effects of diminished gene dosages caused by 11q CNA are obviously balanced by molecular mechanisms yet to be determined.

In line with the proposed model, we observed unfavourable tumours without any detectable structural genomic alterations (Supplementary Table 5), and a substantial concordance of differentially expressed genes between unfavourable and favourable tumours, regardless of the presence of CNA at 11q and other sites (Supplementary Figure 1a) (Westermann *et al*., 2008). These findings corroborate the notion of a common mechanism of malignant transformation in aggressive neuroblastoma. Alternatively, it is possible that loss of 11q represents a first hit in neuroblastoma tumourigenesis and that favourable tumours with 11q CNA constitute an intermediate stage of aggressive neuroblastoma development before a second transforming event. This notion might be supported by the fact that

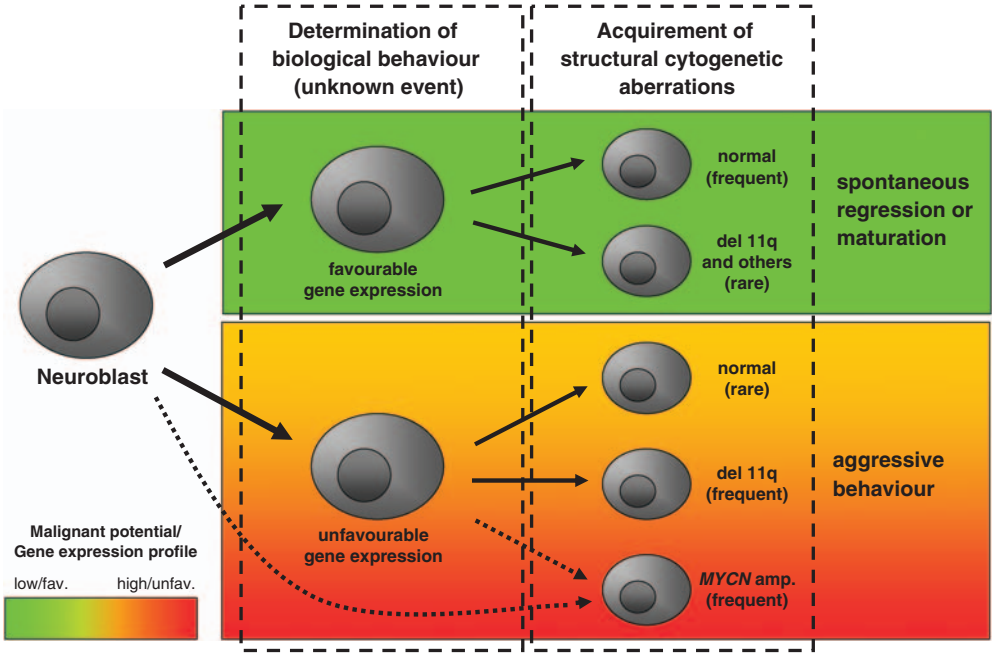


Figure 5 Proposed model of neuroblastoma tumourigenesis modified after Brodeur (2003). The biological and clinical phenotype of neuroblastoma is determined at an early stage of tumour development and is reflected by distinct gene expression patterns. Afterwards, structural genomic alterations occur preferentially in the unfavourable subgroup and may contribute to specific properties of the tumour cells. However, structural alterations may also occasionally occur in the favourable subgroup, in which they are silent events. As an exception, amplification of *MYCN* may represent an early event of tumourigenesis.

4/7 tumours of the del11q_fav subgroup were detected in the neuroblastoma mass screening program. However, as favourable tumours are known to rarely develop into unfavourable neuroblastoma (Brodeur, 2003), this hypothesis seems to be rather unlikely.

It is important to emphasize that this study did not aim at evaluating the prognostic significance but the biological relevance of 11q loss for the pathogenesis of neuroblastoma. It has been shown in large patient cohorts that loss of 11q is statistically associated with adverse clinical courses (Attiyeh *et al*., 2005; Spitz *et al*., 2006a). Nevertheless, our results provide further evidence that gene expression patterns are strongly correlated with the biological and clinical behaviour in neuroblastoma (Ohira *et al*., 2005; Asgharzadeh *et al*., 2006; Oberthuer *et al*., 2006, 2008). In general, our study shows that the presence of prognostically meaningful genomic aberrations does not necessarily form a homogeneous clinico-genetic subgroup of cancer, and exemplifies that integration of detailed clinical information of the patients with genetic, transcriptomic and epigenetic characteristics of the tumours can contribute to the establishment of models for tumour development.

Materials and methods

Characteristics of patients and tumours

In this study, analyses were performed in two patient cohorts that were defined by different criteria: (1) For analysis of the power of the PAM classifier to predict the outcome of patients with tumours showing 11q CNA, all available tumours with loss of 11q, as determined using FISH, were analysed (total, $n=61$; deletion, $n=44$; imbalance, $n=17$; stage 1, $n=8$; stage 2, $n=4$; stage 3, $n=5$; stage 4, $n=40$; stage 4S, $n=4$; *MYCN* amplification, $n=6$; loss of 1p, $n=19$, 1p status not evaluable in 1 case). Patients were enrolled in the German neuroblastoma trials NB90-NB04 with informed consent. Median age at diagnosis was 31 months, and the median follow-up of patients who were alive was 56 months. (2) For the comparison of clinico-genetic subgroups, we selected all available tumours (total, $n=110$; stage 1, $n=50$; stage 2, $n=24$; stage 3, $n=5$; stage 4, $n=14$; stage 4S, $n=17$) from our previously described expression array cohort (Oberthuer *et al*., 2006) that fitted to the following clinical and genomic criteria: (i) Samples from patients with clinical courses defined as either survival without event for at least 2 years without any chemotherapy ($n=88$, referred to as 'favourable' throughout the paper), or as progression/relapse into stage 4 ($n=7$), multiple loco-regional relapses ($n=1$) or death of disease ($n=14$), which were in total referred to as 'unfavourable'. (ii) In addition, the selected tumours were characterized by either the presence ($n=21$) or the absence of loss of 11q ($n=89$) according to FISH analysis. Tumours with *MYCN* amplification or loss of 1p were intentionally excluded from these sets. All patients were enrolled in the German neuroblastoma trials NB90-NB97 with informed consent. Median age at diagnosis was 14 months, and median follow-up of patients who were alive was 89 months. The characteristics of the patients and tumours are summarized in Supplementary Table 1.

Fluorescence in situ *hybridization (FISH) analyses*

CNA of the chromosome arms 1p, 11q and the *MYCN* status were determined by interphase FISH, as described elsewhere (Spitz *et al*., 2003), using the DNA probes D1Z2 (1p36), n-myc (2p24) and MLL (11q23), together with the centromeric probes D1Z1, D2Z and D11Z1, respectively. Cell nuclei were counterstained using 4,6-diamidino-2-phenylindole. According to the ENQUA guidelines (Ambros and Ambros, 2001), chromosomal aberrations were defined as deletion by monosomy of the specific region, imbalance by at least two intact copies of the chromosome and additional copies with deletions in the specific region and *MYCN* amplification by at least a fivefold increase in *MYCN* signal numbers in relation to the number of chromosome 2.

Gene expression profiling analyses

Gene expression profiling experiments were carried out using a customized neuroblastoma-related oligonucleotide microarray (Agilent Technologies, Santa Clara, CA, USA) that comprised 10 163 probes covering 8155 Unigene clusters. A total of 486 of the probes on the microarray refer to genes annotated on chromosome 11q. Expression profiles from each neuroblastoma sample were generated as dye-flipped duplicates in dual-colour experiments as described elsewhere (Oberthuer *et al*., 2006). Data from dye-flipped chip pairs were averaged to yield one intensity value for every gene probe of each patient after quality control of raw microarray data and normalization of expression profiles had been performed. All microarray data are available at the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress accession: E-TABM-38).

Array-based comparative genomic hybridization (aCGH)

High-resolution oligonucleotide aCGH was performed using either 44 or 105 K microarrays (43 000 and 99 000 human sequence probes, respectively) as described elsewhere (Spitz *et al*., 2006b). In brief, 2.5–5 μg genomic tumour and reference DNA were labelled and processed according to the manufacturer protocol for each hybridization (Agilent Technologies) and scanned. Images were extracted using Feature Extraction 9.5 software and visualized using CGH-Analytics 3.5 software (Agilent Technologies). The boundaries of chromosome gains and losses were delineated using the ADM-2 algorithm of the CGH-Analytics software.

Analysis of the methylation status of promoter regions

Analysis of the methylation pattern of DNA promoter regions was performed by Sequenom Inc. (Hamburg, Germany) as described elsewhere (Ehrich *et al*., 2005). In brief, 1 μg genomic DNA of each sample was treated with sodium bisulphite using EZ-96 DNA methylation kit according to the alternative conversion protocol of the manufacturer (Zymo Research, Orange, CA, USA). Genomic regions of interest were amplified by PCR using reverse primers that incorporate the T7 promoter sequence for *in vitro* transcription. Oligonucleotides used as primers were designed by using Methprimer (www.urogene.org/methprimer/, Supplementary Table 2). Quantitative methylation analysis was then performed using Sequenom's MassARRAY platform, which uses MALDI-TOF mass spectrometry in combination with RNA base-specific cleavage (MassCLEAVE). Mass spectra were acquired by using a MassARRAY Compact MALDI-TOF (Sequenom) and spectra's methylation ratios were generated using the Epityper software 1.0 (Sequenom).

Bioinformatics and statistical analyses

For survival analysis, Kaplan–Meier estimates were calculated and compared by log rank-test. Death resulting from therapy complications was censored for event-free survival and overall

survival analysis. Progression, relapse and death from disease were considered as events. PCA and hierarchical clustering of expression data were performed using the Rosetta Resolver Software (Version 7.2; Rosetta Inpharmatics LCC, Seattle, WA, USA). Pairwise comparison of centroid distances in global gene expression patterns was performed in three steps (Classen *et al*., 2007). First, the group mean expression values were calculated for each gene. Next, the distances between the mean values were quantified for each gene in the two groups under investigation using the Euclidean distance. The *centroid distance* between two groups is defined as the sum of distances for all genes. Finally, the significance of the calculated centroid distance of all group-pairs was analysed. The group labels were randomly permuted 10 000 times followed by a recalculation of centroid distances. The *P*-value is given by the fraction of iterations that yield centroid distances at least as great as the original centroid distance between two groups. To determine differentially expressed genes, pairwise comparison of subgroups were performed using one-way ANOVA with Benjamini–Hochberg correction for multiple testing (Rosetta Resolver Software) and SAM (Tusher *et al*., 2001) with the *samr* package v1.23 for R open-source software (http://www.R-project.org). The SAM statistics was computed with 1000 permutations and the Δ-value chosen for a 90th percentile false-discovery rate <0.05. Fisher's exact test was applied to analyse for an enrichment of differentially expressed genes on 11q against the null hypothesis that differentially expressed genes are randomly distributed over all chromosomes. The quantitative methylation data were analysed in an unsupervised hierarchical clustering using the Euclidian distance and complete linkage. All calculations were performed using the gplots, Hmisc, lattice and gmodels package in R. Hypothesis-based significance testing was evaded, because the observed differences were negligible.

Conflict of interest

Mathias Ehrich is a shareholder and an employee of SEQUENOM, Inc.

Acknowledgements

We are grateful to Yvonne Kahlert for excellent technical assistance and to Dr Roman Thomas for critical reading of the paper. This work was supported by grants from the Deutsche Krebshilfe (Grant 50-2719), the Bundesministerium für Bildung und Forschung (BMBF) through the National Genome Research Network 2 (NGFN2, Grants 01GS0456 and 01GR0450) and the Competence Network Paediatric Oncology and Hematology (KPOH) as well as the Fördergesellschaft Kinderkrebs-Neuroblastom-Forschung e.V.

References

Alaminos M, Mora J, Cheung NK, Smith A, Qin J, Chen L *et al*. (2003). Genome-wide analysis of gene expression associated with MYCN in human neuroblastoma. *Cancer Res* **63**: 4538–4546.

Ambros PF, Ambros IM. (2001). Pathology and biology guidelines for resectable and unresectable neuroblastic tumors and bone marrow examination guidelines. *Med Pediatr Oncol* **37**: 492–504.

Asgharzadeh S, Pique-Regi R, Sposto R, Wang H, Yang Y, Shimada H *et al*. (2006). Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. *J Natl Cancer Inst* **98**: 1193–1203.

Attiyeh EF, London WB, Mosse YP, Wang Q, Winter C, Khazi D *et al*. (2005). Chromosome 1p and 11q deletions and outcome in neuroblastoma. *N Engl J Med* **353**: 2243–2253.

Berwanger B, Hartmann O, Bergmann E, Bernard S, Nielsen D, Krause M *et al*. (2002). Loss of a FYN-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. *Cancer Cell* **2**: 377–386.

Bilke S, Chen QR, Westerman F, Schwab M, Catchpoole D, Khan J. (2005). Inferring a tumor progression model for neuroblastoma from genomic data. *J Clin Oncol* **23**: 7322–7331.

Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P *et al*. (2001). N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO J* **20**: 1383–1393.

Brodeur GM. (2003). Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer* **3**: 203–216.

Chen W, Salto-Tellez M, Palanisamy N, Ganesan K, Hou Q, Tan LK *et al*. (2007). Targets of genome copy number reduction in primary breast cancers identified by integrative genomics. *Genes Chromosomes Cancer* **46**: 288–301.

Classen S, Zander T, Eggle D, Chemnitz JM, Brors B, Buchmann I *et al*. (2007). Human resting CD4+ T cells are constitutively inhibited by TGF beta under steady-state conditions. *J Immunol* **178**: 6931–6940.

Cohn SL, Pearson AD, London WB, Monclair T, Ambros PF, Brodeur GM *et al*. (2009). The International Neuroblastoma Risk Group (INRG) classification system: an INRG Task Force report. *J Clin Oncol* **27**: 289–297.

Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G *et al*. (2005). Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci USA* **102**: 15785–15790.

Fischer M, Oberthuer A, Brors B, Kahlert Y, Skowron M, Voth H *et al*. (2006). Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome. *Clin Cancer Res* **12**: 5118–5128.

Fischer M, Spitz R, Oberthur A, Westermann F, Berthold F. (2008). Risk estimation of neuroblastoma patients using molecular markers. *Klin Padiatr* **220**: 137–146.

Gallegos Ruiz MI, Floor K, Roepman P, Rodriguez JA, Meijer GA, Mooi WJ *et al*. (2008). Integration of gene dosage and gene expression in non-small cell lung cancer, identification of HSP90 as potential target. *PLoS ONE* **3**: e0001722.

Henrich KO, Fischer M, Mertens D, Benner A, Wiedemeyer R, Brors B *et al*. (2006). Reduced expression of CAMTA1 correlates with adverse outcome in neuroblastoma patients. *Clin Cancer Res* **12**: 131–138.

Huang J, Sheng HH, Shen T, Hu YJ, Xiao HS, Zhang Q *et al*. (2006). Correlation between genomic DNA copy number alterations and transcriptional expression in hepatitis B virus-associated hepatocellular carcinoma. *FEBS Lett* **580**: 3571–3581.

Janoueix-Lerosey I, Schleiermacher G, Michels E, Mosseri V, Ribeiro A, Lequin D *et al*. (2009). Overall genomic pattern is a predictor of outcome in neuroblastoma. *J Clin Oncol* **27**: 1026–1033.

Lastowska M, Viprey V, Santibanez-Koref M, Wappler I, Peters H, Cullinane C *et al*. (2007). Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. *Oncogene* **26**: 7432–7444.

Maris JM, Hogarty MD, Bagatell R, Cohn SL. (2007). Neuroblastoma. *Lancet* **369**: 2106–2120.

McArdle L, McDermott M, Purcell R, Grehan D, O'Meara A, Breatnach F *et al*. (2004). Oligonucleotide microarray analysis of gene expression in neuroblastoma displaying loss of chromosome 11q. *Carcinogenesis* **25**: 1599–1609.

Mosse YP, Diskin SJ, Wasserman N, Rinaldi K, Attiyeh EF, Cole K *et al*. (2007). Neuroblastomas have distinct genomic DNA profiles that predict clinical phenotype and regional gene expression. *Genes Chromosomes Cancer* **46**: 936–949.

Mosse YP, Laudenslager M, Longo L, Cole KA, Wood A, Attiyeh EF *et al*. (2008). Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455**: 930–935.

Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, Griffin C *et al*. (2005). Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res* **65**: 1678–1686.

Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R *et al*. (2006). Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol* **24**: 5070–5078.

Oberthuer A, Kaderali L, Kahlert Y, Hero B, Westermann F, Berthold F *et al*. (2008). Subclassification and individual survival time prediction from gene expression data of neuroblastoma patients by using CASPAR. *Clin Cancer Res* **14**: 6590–6601.

Ohira M, Oba S, Nakamura Y, Isogai E, Kaneko S, Nakagawa A *et al*. (2005). Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell* **7**: 337–350.

Platzer P, Upender MB, Wilson K, Willis J, Lutterbaugh J, Nosrati A *et al*. (2002). Silence of chromosomal amplifications in colon cancer. *Cancer Res* **62**: 1134–1138.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE *et al*. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* **99**: 12963–12968.

Potter N, Karakoula A, Phipps KP, Harkness W, Hayward R, Thompson DN *et al*. (2008). Genomic deletions correlate with underexpression of novel candidate genes at six loci in pediatric pilocytic astrocytoma. *Neoplasia* **10**: 757–772.

Savelyeva L, Schwab M. (2001). Amplification of oncogenes revisited: from expression profiling to clinical application. *Cancer Lett* **167**: 115–123.

Spitz R, Hero B, Ernestus K, Berthold F. (2003). Deletions in chromosome arms 3p and 11q are new prognostic markers in localized and 4s neuroblastoma. *Clin Cancer Res* **9**: 52–58.

Spitz R, Hero B, Simon T, Berthold F. (2006a). Loss in chromosome 11q identifies tumors with increased risk for metastatic relapses in localized and 4S neuroblastoma. *Clin Cancer Res* **12**: 3368–3373.

Spitz R, Oberthuer A, Zapatka M, Brors B, Hero B, Ernestus K *et al*. (2006b). Oligonucleotide array-based comparative genomic hybridization (aCGH) of 90 neuroblastomas reveals aberration patterns closely associated with relapse pattern and outcome. *Genes Chromosomes Cancer* **45**: 1130–1142.

Stallings RL. (2007). Origin and functional significance of large-scale chromosomal imbalances in neuroblastoma. *Cytogenet Genome Res* **118**: 110–115.

Stallings RL, Nair P, Maris JM, Catchpoole D, McDermott M, O'Meara A *et al*. (2006). High-resolution analysis of chromosomal breakpoints and genomic instability identifies PTPRD as a candidate tumor suppressor gene in neuroblastoma. *Cancer Res* **66**: 3673–3680.

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

Wang Q, Diskin S, Rappaport E, Attiyeh E, Mosse Y, Shue D *et al*. (2006). Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res* **66**: 6050–6062.

Westermann F, Muth D, Benner A, Bauer T, Henrich KO, Oberthuer A *et al*. (2008). Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol* **9**: R150.

Yoshimoto T, Matsuura K, Karnan S, Tagawa H, Nakada C, Tanigawa M *et al*. (2007). High-resolution analysis of DNA copy number alterations and gene expression in renal clear cell carcinoma. *J Pathol* **213**: 392–401.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc)

Cancer Research

CAMTA1, a 1p36 Tumor Suppressor Candidate, Inhibits Growth and Activates Differentiation Programs in Neuroblastoma Cells

Kai-Oliver Henrich[1], Tobias Bauer[4], Johannes Schulte[6], Volker Ehemann[5], Hedwig Deubzer[2], Sina Gogolin[1], Daniel Muth[1], Matthias Fischer[7], Axel Benner[3], Rainer König[4], Manfred Schwab[1], and Frank Westermann[1]

Abstract

A distal portion of human chromosome 1p is often deleted in neuroblastomas and other cancers and it is generally assumed that this region harbors one or more tumor suppressor genes. In neuroblastoma, a 261 kb region at 1p36.3 that encompasses the smallest region of consistent deletion pinpoints the locus for *calmodulin binding transcription activator 1* (*CAMTA1*). Low *CAMTA1* expression is an independent predictor of poor outcome in multivariate survival analysis, but its potential functionality in neuroblastoma has not been explored. In this study, we used inducible cell models to analyze the impact of CAMTA1 on neuroblastoma biology. In neuroblastoma cells that expressed little endogenous *CAMTA1*, its ectopic expression slowed cell proliferation, increasing the relative proportion of cells in G_1/G_0 phases of the cell cycle, inhibited anchorage-independent colony formation, and suppressed the growth of tumor xenografts. CAMTA1 also induced neurite-like processes and markers of neuronal differentiation in neuroblastoma cells. Further, retinoic acid and other differentiation-inducing stimuli upregulated *CAMTA1* expression in neuroblastoma cells. Transciptome analysis revealed 683 genes regulated on CAMTA1 induction and gene ontology analysis identified genes consistent with CAMTA1-induced phenotypes, with a significant enrichment for genes involved in neuronal function and differentiation. Our findings define properties of *CAMTA1* in growth suppression and neuronal differentiation that support its assignment as a 1p36 tumor suppressor gene in neuroblastoma. *Cancer Res; 71(8); 3142–51. ©2011 AACR.*

Introduction

Neuroblastoma is a childhood tumor derived from precursor cells of the sympathetic nervous system. The clinical and biological behavior of this tumor is remarkably heterogeneous, encompassing fatal tumor progression, as well as spontaneous regression and differentiation into mature ganglioneuroma. Deletion within distal 1p characterizes about 30% of neuroblastomas and also frequently occurs in a broad range of other human malignancies including colorectal cancer, glioma, breast cancer, and melanoma. Further, 1p36 deletion is an independent predictor of neuroblastoma progression (1). Thus, it is widely assumed that distal 1p harbors genetic information mediating tumor suppression. The combination of recent fine mapping studies (2, 3) defines a 1p36.3 smallest region of consistent deletion shared by virtually all 1p-deleted neuroblastomas that spans 261 kb between *D1S2731* and *D1S214*, encompassing the *CAMTA1* locus (4). *CAMTA1* encodes a calmodulin-binding transcription activator (5, 6) that is predominantly expressed in neuronal tissues (7). There is no evidence for *CAMTA1* mutations in neuroblastoma (8), however, low *CAMTA1* expression is significantly associated with markers of unfavorable tumor biology and poor outcome. Intriguingly, the prognostic value of *CAMTA1* expression is independent of established risk markers, including 1p deletion, in multivariate survival analysis (4). Additional evidence supporting *CAMTA1* involvement in tumor development comes from glioma and colon cancer. *CAMTA1* is homozygously deleted in a subset of gliomas (9) and is the only gene mapping to the 1p36 smallest region of overlapping heterozygous deletion in this entity (10). In colorectal cancer, genome-wide copy number analysis revealed that loss of a 2 Mb region encompassing *CAMTA1* has the strongest impact on survival among all identified genomic alterations (11). Further, as in neuroblastoma, low expression of *CAMTA1* is an independent predictor of poor outcome in colorectal cancer (11).

Authors' Affiliations: [1]Division of Tumor Genetics B030, [2]Clinical Cooperation Unit Pediatric Oncology G340, and [3]Division of Biostatistics, German Cancer Research Center; [4]Institute of Pharmacy and Molecular Biotechnology, Bioquant; [5]Department of Pathology, University of Heidelberg, Heidelberg, Germany; [6]Department of Pediatric Oncology and Hematology, University Children's Hospital, Essen, Germany; and [7]Department of Pediatric Oncology, University Children's Hospital, Cologne, Germany

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Kai-Oliver Henrich, Division of Tumor Genetics B030, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. Phone: 49-6221-423220; Fax: 49-6221-423277; E-mail: k.henrich@dkfz.de

doi: 10.1158/0008-5472.CAN-10-3014

In this study, we explore the effect of CAMTA1 on neuroblastoma biology using inducible cell models. Our data imply that *CAMTA1* is a 1p36 tumor suppressor candidate that inhibits features of malignant cells and is involved in neuroblastoma cell differentiation.

Materials and Methods

Cell culture

Culture of the neuroblastoma cell lines SH-EP, IMR5-75, and Be(2)-C, was described previously (12). All lines were kindly provided by Dr. Larissa Savelyeva (German Cancer Research Center) and authenticated by multiplex-FISH karyotyping at the start of the project. Cells were tested for mycoplasma, viral, and foreign cell contamination using the Multiplex cell Contamination Testing (McCT) Service (13). Drugs were added at the following concentrations: all-*trans* retinoic acid (Sigma), 10 μmol/L (in ethanol, end concentration did not exceed 0.1%); valproic acid (Sigma), 1 mmol/L (in Dulbecco's PBS); *Helminthosporium carbonum*-toxin (HC-toxin; Sigma, Lot #054K4121), 15 nmol/L (in methanol, end concentration did not exceed 0.02%).

Polyclonal antibody production and Western blotting

A custom polyclonal CAMTA1 antibody was raised in rabbits against the epitope peptide NH^2-CHRLYKRSER-IEKGQGT-COOH, representing the COOH-terminal CAMTA1 region (Pineda Antikoerper-Service). Final bleeds were affinity purified, and antibody specificity was confirmed via Western blotting detection of inducible ectopically expressed CAMTA1 and specific epitope blocking. Protein expression was assessed in cell lysates as previously described (12). Protein loading was controlled by detecting β-actin (clone AC-15, Sigma-Aldrich).

Inducible stable cell lines

Stable neuroblastoma cell lines overexpressing *CAMTA1* under control of the Tet-repressor were generated using the T-REX Tetracycline-Regulated Mammalian Expression System and Gateway Technology (Invitrogen). To generate a Tet/Doxy-responsive *CAMTA1*-expression plasmid, we followed the One-Tube Protocol for Cloning attB-PCR Products into Destination Vectors from the Gateway Technology with Clonase II manual. The full-length coding sequence of *CAMTA1* was *att*B-PCR-amplified from the hg01719s1/KIAA0833 cDNA clone (7), and cloned into pT-REx-DEST30 (Invitrogen). The resulting construct was transfected into SH-EP and IMR5-75 cells stably expressing the tetracycline repressor protein (pcDNA/6TR; Invitrogen). Individual clones were selected, expanded and assayed for CAMTA1 expression on tetracycline treatment (1 μg/mL) using Western blotting. Control clones allowing inducible LacZ expression were generated in parallel (vector pT-REx/GW-30/LacZ; Invitrogen). Blasticidin (7.5 μg/mL) and geneticin/G418 sulfate (200 μg/mL) were used for clone selection.

Cell cycle, growth, and colony formation assays

Cell cycle distribution was assessed by fluorescence-activated cell sorting (FACS) as previously described (14). For growth assays, cells were seeded in triplicate onto 6-well plates (1,500 per well), and growth rates were determined by Alamar Blue assay (AbD Serotec) on days 0, 2, 4, and 6 according to the manufacturer's instructions. Cells were formaldehyde-fixed and stained with 10% Giemsa solution to visualize colonies 2 weeks after seeding.

For soft agar assays, 6-well plates were precoated with 0.7% agarose in full medium (RPMI-1640 supplemented with 10% FCS), and 4,000 cells were seeded into 0.35% agarose in full medium per well in triplicate. Cells were fed weekly and stained with crystal violet 4 weeks after seeding.

Growth of xenograft tumors in nude mice

IMR5-75 cells were cultured to 80% confluency, harvested, and suspended in Matrigel (BD Bioscience). Eight-week-old athymic NCR (nu/nu) mice were inoculated s.c. in the flank with 2×10^7 cells in 200 μL Matrigel (sample size: 8 mice inoculated with IMR5-75-CAMTA1 and 6 mice inoculated with IMR5-75-LacZ control cells). Doxycycline was administered via drinking water (2 mg/mL) and orogastric lavage (2 mg/mouse) when all tumors were progressive and reached a volume of at least 100 mm^3. Tumor size was measured with a digital calliper to calculate tumor volume. Mice were sacrificed at day 4 after induction.

Quantitative real-time RT-PCR

The quantitative real-time reverse transcriptase PCR (QPCR) protocol and primers for *CAMTA1* and housekeeping genes were described previously (4). QuantiTect Primer Assays (Qiagen) were used for amplification of *tubulin, beta 3* (*TUBB3*, Hs_TUBB3_1_SG); *neurofilament, light polypeptide* (*NEFL*, Hs_NEFL_1_SG); *microtubule-associated protein 2* (*MAP2*, Hs_MAP2_1_SG), *cyclin-dependent kinase inhibitor 1C* ($p57^{Kip2}$) (*CDKN1C*, Hs_CDKN1C_1_SG); *tropomodulin 2 (neuronal)* (*TMOD2*, Hs_TMOD2_1_SG); *sodium channel, voltage gated, type VIII, alpha subunit* (*SCN8A*, HsSCN8A_1_SG); *S100 calcium binding protein B* (*S100B*, Hs_S100B_1_SG); and *stathmin-like 3* (*STMN3*, Hs_ STMN3_1_SG).

Microarray analysis

Total RNA was isolated using Trizol (Invitrogen) from CAMTA1 expressing SH-EP cells at 0, 3, 6, 12, and 24 hours after CAMTA1 induction and at 12 hours from noninduced controls. Two biological replicates were carried out for time-series experiments. RNA was converted to cRNA, labeled, and hybridized to Agilent whole human genome 4×44 K (G4112F) microarrays according to the Two-Color Microarray-Based Gene Expression Analysis protocol (Agilent Technologies). Raw data were background-corrected using the "normexp"-method and quantile-normalized employing the "limma" package included in the Bioconductor release 2.4 (www.bioconductor.org) for R statistical software v2.9.0 (www.r-project.org). Unspecific filtering was applied to the normalized data as implemented in the "genefilter" R package (Bioconductor release 2.4) for each biological replicate separately (15). Probes were selected for which expression values were greater than or equal to the first quartile of the expression range for at least 2 time points (rather than selecting an absolute expression

value as a cutoff) to select genes with quantifiable expression in at least 2 measurements. A relaxed threshold for the interquartile range (IQR) filter was selected (0.3) and applied to exclude genes with low variability. Only probes which showed a positive Pearson correlation between biological replicates ($r > 0.8$) were included in analyses. The remaining intersection of filtered probes for both biological replicates was used in unsupervised hierarchical clustering of Pearson correlation distances (1 − Pearson correlation coefficients) to obtain clusters with common expression profiles. All arrayed probes are henceforth referred to as genes for simplification. Gene Ontology Tree Machine (GOTM) was used for functional annotation of expression data (16). GOTM compares the distribution of GO terms within a gene set (defined here as gene clusters with common expression profiles) to that in a reference gene set (defined here as all genes represented on the G4112 array). To test for a statistically significant enrichment of GO terms within gene sets, a hypergeometric test was used with a significance level of 0.01 (16).

Survival analysis

CAMTA1 expression was derived from expression profiling data from a cohort of 251 neuroblastomas (17). Of 251 tumors, 70 were previously analyzed for *CAMTA1* expression by cDNA microarray or QPCR (4). All patients were enrolled in the German Neuroblastoma Trial and diagnosed between 1989 and 2004 ($n = 68$ stage 1, $n = 46$ stage 2, $n = 39$ stage 3, $n = 67$ stage 4, $n = 31$ stage 4s; $n = 31$ *MYCN* amplified, $n = 220$ *MYCN* nonamplified; $n = 168$ age at diagnosis <1.5 years, $n = 83$ age at diagnosis ≥1.5 years). Criteria for sample selection were availability of sufficient amounts of tumor material, 60% or more tumor content, and RNA integrity number more than 7.5. The composition of the cohort in terms of tumor stage, *MYCN* status and age at diagnosis was in agreement with the composition of an unselected cohort of 940 patients diagnosed between 1995 and 2001 in Germany (data not shown). Univariate survival analysis was done to validate established prognostic variables as described previously (4). Multivariate Cox regression was used to investigate the prognostic power of *CAMTA1* expression adjusting for established prognostic variables as described previously (4). The cutoff value for dichotomization of *CAMTA1* expression was estimated by maximally selected log-rank statistics (18). Parameter estimate shrinkage was applied to correct for potential overestimation of the hazard ratio estimate due to cutoff selection (19). Bootstrap resampling, together with a shrinkage procedure, was used to correct confidence limits and *P* values (20). Event-free survival (EFS) was measured from date of diagnosis until occurrence of disease progression, relapse, or death from disease. EFS times of patients who experienced no events within the follow-up time were censored.

Results

Low *CAMTA1* expression predicts poor neuroblastoma outcome

Low *CAMTA1* expression was previously identified as a predictor of poor outcome (4). To validate the prognostic value of *CAMTA1* in a larger set of patients, *CAMTA1* expression was derived from expression profiling data from a cohort of 251 neuroblastomas (17) and analyzed. Multivariate survival analysis confirmed low *CAMTA1* expression as a predictor of poor outcome, independent of established risk markers, including 1p status, *MYCN* status, tumor stage and age of the patient at diagnosis (Table 1). Even within the cohort of 1p nondeleted tumors, *CAMTA1* expression emerged as an independent prognostic factor (Table 2).

CAMTA1 suppresses growth of neuroblastoma cells *in vitro* and *in vivo*

The effect of CAMTA1 on neuroblastoma cell growth was explored in stable clones allowing tetracycline-inducible CAMTA1 expression in the SH-EP cell line, which has low endogenous *CAMTA1* expression (21) (validated by QPCR, data not shown). Induction of CAMTA1 in SH-EP cells significantly decreased colony formation ability and growth rate

Table 1. Cox proportional hazards regression for event-free survival (251 neuroblastomas)

Factor	Effect	Hazard ratio (95% confidence limits)	*P*
CAMTA1 expression[a]	Low vs. high	5.23 (2.29–10.3)	<0.001
1p deletion	Yes vs. no	1.26 (0.62–2.55)	0.52
Stage	3, 4 vs. 1, 2, 4s	1.16 (0.58–2.32)	0.67
Age	≥1.5 years vs. <1.5 years	1.32 (0.68–2.54)	0.41
MYCN amplification	Yes vs. no	0.85 (0.41–1.76)	0.66

NOTE: *CAMTA1* expression was derived from array expression data from a cohort of 251 neuroblastomas (17). Results specific for oligo probe A_32_P4981 are shown. Two other independent *CAMTA1*-specific probes (A_32_P4985 and A_24_P220921) revealed similar results. Established risk factors included in the model were all associated with decreased event-free survival in univariate survival analysis: 1p deletion (HR 4.05, $P < 0.001$), higher stage (3 and 4; HR 3.36, $P < 0.001$), age ≥ 1.5 years (HR 3.8, $P < 0.001$), and *MYCN* amplification (HR 3.55, $P < 0.001$).

[a]To correct for potential hazard ratio overestimation due to cutoff selection, parameter estimate shrinkage was applied. To correct confidence limits and *P* values, bootstrap resampling, together with a shrinkage procedure, was used.

Table 2. Cox proportional hazards regression for event-free survival in patients without 1p deletion (195 neuroblastomas)

Factor	Effect	Hazard ratio (95% confidence limits)	*P*
CAMTA1 expression[a]	Low vs. high	4.42 (1.64–9.67)	0.002
Stage	3, 4 vs. 1, 2, 4s	1.14 (0.5–2.63)	0.76
Age	≥1.5 years vs. <1.5 years	1.74 (0.76–3.99)	0.19
MYCN amplification	Yes vs. no	2.38 (0.27–20.99)	0.43

NOTES: *CAMTA1* expression was derived from array expression data from 195 neuroblastomas (17). Results specific for oligo probe A_32_P4981 are shown. Two other independent *CAMTA1*-specific probes (A_32_P4985 and A_24_P220921) revealed similar results.
[a]To correct for potential hazard ratio overestimation due to cutoff selection, parameter estimate shrinkage was applied. To correct confidence limits and *P* values, bootstrap resampling, together with a shrinkage procedure, was used.

(Fig. 1A and B). The proportion of cells in the G_1/G_0 phase significantly increased on CAMTA1 induction (OFF: 59.3% ± 1.7%, ON: 69.9% ± 0.4%, mean ± SD, $P < 0.001$ for 3 replicates) with a concomitant decrease in the proportion of cells in the S-phase (OFF: 29.2% ± 1%, ON: 19.5% ± 0.7%, mean ± SD, $P < 0.001$ for 3 replicates; exemplary FACS analysis in Fig. 1C). Induction of LacZ in a negative control clone had no significant effect on cell growth or cell cycle distribution. To analyze the effect of CAMTA1 on anchorage-independent and *in vivo* growth, we used stable clones allowing tetracycline/doxycycline-inducible CAMTA1 expression in IMR5-75, a *MYCN*-amplified neuroblastoma cell line with low endogenous *CAMTA1* levels (assessed by QPCR, data not shown) and colony-forming ability in soft agar. CAMTA1 induction in IMR5-75 significantly inhibited anchorage-independent growth, whereas induction of LacZ in a negative control clone had no significant effect on colony-forming ability in soft agar (Fig. 1D). Following subcutaneous inoculation of athymic nude mice with CAMTA1-inducible IMR5-75 cells, CAMTA1 was induced in established tumors via doxycycline administration. Induction of *CAMTA1* in tumors was validated by QPCR (Fig. 1E). CAMTA1 induction resulted in significantly reduced tumor volume, whereas induction of LacZ in the negative controls had no significant effect (Fig. 1E). Taken together, these data show that higher CAMTA1 expression shifts the cell cycle away from proliferation and suppresses both *in vitro* and *in vivo* growth of neuroblastoma cells.

CAMTA1 induces markers of neuronal differentiation and is upregulated during neuroblastoma cell differentiation

Microscopic inspection of CAMTA1-induced SH-EP cells revealed a higher degree of morphological differentiation, including acquisition of neurite-like processes (Fig. 2A). To investigate whether this morphology is associated with induction of neuron-specific markers, we measured expression of genes encoding the early neuronal marker β3 tubulin (*TUBB3*), and the later neuronal markers microtubule associated protein 2 (*MAP2*) and neurofilament light chain (*NEFL*). CAMTA1 induction in SH-EP cells significantly increased the expression levels of all 3 neuronal markers compared with noninduced controls (Fig. 2B). To test whether *CAMTA1* is upregulated during neuronal differentiation, we assessed *CAMTA1* expression in 4 established neuroblastoma *in vitro* differentiation models that were extensively characterized in previous studies (22, 23): (i) Be(2)-C treated with retinoic acid, (ii) SH-EP treated with valproic acid, (iii) Be(2)-C treated with HC-toxin, and (iv) SH-EP treated with HC-toxin. Morphological differentiation and induction of the neuronal marker *MAP2* were associated with a significant increase of *CAMTA1* expression levels in all tested neuroblastoma differentiation models (Fig. 3). Together, these data support that CAMTA1 regulation is part of the response to differentiation signals and induces genes characteristic of neuronal differentiation.

CAMTA1 induces genetic programs mediating neuronal functions and growth inhibition

We investigated time-resolved genome wide transcription profiles of CAMTA1-induced SH-EP cells to analyze the global molecular changes induced by the transcription factor CAMTA1 and to elucidate the biological basis of the observed CAMTA1-associated phenotype. We identified a total of 683 genes regulated on CAMTA1 induction (Supplementary Table S1). Unsupervised clustering resulted in 5 clusters comprising genes with common dynamic expression patterns (Fig. 4). Of these, 2 clusters (A, 368 genes and B, 133 genes) contained genes that were time-dependently upregulated by CAMTA1 induction, and whose expression was unchanged in the 12 hours noninduced control. Cluster E comprised 88 genes downregulated on CAMTA1 induction (expression unchanged in noninduced control). All CAMTA1-regulated genes from clusters A, B, and E were associated to GO annotations. This categorization of genes into functional classes was used to provide insight into the molecular processes contributing to the CAMTA1-induced phenotype. We also tested whether specific GO terms were enriched among CAMTA1-induced or CAMTA1-repressed genes. A given GO term was considered enriched when the observed number of genes from that category was significantly greater than the number expected by chance ($P < 0.01$). GO terms enriched among CAMTA1-induced genes (Fig. 4, clusters A and B) reflected the CAMTA1-associated differentiation phenotype.

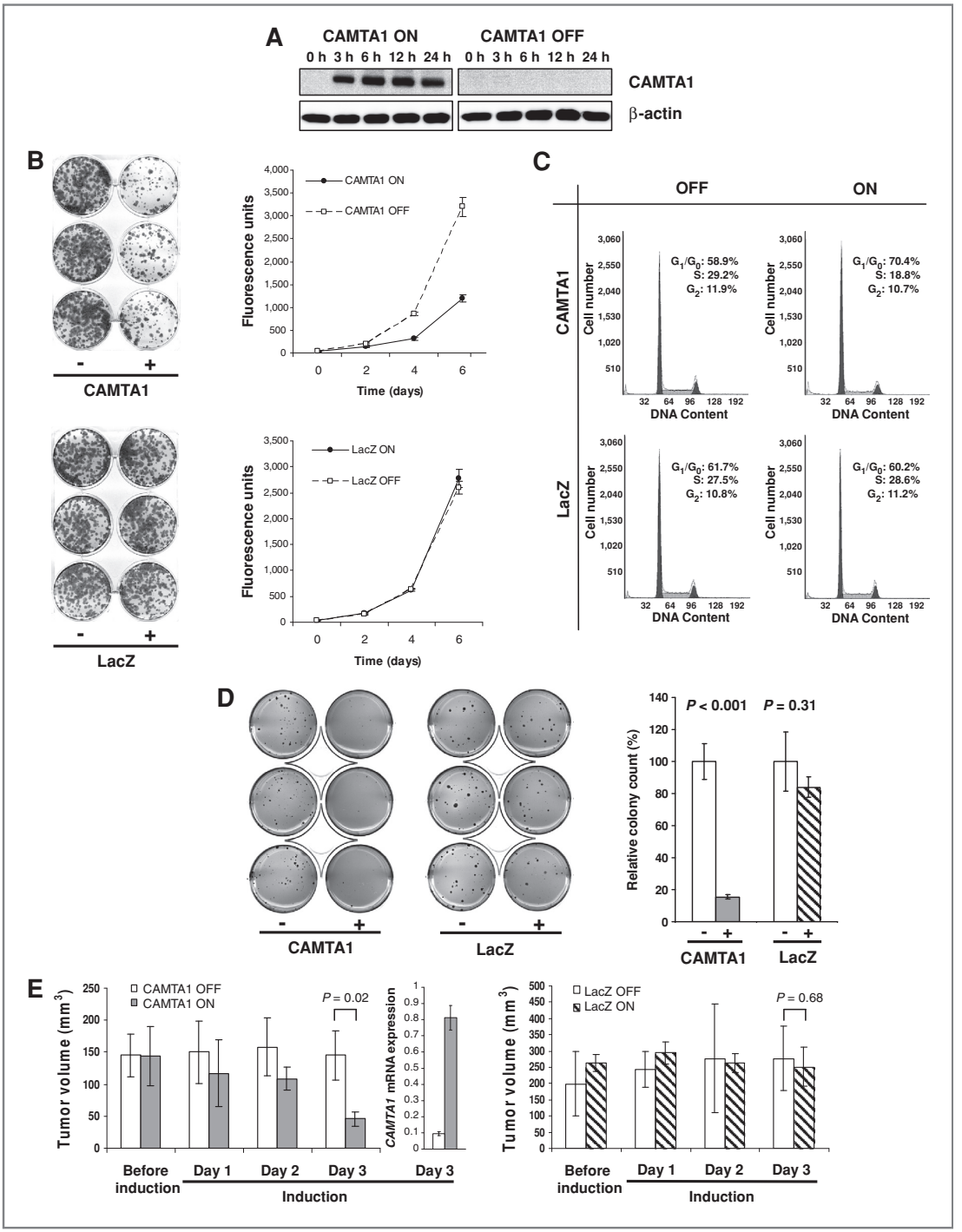


Figure 1. A, tetracycline-inducible CAMTA1 expression and detection via custom-made polyclonal antibody (here in SH-EP). B, CAMTA1 expression suppresses growth of SH-EP cells as determined by colony formation assay and Alamar Blue viability assay (mean ± SD, 3 replicates). C, CAMTA1 expression in SH-EP cells results in an increased proportion of cells in G_1/G_0 phase 48 hours after induction as determined by FACS analysis (1 of 3 replicates is shown). D, CAMTA1 suppresses anchorage-independent growth of IMR5–75 cells in soft agar. E, CAMTA1 induction suppresses growth of subcutaneous IMR5–75 tumors in nude mice. Sample size: 8 mice inoculated with IMR5–75-CAMTA1 and 6 mice inoculated with IMR5–75-LacZ negative control cells. Doxycycline was administered via drinking water (2 mg/mL) and orogastric lavage (2 mg/mouse) when all tumors were progressive and reached a volume of at least 100 mm^3. QPCR was performed on total RNA isolated from 1 CAMTA1 ON and 1 CAMTA1 OFF tumor, respectively, to validate *CAMTA1* induction via doxycycline administration. Cells allowing inducible LacZ expression were used as negative controls.

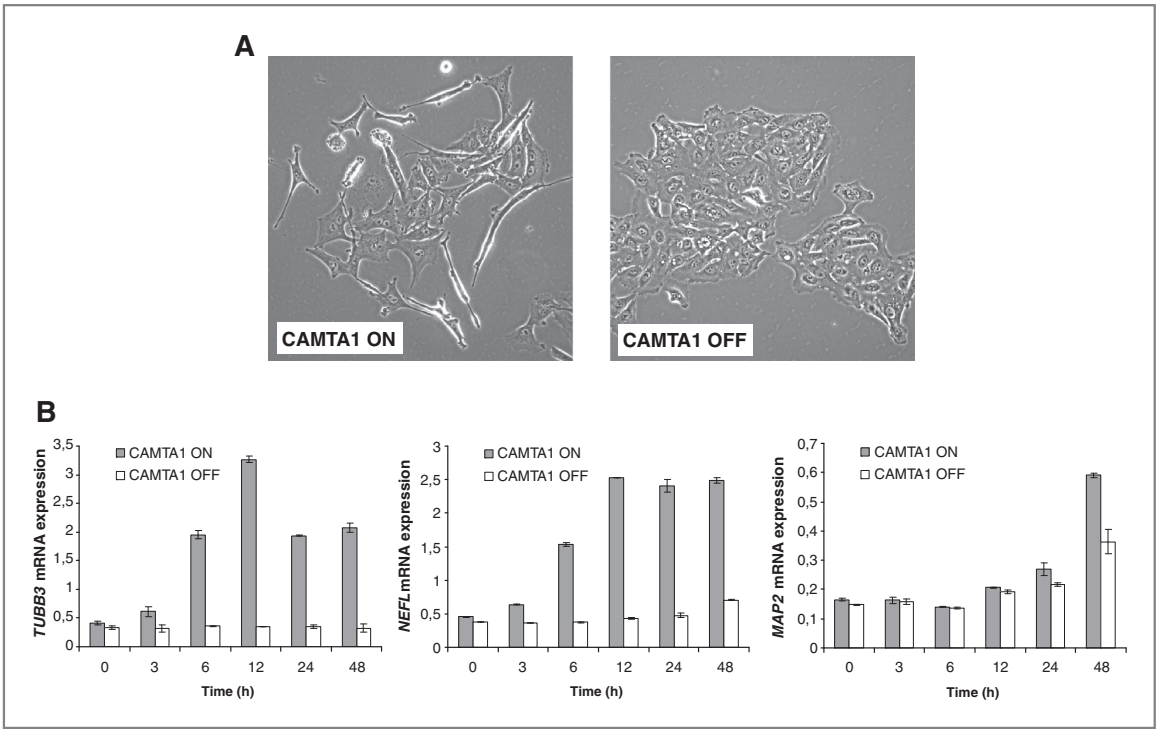


Figure 2. A, morphological changes on CAMTA1 induction in SH-EP cells. B, QPCR reveals induction of neuron-specific marker genes on CAMTA1 induction in SH-EP cells (mean ± SD, 3 replicates).

A large fraction of enriched GO terms related to neuronal differentiation or function (e.g., "nervous system development," "transmission of nerve impulse," "voltage-gated sodium channel activity," and "neurofilament"). Overrepresentation of the GO term "kinase inhibitor activity" among CAMTA1-induced genes was in line with the observed inhibitory effect on cell cycle progression. The enrichment of the GO term "Ca^{2+}/calmodulin-dependend protein kinase complex" was of particular note considering the Ca^{2+}/calmodulin-dependent activity of CAMTAs (24). Among CAMTA1-repressed genes (Fig. 4, cluster E), the majority of enriched GO terms related to cell cycle associated processes. General inhibition of the cell cycle was indicated by overrepresentation of the GO term "regulation of cyclin-dependent kinase activity." Mitotic inhibition was reflected by enrichment of GO terms, such as "mitosis" and "spindle organization and biogenesis." Inhibition of DNA synthesis was indicated by overrepresentation of GO terms, such as "DNA replication initiation." We chose 5 CAMTA1 targets that are representative of the functional classes "neuronal differentiation" and "cell cycle inhibition" for validation by QPCR in an independent SH-EP-*CAMTA1* clone (Fig. 5): *CDKN1C* ($p57^{Kip2}$) that is involved in G_1 phase arrest and is a critical terminal effector of pathways controlling differentiation, *TMOD2*, encoding a neuron-specific member of the tropomodulin family of actin-regulatory proteins, *SCN8A*, encoding a subunit of voltage gated sodium channels, *S100B*, encoding a Ca^{2+} binding protein involved in neurite extension and axonal proliferation and *STMN3*, a paralog of *STMN2 (SCG10)*, which is implicated in terminal differentiation of sympathetic neurons (25). The consistent CAMTA1-dependent regulation of the tested genes in this independent setting supports the robustness of our approach. Overall, time-resolved expression profiling in CAMTA1-induced cell models and functional classification using GO term analysis indicate that CAMTA1 induces differentiation programs and inhibits effectors of cell cycle progression.

Discussion

CAMTA1 is pinpointed by a 1p36.3 smallest region of consistent deletion in neuroblastoma (2–4). In the absence of somatic mutations (8), low *CAMTA1* expression is an independent predictor of poor survival as determined by QPCR and multivariate survival analysis in a cohort of 102 neuroblastomas (4). Here, we further confirmed this result in an extended cohort of 251 neuroblastomas employing oligonucleotide array expression data (17), supporting the robustness of this prognostic marker independent of the technical platform used. *CAMTA1* is also included in most of the recently reported prognostic neuroblastoma expression classifier gene sets, highlighting its predictive power (17, 26–28). The consistently low expression of *CAMTA1* in aggressive

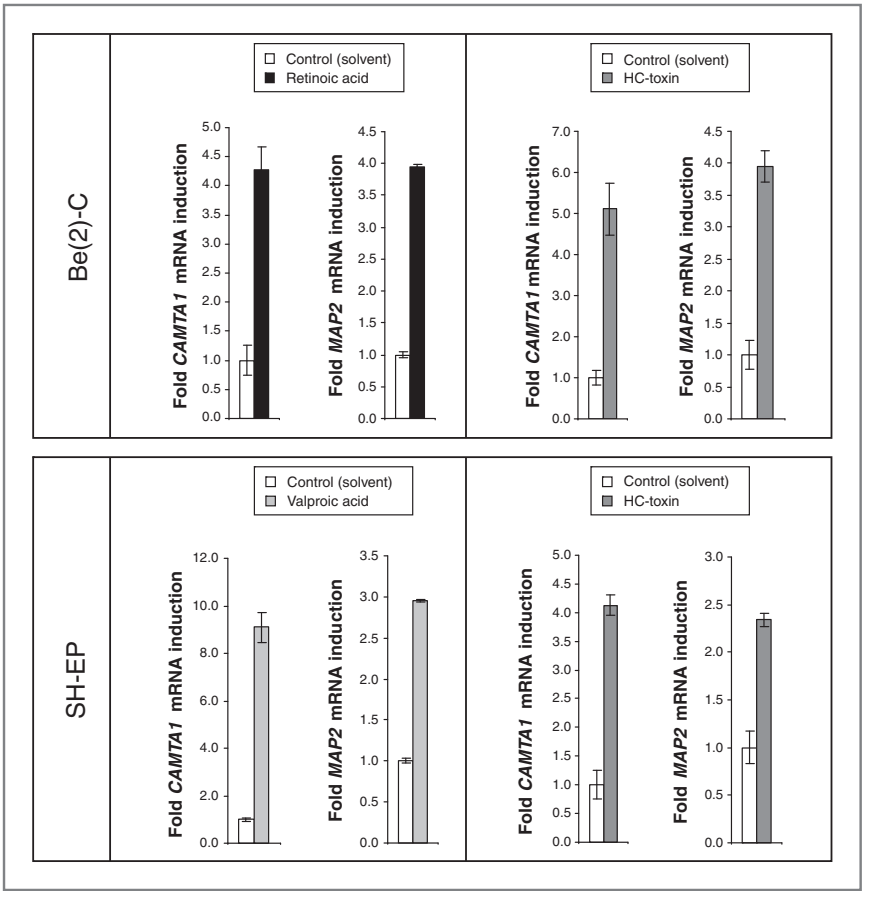


Figure 3. *CAMTA1* is induced in neuroblastoma differentiation models. QPCR reveals induction of *CAMTA1* and the neuronal marker *MAP2* in Be(2)-C cells treated with all-*trans* retinoic acid or *Helminthosporium carbonum* (HC)-toxin, and SH-EP cells treated with valproic acid or HC-toxin (mean ± SD, 3 replicates). Incubation time was 5 days for valproic acid or HC-toxin and 7 days for all-*trans* retinoic acid. Morphological differentiation was confirmed microscopically.

neuroblastomas led us to hypothesize that (i) downregulation of *CAMTA1* mediates a selective advantage of malignant neuroblastoma cells and (ii) reexpression of CAMTA1 in neuroblastoma cells with low endogenous CAMTA1 levels may inhibit features of malignancy. In line with this hypothesis, CAMTA1 induction in SH-EP cells suppressed colony formation and growth rate and induced accumulation of cells in the G_1/G_0 phase of the cell cycle. In IMR5-75 cells, CAMTA1 inhibited anchorage independent growth and *in vivo* growth in nude mice, further strengthening the role of CAMTA1 as a tumor suppressor candidate.

The induction of neurite-like processes and neuronal marker genes (*TUBB3*, *MAP2*, and *NEFL*) on CAMTA1 induction in SH-EP points to a role of CAMTA1 in neuroblastoma cell differentiation. This is further supported by *CAMTA1* upregulation in different *in vitro* models of neuroblastoma differentiation. The histone deacetylase inhibitors used here, valproic acid and HC-toxin, exhibit antineuroblastoma activity (23, 29) and are candidates for future clinical use. Retinoic acid is already implemented in the postconsolidation therapy of stage 4 neuroblastomas (30). Whether upregulation of CAMTA1 contributes to the antineuroblastoma properties of these drugs needs to be addressed in further studies.

CAMTA1 acts as a transcription activator (6). Our results from integrating CAMTA1-induced transcription profiles and corresponding GO annotations are consistent with the idea that CAMTA1 regulates effectors of neuronal function and cell cycle inhibition. High expression of neuron-specific genes is a feature of localized tumors (stages 1 and 2) (31) and, despite poor histological differentiation, disseminated 4s tumors (32). The high expression of *CAMTA1* in stages 1, 2, and 4s tumors (4) may indicate that the genetic programs induced by CAMTA1 *in vitro* contribute to the favorable phenotype of this subgroup *in vivo*. Intracellular Ca^{2+} fulfils a pleiotropic role in both the physiology and differentiation of neuronal cells, and neuritic outgrowth can be induced in neuroblastoma cells by promoting Ca^{2+} influx (33). CAMTA family members respond to Ca^{2+} signaling by binding to calmodulin (24), and the GO term "Ca^{2+}/calmodulin-dependend protein kinase complex" was enriched among CAMTA1-induced genes. This suggests that CAMTA1 acts as both integrator and effector of Ca^{2+} signaling and may mediate Ca^{2+}-dependent processes in neuronal differentiation. GO terms enriched among CAMTA1-repressed genes indicate that processes of mitosis and DNA replication are inhibited by CAMTA1. Together with the previous observation that *CAMTA1* is

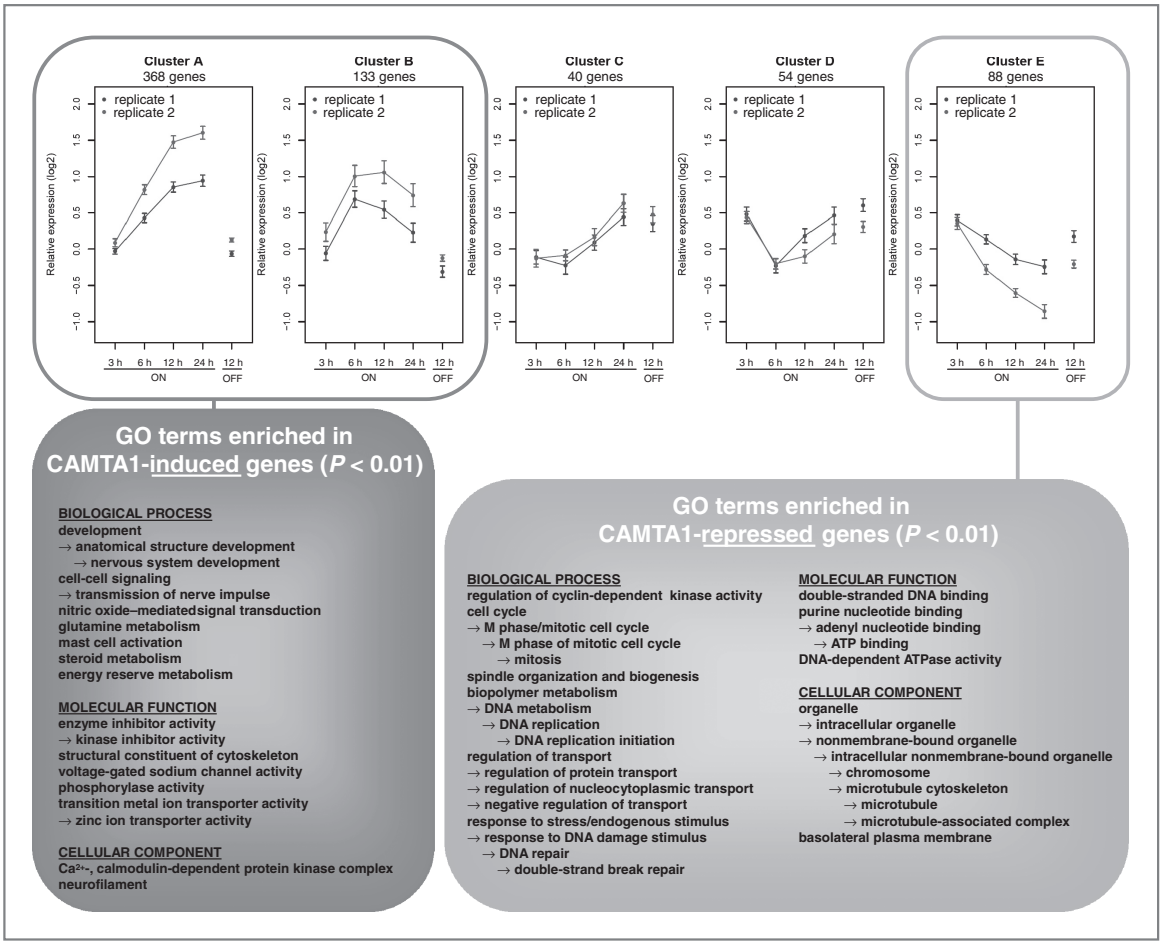


Figure 4. Genetic programs induced by CAMTA1 in SH-EP cells as determined by time-resolved whole-genome microarray expression analysis. RNA harvested at 3, 6, 12, and 24 hours after CAMTA1 induction and at 12 hours without induction (control) was hybridized against RNA harvested at time point 0 hour (uninduced). Experiments were done in 2 biological replicates. Gene clustering based on Pearson correlation coefficients revealed 5 clusters containing genes with common time-dependent expression profiles on CAMTA1 induction (Clusters A–E). Left panel, GO terms significantly enriched among CAMTA1 induced genes (clusters A + B) according to GOTM analysis ($P < 0.01$). Right panel, GO terms significantly enriched among CAMTA1 repressed genes (cluster E) according to GOTM analysis ($P < 0.01$).

expressed in a cell cycle dependent manner with highest levels in S and M phase (21), this may indicate that CAMTA1 acts as a negative regulatory factor during DNA synthesis and mitosis.

A variety of regulatory mechanisms could be responsible for the downregulation of *CAMTA1* in unfavorable neuroblastomas. In line with a haploinsufficiency model, *CAMTA1* expression is lower in 1p deleted neuroblastomas (4). However, low *CAMTA1* expression predicts poor outcome also within the subgroup of 1p nondeleted neuroblastomas, which calls for additional negative regulators of *CAMTA1* expression. A common mechanism mediating transcriptional repression of growth-regulating genes in tumors is methylation of cytosine residues in gene-associated CpG islands. However, we found no evidence for *CAMTA1* promotor methylation using methylation specific PCR on bisulfite treated neuroblastoma DNA samples (data not shown). Further epigenetic factors may play a role. It has been reported that histone deacetylase inhibitors reexpress silenced tumor suppressors including $p21^{WAF1/CIP1}$, p16, $p57^{Kip2}$, and $p19^{INK4d}$ (34). The induction of *CAMTA1* on treatment with the histone deacetylase inhibitors valproic acid and HC-toxin is in line with a similar regulatory model. Whether *CAMTA1* induction by HDAC inhibitors involves chromatin remodeling at the *CAMTA1* locus or whether factors upstream of *CAMTA1* are activated, remains to be investigated.

To pinpoint 1p36 tumor suppressor genes, a previous study used chromosome engineering generating mouse models with gain and loss of a region corresponding to human 1p36 (35). Gain of this region inhibited proliferation, whereas loss of the same region rendered cells sensitive to oncogenic transformation. In search of the gene(s) mediating this phenotype, several candidates, including *CAMTA1*, were knocked down to test whether their depletion could reverse the proliferation defect

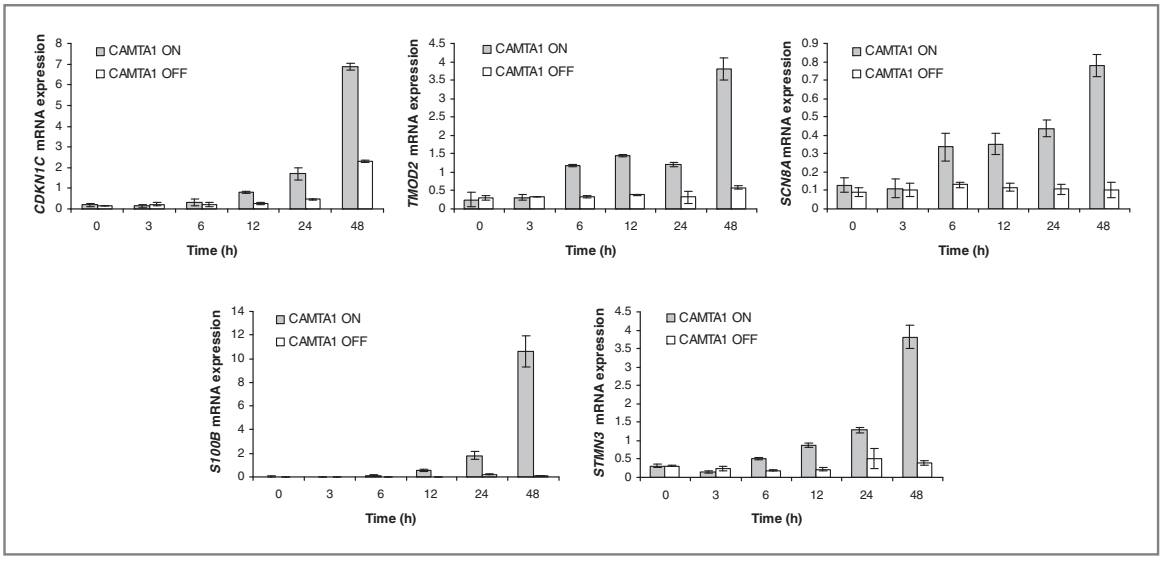


Figure 5. Validation of selected CAMTA1 targets by QPCR. Induction was tested in an independent SH-EP-CAMTA1 clone (mean ± SD, 3 replicates).

of mouse embryonic fibroblasts with gain of the 1p36 homologous region. In this assay, knockdown of most tested genes, including *CAMTA1*, had no significant effect, whereas knockdown of another gene (*CHD5*), functionally rescued the proliferation defect. The failure of *CAMTA1* to show an effect in this context is likely to be due to tissue specificity. *CAMTA1* is predominantly expressed in neuronal tissues and, in light of the data presented here, its growth suppressive effect in neuroblastoma cells is closely linked to its ability to induce effectors of neuronal differentiation. In a mouse embryonic fibroblast background, both the expression of *CAMTA1* and the potential to induce differentiation are likely to be limited.

Together, our data suggest that *CAMTA1* is a 1p36 tumor suppressor candidate that inhibits key features of malignant cells and is involved in neuronal differentiation. Understanding the function of CAMTA1 may help develop diagnostic tools and/or effective therapeutic strategies for children with unfavorable neuroblastoma. Further dissection of CAMTA1 downstream signaling and identification of mechanisms regulating *CAMTA1* will be the points of departure to reach this goal.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Magaly Santiago-Reichelt, Laura Sieber, and Yvonne Kahlert for excellent technical assistance, Kathy Astrahantseff for critical reading of the manuscript, and Barbara Hero and the German Neuroblastoma Study Group for providing clinical data.

Grant Support

BMBF: NGFN^{Plus} #01GS0896 (K.O. Henrich, H. Deubzer, M. Schwab, F. Westermann), #01GS0898 (T. Bauer, R. Konig), #01GS0895 (M. Fischer). BMBF: FORSYS Consortium Viroquant #0313923 (T. Bauer, R. Konig). EU (FP6): E.E.T. Pipeline #037260 (S. Gogolin, D. Muth, M. Schwab, F. Westermann).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 16, 2010; revised January 10, 2011; accepted February 7, 2011; published OnlineFirst March 8, 2011.

References

1. Maris JM, Weiss MJ, Guo C, Gerbing RB, Stram DO, White PS, et al. Loss of heterozygosity at 1p36 independently predicts for disease progression but not decreased overall survival probability in neuroblastoma patients: a Children's Cancer Group study. J Clin Oncol 2000;18:1888–99.
2. Bauer A, Savelyeva L, Claas A, Praml C, Berthold F, Schwab M. Smallest region of overlapping deletion in 1p36 in human neuroblastoma: a 1 Mbp cosmid and PAC contig. Genes Chromosomes Cancer 2001;31:228–39.
3. White PS, Thompson PM, Gotoh T, Okawa ER, Igarashi J, Kok M, et al. Definition and characterization of a region of 1p36.3 consistently deleted in neuroblastoma. Oncogene 2005;24:2684–94.
4. Henrich KO, Fischer M, Mertens D, Benner A, Wiedemeyer R, Brors B, et al. Reduced expression of CAMTA1 correlates with adverse outcome in neuroblastoma patients. Clin Cancer Res 2006;12:131–8.
5. Henrich KO, Westermann F. CAMTA1. In:Schwab M, editor. Encyclopedia of Cancer. 2nd ed. Berlin, New York, Heidelberg: Springer; 2008.
6. Bouche N, Scharlat A, Snedden W, Bouchez D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. J Biol Chem 2002;277:21851–61.
7. Nagase T, Ishikawa K, Suyama M, Kikuno R, Hirosawa M, Miyajima N, et al. Prediction of the coding sequences of unidentified human

genes. XII. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res 1998;5:355–64.
8. Henrich KO, Claas A, Praml C, Benner A, Mollenhauer J, Poustka A, et al. Allelic variants of CAMTA1 and FLJ10737 within a commonly deleted region at 1p36 in neuroblastoma. Eur J Cancer 2007;43:607–16.
9. Ichimura K, Vogazianou AP, Liu L, Pearson DM, Backlund LM, Plant K, et al. 1p36 is a preferential target of chromosome 1 deletions in astrocytic tumours and homozygously deleted in a subset of glioblastomas. Oncogene 2008;27:2097–108.
10. Barbashina V, Salazar P, Holland EC, Rosenblum MK, Ladanyi M. Allelic losses at 1p36 and 19q13 in gliomas: correlation with histologic classification, definition of a 150-kb minimal deleted region on 1p36, and evaluation of CAMTA1 as a candidate tumor suppressor gene. Clin Cancer Res 2005;11:1119–28.
11. Kim MY, Yim SH, Kwon MS, Kim TM, Shin SH, Kang HM, et al. Recurrent genomic alterations with impact on survival in colorectal cancer identified by genome-wide array comparative genomic hybridization. Gastroenterology 2006;131:1913–24.
12. Muth D, Ghazaryan S, Eckerle I, Beckett E, Pohler C, Batzler J, et al. Transcriptional repression of SKP2 is impaired in MYCN-amplified neuroblastoma. Cancer Res 2010;70:3791–802.
13. Schmitt M, Pawlita M. High-throughput detection and multiplex identification of cell contaminations. Nucleic Acids Res 2009;37:e119.
14. Ehemann V, Hashemi B, Lange A, Otto HF. Flow cytometric DNA analysis and chromosomal aberrations in malignant glioblastomas. Cancer Lett 1999;138:101–6.
15. von Heydebreck A, Huber W, Gentleman R. Differential Expression with the Bioconductor Project. Bioconductor Project Working Papers 2004;Working Paper 7.
16. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004;5:16.
17. Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R, et al. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. J Clin Oncol 2006;24:5070–8.
18. Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics 1992;48:73–85.
19. Schumacher M, Holländer N, Schwarzer G, Sauerbrei W. Prognostic factor studies. In:Crowley J, editor. Handbook of statistics in clinical oncology. New York: Marcel Dekker; 2001.
20. Hollander N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. Stat Med 2004;23:1701–13.
21. Nakatani K, Nishioka J, Itakura T, Nakanishi Y, Horinouchi J, Abe Y, et al. Cell cycle-dependent transcriptional regulation of calmodulin-binding transcription activator 1 in neuroblastoma cells. Int J Oncol 2004;24:1407–12.
22. Deubzer HE, Ehemann V, Kulozik AE, Westermann F, Savelyeva L, Kopp-Schneider A, et al. Anti-neuroblastoma activity of Helminthosporium carbonum (HC)-toxin is superior to that of other differentiating compounds in vitro. Cancer Lett 2008;264:21–8.
23. Deubzer HE, Ehemann V, Westermann F, Heinrich R, Mechtersheimer G, Kulozik AE, et al. Histone deacetylase inhibitor Helminthosporium carbonum (HC)-toxin suppresses the malignant phenotype of neuroblastoma cells. Int J Cancer 2008;122:1891–900.
24. Finkler A, Ashery-Padan R, Fromm H. CAMTAs: calmodulin-binding transcription activators from plants to human. FEBS Lett 2007;581:3893–8.
25. Nakagawara A. Molecular and developmental biology of neuroblastoma. In:Cheung N-K, Cohn S, editors. Neuroblastoma. Berlin, Heidelberg, New York: Springer; 2005.
26. Asgharzadeh S, Pique-Regi R, Sposto R, Wang H, Yang Y, Shimada H, et al. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. J Natl Cancer Inst 2006;98:1193–203.
27. Vermeulen J, De Preter K, Naranjo A, Vercruysse L, Van Roy N, Hellemans J, et al. Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study. Lancet Oncol 2009;10:663–71.
28. Warnat P, Oberthuer A, Fischer M, Westermann F, Eils R, Brors B. Cross-study analysis of gene expression data for intermediate neuroblastoma identifies two biological subtypes. BMC Cancer 2007;7:89.
29. Cinatl J Jr., Cinatl J, Driever PH, Kotchetkov R, Pouckova P, Kornhuber B, et al. Sodium valproate inhibits in vivo growth of human neuroblastoma cells. Anticancer Drugs 1997;8:958–63.
30. Matthay KK, Villablanca JG, Seeger RC, Stram DO, Harris RE, Ramsay NK, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. Children's Cancer Group. N Engl J Med 1999;341:1165–73.
31. Ohira M, Morohashi A, Inuzuka H, Shishikura T, Kawamoto T, Kageyama H, et al. Expression profiling and characterization of 4200 genes cloned from primary neuroblastomas: identification of 305 genes differentially expressed between favorable and unfavorable subsets. Oncogene 2003;22:5525–36.
32. Fischer M, Oberthuer A, Brors B, Kahlert Y, Skowron M, Voth H, et al. Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome. Clin Cancer Res 2006;12:5118–28.
33. Wu G, Fang Y, Lu ZH, Ledeen RW. Induction of axon-like and dendrite-like processes in neuroblastoma cells. J Neurocytol 1998;27:1–14.
34. Emanuele S, Lauricella M, Tesoriere G. Histone deacetylase inhibitors: apoptotic effects and clinical implications (Review). Int J Oncol 2008;33:637–46.
35. Bagchi A, Papazoglu C, Wu Y, Capurso D, Brodt M, Francis D, et al. CHD5 is a tumor suppressor at human 1p36. Cell 2007;128:459–75.

Research

Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas

Frank Westermann[¤*], Daniel Muth[¤*], Axel Benner[†], Tobias Bauer[‡], Kai-Oliver Henrich[*], André Oberthuer[§], Benedikt Brors[‡], Tim Beissbarth[¶], Jo Vandesompele[¥], Filip Pattyn[¥], Barbara Hero[§], Rainer König[‡], Matthias Fischer[§] and Manfred Schwab[*]

Addresses: [*]Department of Tumor Genetics, German Cancer Research Center, Im Neuenheimer Feld 280, Heidelberg, 69120, Germany. [†]Department of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, Heidelberg, 69120, Germany. [‡]Theoretical Bioinformatics, German Cancer Research Center, Im Neuenheimer Feld 280, Heidelberg, 69120, Germany. [§]Department of Pediatric Oncology, University Children's Hospital of Cologne, Kerpener Strasse 62, Cologne, 50924, Germany. [¶]Division of Molecular Genome Analysis, German Cancer Research Center, Im Neuenheimer Feld 580, Heidelberg, 69120, Germany. [¥]Center for Medical Genetics, Ghent University Hospital, De Pintelaan 185, Ghent, 9000, Belgium.

¤ These authors contributed equally to this work.

Correspondence: Frank Westermann. Email: f.westermann@dkfz.de

Published: 13 October 2008

Genome ***Biology*** 2008, **9:**R150 (doi:10.1186/gb-2008-9-10-r150)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/10/R150

Received: 6 August 2008
Revised: 19 September 2008
Accepted: 13 October 2008

Abstract

Background: Amplified *MYCN* oncogene resulting in deregulated MYCN transcriptional activity is observed in 20% of neuroblastomas and identifies a highly aggressive subtype. In *MYCN* single-copy neuroblastomas, elevated *MYCN* mRNA and protein levels are paradoxically associated with a more favorable clinical phenotype, including disseminated tumors that subsequently regress spontaneously (stage 4s-non-amplified). In this study, we asked whether distinct transcriptional MYCN or c-MYC activities are associated with specific neuroblastoma phenotypes.

Results: We defined a core set of direct MYCN/c-MYC target genes by applying gene expression profiling and chromatin immunoprecipitation (ChIP, ChIP-chip) in neuroblastoma cells that allow conditional regulation of *MYCN* and c-*MYC*. Their transcript levels were analyzed in 251 primary neuroblastomas. Compared to localized-non-amplified neuroblastomas, MYCN/c-MYC target gene expression gradually increases from stage 4s-non-amplified through stage 4-non-amplified to *MYCN* amplified tumors. This was associated with MYCN activation in stage 4s-non-amplified and predominantly c-MYC activation in stage 4-non-amplified tumors. A defined set of MYCN/c-MYC target genes was induced in stage 4-non-amplified but not in stage 4s-non-amplified neuroblastomas. In line with this, high expression of a subset of MYCN/c-MYC target genes identifies a patient subtype with poor overall survival independent of the established risk markers amplified *MYCN*, disease stage, and age at diagnosis.

Conclusions: High MYCN/c-MYC target gene expression is a hallmark of malignant neuroblastoma progression, which is predominantly driven by c-MYC in stage 4-non-amplified tumors. In contrast, moderate MYCN function gain in stage 4s-non-amplified tumors induces only a restricted set of target genes that is still compatible with spontaneous regression.

Background

Neuroblastoma is the most common extracranial malignant solid tumor in early childhood. Clinical courses are highly variable, ranging from spontaneous regression to therapy-resistant progression. Clinical and biological features, such as age at diagnosis, disease stage, numerical (ploidy) and structural chromosomal alterations (*MYCN* gene amplification; 1p, 3p, 11q deletions; 17q gain), are associated with patient outcome [1,2]. Amplified *MYCN* oncogene identifies a subtype with poor prognosis [3] and is consistently associated with high *MYCN* mRNA and protein levels. There is strong experimental evidence (ectopic *MYCN* expression in cell lines, N-*myc* transgenic neuroblastoma mouse model) that increased MYCN activity is involved in tumor initiation and progression of at least a subset of neuroblastomas [4,5].

The *MYC* gene family members, c-*MYC*, *MYCN* and *MYCL*, are involved in the biology of many cancer types. They encode basic helix-loop-helix leucine zipper proteins that are found as heterodimers with their obligate partner protein, MAX [6]. The MYC-MAX heterodimer binds to DNA consensus core binding sites, 5'-CACGTG-3' or variants thereof (E-boxes), which preferentially leads to transcriptional activation of target genes. Repression of target genes by MYC proteins has also been described [7]. This seems to be independent of the binding of MYC proteins to E-boxes, but involves a cofactor, Miz-1, that tethers MYC-MAX to gene promoters, such as *p15* and *p21*. Enhanced activity of MYC transcription factors contributes to almost every aspect of tumor formation: unrestricted cell proliferation, inhibition of differentiation, cell growth, angiogenesis, reduced cell adhesion, metastasis, and genomic instability [6,8]. In contrast, MYC transcription factors, including MYCN, also sensitize cells for apoptosis, a function that should inhibit tumor formation and that could also be involved in spontaneous tumor regression [9].

Spontaneous tumor regression does occur in neuroblastoma, at a higher frequency than in any other cancer type. This process resembles the physiological concurrence of massive cellular suicide (apoptosis) and differentiation of a few neurons along the sympathoadrenal cell lineage in the normal development of the sympathetic nervous system. Spontaneous regression is most frequently observed in a subset of disseminated *MYCN* single-copy neuroblastomas (non-amplified (NA)), termed stage 4s (stage 4s-NA) [10]. However, population-based screening studies for neuroblastomas in Japan, Quebec and Germany suggest that spontaneous regression also occurs in other neuroblastoma subtypes, predominantly localized (stages 1, 2, 3) neuroblastomas (localized-NA) [11-13]. Paradoxically, *MYCN* mRNA and protein levels are higher in favorable localized-NA and, particularly, in stage 4s-NA tumors than in stage 4-NA tumors with poor outcome [14-16], but they do not reach the levels observed in *MYCN* amplified tumors. In line with this, neuroblastoma cells with elevated *MYCN* expression retain their capacity to undergo apoptosis [17] or neuronal differentiation [18]. Thus, it has been speculated that MYCN does not only mediate malignant progression in *MYCN* amplified tumors, but is also either involved or at least compatible with spontaneous regression in favorable neuroblastomas. In contrast, a functional role of MYCN in stage 4-NA tumors with low *MYCN* levels is questionable. Here, other transcription factors or pathways within or outside the MYC family of transcription factors could be more relevant. Neuroblastoma-derived cell lines that lack amplified *MYCN* generally express c-MYC rather than MYCN, often at higher levels than normal tissues [19,20]. However, transcriptional activity of MYCN or c-MYC as reflected by the transcript levels of direct MYCN/c-MYC target genes in relation to *MYCN* and c-*MYC* levels has not yet been defined in neuroblastoma subtypes.

Here, we defined a core set of MYCN and c-MYC target genes by using oligonucleotide microarrays and a neuroblastoma cell line that allows conditional expression of *MYCN* or c-*MYC*. Direct regulation of these target genes by MYCN/c-MYC was assessed by analyzing the binding of MYCN and c-MYC protein to target gene promoters using PCR- and array-based chromatin immunoprecipitation (ChIP and ChIP-chip, respectively) in different neuroblastoma cell lines. We further investigated the expression of these direct MYCN/c-MYC target genes in relation to *MYCN* and c-*MYC* expression in different clinical neuroblastoma subtypes. In addition, the association of MYCN/c-MYC target gene expression with overall survival independent of the well-established markers - amplified *MYCN*, disease stage and age at diagnosis - was demonstrated.

Results

Inverse correlation of *MYCN* and c-*MYC* expression in neuroblastoma subtypes

c-*MYC* mRNA levels are very low in *MYCN* amplified tumors (Figure 1), which is due to high MYCN protein repressing c-*MYC* mRNA expression [20]. Previous quantitative PCR analyses in a cohort of 117 neuroblastoma patients revealed that mRNA levels of *MYCN* are significantly lower in stage 4-NA than in stage 4s-NA ($p = 0.008$) and localized-NA neuroblastomas (stages 1, 2, 3; $p = 0.03$) [14]. To test whether this lower expression of *MYCN* in stage 4-NA tumors is due to elevated c-MYC activity that represses *MYCN* expression, we analyzed c-*MYC* and *MYCN* mRNA levels in a cohort of 251 primary neuroblastoma tumors using a customized 11K oligonucleotide microarray (other *MYC* gene family members were not differently expressed (data not shown)). Although c-*MYC* mRNA levels were not significantly higher in stage 4-NA (n = 52) than in localized-NA tumors (n = 138), we found an inverse correlation of *MYCN* and c-*MYC* expression between stage 4s-NA (n = 30) and stage 4-NA tumors. Stage 4-NA tumors showed lower expression of *MYCN* and higher expression of c-*MYC*, whereas stage 4s-NA tumors showed lower expression of c-*MYC* and higher expression of *MYCN* (Figure 1; $p = 0.008$ for c-*MYC*, $p = 0.07$ for *MYCN*).

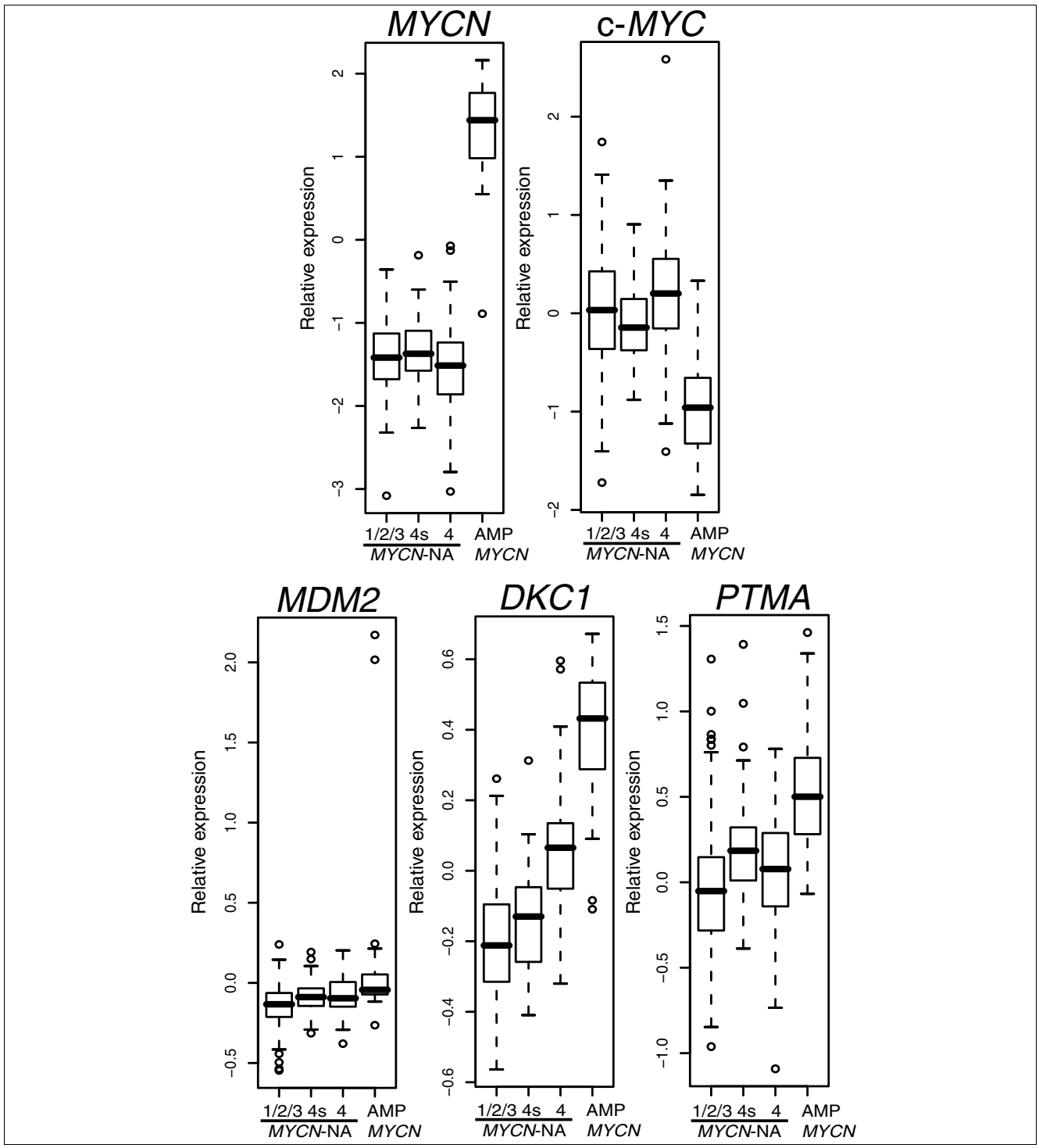


Figure 1
Inverse correlation of *MYCN* and c-*MYC* mRNA levels in neuroblastoma subtypes. Relative mRNA expression is shown for *MYCN* and c-*MYC* as well as for *MDM2*, *DKC1*, and *PTMA*, three direct targets of MYCN/c-MYC. Data are represented as box plots: horizontal boundaries of boxes represent the 25th and 75th percentile. The 50th percentile (median) is denoted by a horizontal line in the box and whiskers above and below extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box. A set of 251 primary neuroblastoma tumors was analyzed consisting of 138 localized-NA (stage 1/2/3), 30 stage 4s-NA, 52 stage 4-NA and 31 *MYCN* amplified (AMP) neuroblastoma tumors. Gene expression levels from stage 4s-NA, stage 4-NA, and *MYCN* amplified tumors were compared pair-wise with those of localized-NA tumors as reference. Differential gene expression was assessed for each gene by using the Mann-Whitney test (cut-off of $p < 0.05$).

Because increased activity of MYCN in stage 4s-NA or c-MYC in stage 4-NA tumors should both result in high expression of shared target genes compared to localized-NA neuroblastomas, we analyzed known direct MYCN/c-MYC target genes, namely *MDM2* [21], *DKC1* [22], and *PTMA* [23], in neuroblastoma subtypes. As expected, the highest expression of all three transcripts was observed in *MYCN* amplified tumors (Figure 1; $p < 0.001$ for all three transcripts, n = 31). *MDM2* mRNA levels were higher in stage 4-NA ($p = 0.005$) and stage 4s-NA ($p = 0.03$) than in localized-NA tumors (the expression range of *MDM2* is large because of two *MYCN* amplified tumors with non-syntenic co-amplification of *MDM2* (data not shown)). Similarly, *DKC1* and *PTMA* expression was higher in stage 4-NA ($p < 0.001$ for *DKC1*, $p = 0.02$ for *PTMA*) and in stage 4s-NA ($p = 0.03$ for *DKC1*, $p = 0.007$ for *PTMA*) than in localized-NA tumors. These results suggest an increased MYCN/c-MYC activity also in stage 4s-NA (MYCN) and in stage 4-NA (predominantly c-MYC) compared to localized-NA tumors. However, higher *DKC1* mRNA levels in stage 4-NA tumors and higher *PTMA* mRNA levels in stage 4s-NA tumors also suggest differential regulation of MYCN/c-MYC target genes in these subtypes. To further analyze MYCN/c-MYC activity as well as differential regulation of MYCN/c-MYC target genes in neuroblastoma subtypes, we thought to define a comprehensive set of target genes directly regulated by MYCN and/or c-MYC in neuroblastoma cells.

Repression of endogenous c-*MYC* by targeted expression of a *MYCN* transgene in SH-EP^{MYCN} cells defines c-MYC- and MYCN-regulated genes

To identify MYCN/c-MYC-regulated genes in neuroblastoma cells, we employed the experimental system SH-EP^{MYCN}, which stably expresses a tetracycline-regulated *MYCN* transgene [23]. Exponentially growing SH-EP^{MYCN} cells cultured with tetracycline express c-MYC but almost no MYCN protein (Figure 2a). Induction of MYCN by removing tetracycline from the medium is associated with a rapid reduction of c-MYC at the mRNA and protein levels. c-MYC reduction occurs prior to the full expression of ectopically induced MYCN protein (Figure 2a). Accordingly, mRNA levels of direct MYCN/c-MYC targets, such as *PTMA* and *DKC1*, initially decline before accumulating MYCN protein leads to the re-induction of these genes. Similar profiles were observed with direct MYCN target genes, such as *MDM2* and *MCM7* (Additional data file 1).

We used SH-EP^{MYCN} cells for a global search of MYCN and c-MYC target genes in neuroblastoma cells using a customized neuroblastoma oligonucleotide microarray (11K, Agilent) that was enriched with probes for genes differentially expressed in neuroblastoma subtypes and for direct MYCN/c-MYC target genes [14,24]. Gene expression profiles of SH-EP^{MYCN} cells at 2, 4, 8, 12, 24, and 48 hours after targeted *MYCN* expression were generated. Self-organizing maps (SOMs) were used to capture the predominant pattern of gene expression. This analysis yielded 504 clusters (best matching units (BMUs)) consisting, on average, of 20 clones per cluster (Additional data file 1). We searched for clusters with characteristic gene expression profiles of direct MYCN/c-MYC target genes. In addition, known c-MYC target genes from a public database [25] and known MYCN target genes from a literature search were mapped to the 504 clusters (Additional data file 2). A significant enrichment of known MYCN/c-MYC targets was found in 6 clusters (clusters 140, 168, 195, 280, 308, and 336; $p < 0.05$, adjusted for multiple testing), consisting of 167 genes. The genes in these six clusters were induced by MYCN and c-MYC in SH-EP^{MYCN} cells. Based on their average gene expression profiles, we grouped the clusters into two subgroups, I and II. Subgroup I genes (clusters 140, 168, and 195) were expressed at equal levels in SH-EP^{MYCN} cells expressing endogenous c-MYC (2 hours) and in those fully expressing ectopic MYCN (24 and 48 hours), despite the fact that the maximum protein level of MYCN was significantly higher than that of endogenous c-MYC (Figure 2a; Additional data file 1). This indicates that subgroup I genes are regulated by MYCN, and also suggests that they are less responsive to MYCN than to c-MYC in SH-EP^{MYCN} cells. The mRNA levels of subgroup II genes (clusters 280, 308, and 336) were highest in SH-EP^{MYCN} cells fully expressing ectopic MYCN and followed the combined absolute c-MYC and MYCN protein levels during the time course experiment. We also found clusters with MYCN and c-MYC repressed genes (for example, subgroup III; Additional data file 1). However, enrichment of known MYCN/c-MYC repressed genes from the literature/database in defined clusters was not found using our statistical cut-off (after adjustment for multiple testing, no cluster showed $p < 0.05$). This was at least partly due to the fact that in SH-EP^{MYCN} cells, some genes were repressed by MYCN but not by c-MYC (subgroup IV). In addition, c-MYC repressed genes from different experimental systems compiled in the c-MYC target gene database were not necessarily repressed by MYCN and/or c-MYC in SH-EP^{MYCN} cells.

Therefore, we focused on genes for further validation that were induced by both MYCN and c-MYC proteins in SH-EP^{MYCN} cells and grouped into subgroup I and II. We extracted all available promoters from the genes represented on the array and scanned for canonical E-boxes (CACGTG) and for the 12 bp MYCN position-weight matrix [26] within -2 kb and +2 kb of the transcriptional start site. We ranked all 504 clusters according to the relative number of putative MYCN/c-MYC binding sites in each cluster. All clusters from subgroups I and II were among the 15 top-ranked clusters with enrichment of predicted MYCN/c-MYC binding sites (data not shown).

To further validate target gene regulation by MYCN/c-MYC in neuroblastoma cells, we performed ChIP-chip using a 244K oligonucleotide promoter microarray (Agilent). We analyzed the binding of MYCN and c-MYC to the promoters of the 147 subgroup I and II genes that were represented on the 244K promoter microarray. We used five neuroblastoma cell lines

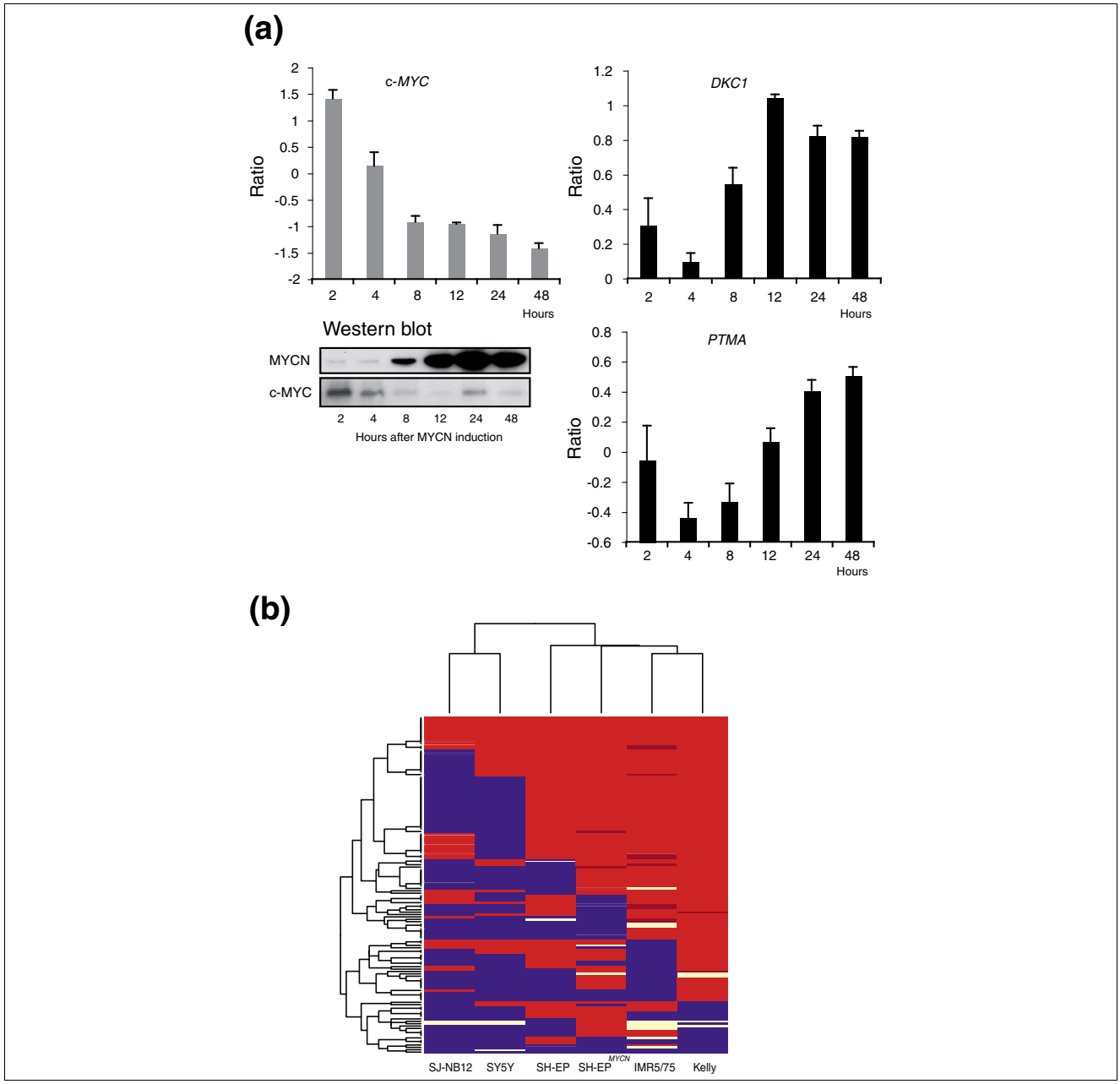


Figure 2
Identification and validation of MYCN/c-MYC target genes in neuroblastoma cell lines. **(a)** Repression of endogenous c-MYC by targeted expression of a *MYCN* transgene in SH-EP^{MYCN} cells defines MYCN/c-MYC-regulated genes. MYCN and c-MYC protein levels were monitored in a time series after removing tetracycline in exponentially growing SH-EP^{MYCN} cells that stably express a tetracycline-regulated *MYCN* transgene. Mean and standard deviation of the relative mRNA levels of *MYC*, *DKC1* and *PTMA* are given from two time series experiments as measured by a customized neuroblastoma oligo microarray. **(b)** Hierarchical clustering of MYCN- and c-MYC binding to 140 target gene promoters as measured by ChIP-chip in 6 neuroblastoma cell lines. ChIP-chip results of 140 MYCN/c-MYC target genes from 5 neuroblastoma cell lines that preferentially express either high levels of MYCN (SH-EP^{MYCN}, IMR5/75 (approximately 75 copies of *MYCN*) and Kelly (approximately 100-120 copies of *MYCN*)) or c-MYC (SJ-NB12 and SY5Y). Additionally, as an intermediate type, parental SH-EP cells were analyzed. SH-EP cells preferentially express c-MYC, but also low levels of MYCN. ChIP-chip experiments were performed with a monoclonal antibody against human MYCN and a polyclonal antibody against human c-MYC for each neuroblastoma cell line. A cut-off for positive binding was set for both transcription factors to >4-fold enrichment for one and >2-fold enrichment of at least one of the two neighboring probes. MYCN/c-MYC-binding is color-coded as follows: blue, c-MYC binding; red, MYCN/c-MYC binding; dark red, MYCN binding; light yellow, lack of MYCN/c-MYC binding. Hierarchical clustering was used to group neuroblastoma cell lines according to their MYCN/c-MYC-binding pattern. Differentiation between MYCN and c-MYC-binding was mainly achieved through the monoclonal MYCN antibody. The polyclonal antibody against c-MYC also gave positive binding signals for a large set of analyzed target gene promoters in neuroblastoma cell lines with high MYCN that lack c-MYC expression (SH-EP^{MYCN}, IMR5/75 and Kelly).

that either preferentially express high levels of MYCN (SH-EP^{MYCN}, IMR5/75 (approximately 75 copies of *MYCN*), and Kelly (approximately 100-120 copies of *MYCN*)) or c-MYC (SJ-NB12 and SY5Y). Additionally, as an intermediate type, parental SH-EP cells were analyzed, which preferentially express c-MYC, but also MYCN at low level [20,23]. ChIP-chip experiments were performed with a monoclonal antibody against human MYCN and a polyclonal antibody against human c-MYC for each of the neuroblastoma cell lines. A cut-off for positive binding was defined as >4-fold enrichment for one probe together with >2-fold enrichment for at least one of the two neighboring probes compared to input control. In addition, we manually inspected each of the MYCN and c-MYC-binding profiles from the 147 genes. Seven genes were excluded from the analysis because the probe sets for the genes mapped within the genes but outside the target gene promoter regions (all profiles for Kelly and SJ-NB12 cell lines are given in Additional data files 3 and 4, respectively; MYCN- and c-MYC-binding results are given in Additional data files 5-7). We also performed PCR-based ChIP for selected candidate genes (n = 13; Additional data file 8), which all showed analogous results to ChIP-chip (data not shown). Almost all 140 target gene promoters showed binding of MYCN and/or c-MYC in the six analyzed neuroblastoma cell lines as measured by ChIP-chip (Figure 2b). Intriguingly, hierarchical clustering of neuroblastoma cell lines according to the MYCN/c-MYC-binding pattern clearly separated MYCN- and c-MYC-expressing neuroblastoma cell lines. Differentiation between MYCN and c-MYC binding was mainly achieved through the monoclonal anti-MYCN antibody. The polyclonal antibody against c-MYC also gave positive binding signals for a large set of target gene promoters in neuroblastoma cell lines with high MYCN that lack detectable c-MYC expression (SH-EP^{MYCN}, IMR5/75 and Kelly). This was most likely due to unspecific binding of the polyclonal c-MYC antibody to MYCN in these cells. Nevertheless, the lack of binding of MYCN to a large set of target gene promoters in the c-MYC-expressing cells, SJ-NB12 and SY5Y, and the positive binding of c-MYC to almost all of these target gene promoters in these cells allowed the distinction between MYCN and c-MYC. Taken together, these results indicate that the genes from subgroups I and II represent a core set of target genes directly regulated by either MYCN or c-MYC in neuroblastoma cells dependent on which MYC protein is expressed.

Gradual increase of MYCN/c-MYC target gene expression from stage 4s-NA through stage 4-NA to *MYCN* amplified tumors

To determine transcriptional activity of MYCN/c-MYC proteins in primary neuroblastomas (n = 251), we analyzed differential expression of subgroup I and II genes in neuroblastoma subtypes using the Global test as proposed by Goeman *et al.* [27]. Almost all these genes (154 of 167; 92%) showed highest expression in *MYCN* amplified tumors, suggesting that regulation of these genes by MYCN is similar in neuroblastoma cell lines and tumors. Compared to localized-NA tumors (stages 1, 2, 3), expression of subgroup I and II genes was significantly associated with stage 4s-NA (p = 0.002), stage 4-NA ($p < 0.001$) and *MYCN* amplified tumors ($p < 0.001$). Global test results further indicated that an increasing number of MYCN/c-MYC target genes was induced from stage 4s-NA through stage 4-NA to *MYCN* amplified tumors (Additional data files 9-11). To further illustrate this, we grouped each of the 154 genes into one of four classes based on pair-wise comparisons (Mann-Whitney test, cut-off $p < 0.05$). These were, compared to localized-NA tumors: overexpressed in *MYCN* amplified and in stage 4s-NA tumors (class 1); overexpressed in *MYCN* amplified, stage 4-NA and stage 4s-NA tumors (class 2); overexpressed in *MYCN* amplified tumors (class 3); overexpressed in *MYCN* amplified and stage 4-NA tumors (class 4) (Figure 3). Compared to localized-NA tumors, 25 (16%) of the 154 MYCN/c-MYC target genes, including *CCT4*, *FBL*, *MDM2*, *NCL*, *NPM1*, *PTMA*, and *TP53*, were expressed at higher levels in stage 4s-NA tumors (Table 1). Eighty-eight (57%) of the 154 MYCN/c-MYC target genes, including 21 of those overexpressed also in stage 4s-NA tumors, were expressed at higher levels in stage 4-NA than in localized-NA tumors (Table 1, class 2; Additional data file 5). Accordingly, stage 4-NA tumors shared overexpression of 68 of 154 direct MYCN/c-MYC target genes (44%), including *AHCY*, *RUVBL1*, *PHB*, *CDK4*, and *MRPL3*, with *MYCN* amplified tumors. Together, this indicates that besides *MYCN* amplified tumors, stage 4-NA tumors, and to a lesser extent stage 4s-NA tumors, also show higher MYCN/c-MYC activity compared to localized-NA tumors. In line with this, we also found lower mRNA levels of an increasing number of MYCN/c-MYC repressed genes from stage 4s-NA (10 out of 68 (15%) *in vitro* validated repressed genes that are also lower in *MYCN* amplified tumors) through stage 4-NA (34 out of 68 (50%)) to *MYCN* amplified tumors (68 out of 102 *in vitro* validated repressed genes had the lowest expression levels in *MYCN* amplified tumors (67%)). Based on the relative expression of *MYCN* and c-*MYC* in neuroblastoma subtypes, we propose that elevated MYCN activity in stage 4s-NA tumors induces only a restricted set of MYCN/c-MYC target genes, whereas elevated c-MYC activity in stage 4-NA tumors induces a larger set of MYCN/c-MYC target genes.

High expression of MYCN/c-MYC target genes is a robust marker of poor overall survival independent of genomic *MYCN* status, age at diagnosis and disease stage

Having shown that MYCN/c-MYC target gene activation is also associated with distinct neuroblastoma subtypes, we wanted to test whether MYCN/c-MYC activity as determined by the expression levels of their target genes is associated with overall survival and improves outcome prediction independent of known risk markers. We used the Global test to test the influence of each of the 504 experimentally defined gene clusters on overall survival directly, without the intermediary of single gene testing. The p-values for each cluster were adjusted for multiple testing and ranked according to their

Table 1

MYCN/c-MYC target genes overexpressed in stage 4s-NA compared to localized-NA tumors (classes 1 and 2)

Probe	Gene name	Class	BMU	Group	MYCN/c-MYC-fold change*	c-MYC target DB†	Validated by ChIP‡
A_24_P311604	*C4orf28*	1	195	I	1.38		+
A_23_P102420	*CCT4*	1	168	I	1.31		+
A_23_P5551	*NCL*	1	308	II	1.69	Up	+
A_23_P44836	*NT5DC2*	1	140	I	1.40		+
A_32_P139196	*C13ORF25V_1*	2	308	II	3.83		ND
A_24_P133488	*CDCA4*	2	140	I	1.45		+
A_23_P137143	*DKC1*	2	308	II	1.93	Up	+
A_23_P216396	*EXOSC2*	2	308	II	1.83		+
A_23_P78892	*FBL*	2	195	I	1.93	Up	+
A_24_P228796	*GAGE7B*	2	195	I	1.27		ND
A_23_P41025	*GNL3*	2	308	II	1.80	Up	ND
A_32_P8120	*GNL3*	2	308	II	1.81	Up	ND
A_23_P398460	*HK2*	2	280	II	1.71	Up	+
Hs172673.9	*Hs172673.9*	2	168	I	1.73		+
A_23_P502750	*MDM2*	2	336	II	1.19	ChIP	+
A_23_P92261	*MGC2408*	2	280	II	2.14		+
A_23_P50897	*MKI67IP*	2	280	II	1.97	Up	+
A_23_P214037	*NPM1*	2	140	I	1.61	Up	+
A_23_P57709	*PCOLCE2*	2	308	II	2.40		+
A_24_P34632	*PTMA*	2	308	II	2.21	Up	+
A_23_P126825	*SLC16A1*	2	195	I	1.22		+
A_23_P126291	*SNRPE*	2	336	II	1.49		+
A_23_P117068	*SNRPF*	2	336	II	1.44		+
A_23_P31536	*SSBP1*	2	336	II	1.24		+
A_23_P26810	*TP53*	2	140	I	1.44	Up	+

*Fold change expression in SH-EP^{MYCN} cells after *MYCN* induction. †c-MYC target gene database entry [25]: Up, upregulated; ChIP, validated by ChIP. ‡Validation of MYCN/c-MYC binding using ChIP in this study (Additional data files 5-7). BMU, best matching unit; ND, not determined.

association with overall survival. Table 2 gives the association with overall survival of the six MYCN/c-MYC target gene clusters and the rank in relation to all other clusters. In a separate analysis, we determined the association with overall survival for each of the 504 experimental gene clusters adjusted for amplified *MYCN*, stage 4 versus stages 1, 2, 3, and 4s, and age at diagnosis ≥1.5 years (Table 2). These well-established risk markers highly correlated with poor outcome in univariate analyses ($p < 0.001$ for each of these three markers). As expected, the Global test without adjustment for co-variables indicated that all MYCN/c-MYC target gene clusters were significantly associated with poor overall survival ($p < 0.001$). Intriguingly, all six MYCN/c-MYC target gene clusters remained significantly associated with overall survival after adjusting for amplified *MYCN*, stage 4 versus stages 1, 2, 3, and 4s, and age at diagnosis ≥1.5 years. Of note, two of the MYCN/c-MYC target gene clusters (clusters 168 and 140, both from subgroup I showing a higher responsiveness to c-MYC than to MYCN in SH-EP^{MYCN}) revealed the strongest association with overall survival of all 504 clusters after adjusting for co-variables (Table 2). Figure 4 shows the association with overall survival for each gene from cluster 168 with and without adjustment for co-variables. Most of the genes within this cluster, such as *AHCY*, *ARD1A*, *CDK4*, *HSPD1*, *PHB*, *RUVBL1*, and *TRAP1*, remained associated with overall survival after adjustment for co-variables. A less significant association with overall survival was observed for clusters with MYCN/c-MYC repressed genes: clusters 454, 482, 484, and 486 were associated with poor overall survival without adjustment for co-variables in the Global test ($p < 0.001$, adjusted for multiple testing), but they showed no significant association with poor overall survival when adjusting for the co-variables amplified *MYCN*, stage 4 versus stages 1, 2, 3, and 4s, and age at diagnosis ≥1.5 years. We also asked whether direct MYCN/c-MYC target genes as defined by our analyses are represented in previously published gene expression-based classifiers that distinguish low-risk from high-risk neuroblastomas independent of other risk markers. Gene lists from these studies hardly overlapped, making interpretation difficult. The overlap with our MYCN/c-MYC target gene list was defined by using the gene names as common identifiers. Indeed, different genes defined by our study as direct MYCN/c-MYC target genes were represented in the gene expression classifier gene lists: from the 44 genes over-

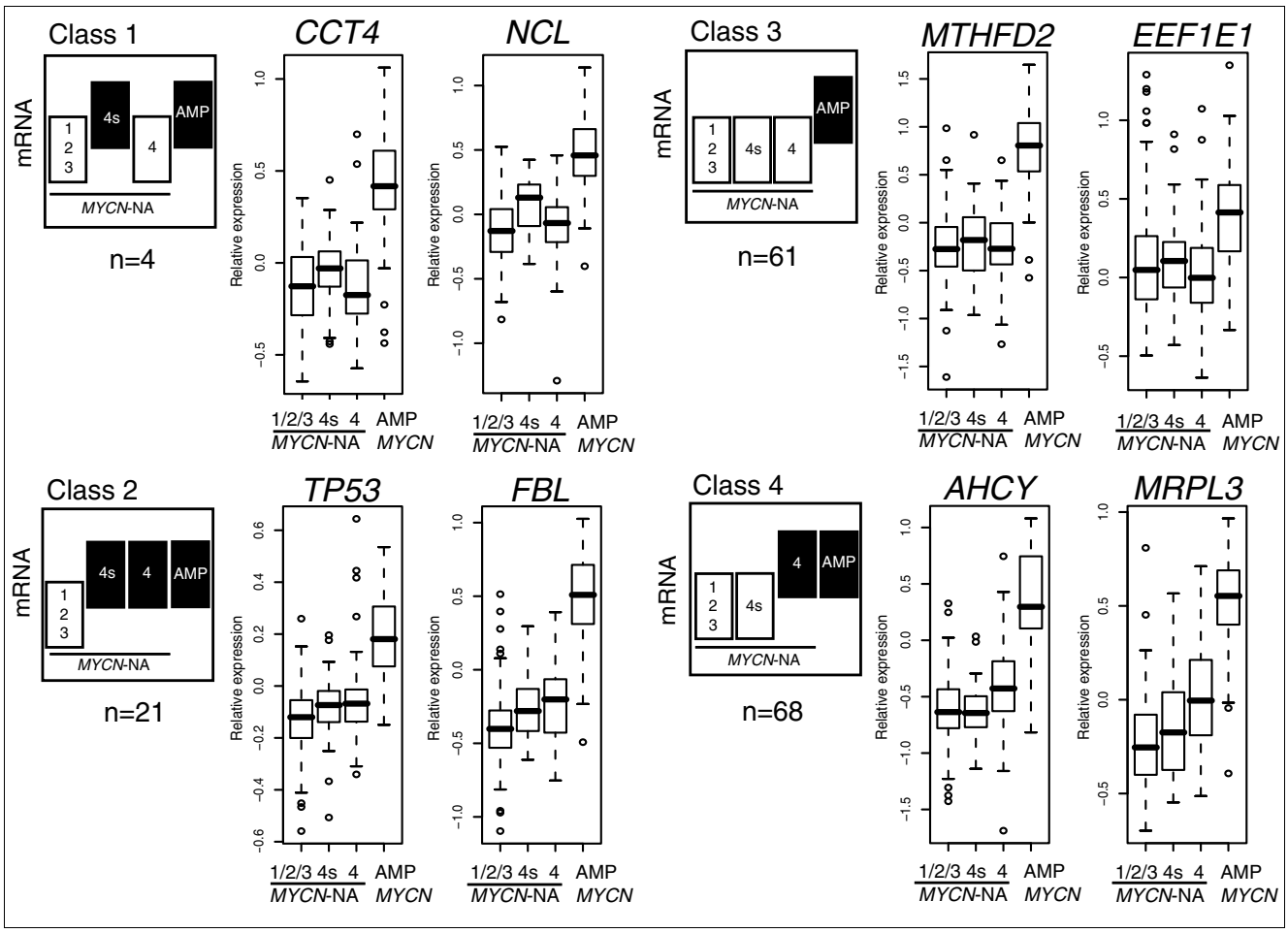


Figure 3
Expression of MYCN/c-MYC target genes in neuroblastoma subtypes. Differential expression was analyzed for each of the genes (n = 154) in *MYCN* amplified (AMP), stage 4s-NA and stage 4-NA tumors using localized-NA (stage 1/2/3) tumors as reference in pair-wise comparisons (Mann-Whitney test, cut-off $p < 0.05$, black). We grouped each of these 154 genes into one of four classes based on their relative expression in clinically relevant neuroblastoma subtypes. These classes were, compared to localized-NA tumors: overexpressed in *MYCN* amplified and in stage 4s-NA tumors (class 1; *CCT4* and *NCL*); overexpressed in *MYCN* amplified, stage 4-NA and stage 4s-NA tumors (class 2; *TP53* and *FBL*); overexpressed in *MYCN* amplified tumors (class 3; *MTHFD2* and *EEF1E1*); and overexpressed in *MYCN* amplified and stage 4-NA tumors (class 4; *AHCY* and *MRPL3*).

expressed in high-risk neuroblastomas independent of other markers described by Schramm *et al.* [28], we identified 10 genes directly regulated by MYCN/c-MYC (*DDX21*, *SCL25A3*, *EIFA4A2*, *NME1*, *NME2*, *TKT*, *LDHA*, *LDHB*, *HSPD1*, *HSPCB*); from the 20 genes overexpressed in high-risk neuroblastomas independent of other markers described by Ohira *et al.* [29], we identified 5 genes directly regulated by MYCN/c-MYC (*EEF1G*, *AHCY*, *TP53*, *ENO1*, *TKT*); and from the 66 genes overexpressed in high-risk neuroblastomas independent of other markers described by Oberthuer *et al.* [24], we identified 7 genes directly regulated by MYCN/c-MYC (*PRDX4*, *MRPL3*, *SNRPE*, *FBL*, *LOC200916*, *PAICS*, *AHCY*; Figure 5). Together, these results show that MYCN/c-MYC activity as determined by the expression status of a subset of MYCN/c-MYC target genes is significantly associated with poor overall survival independent of other established markers and is a consistent element of gene expression-based neuroblastoma risk classification systems.

Discussion

In this study, we analyzed MYCN and c-MYC activity as reflected by the expression levels of a core set of direct MYCN/c-MYC targets in neuroblastoma subtypes. As expected, the highest expression levels of MYCN/c-MYC targets were observed in *MYCN* amplified tumors. However, we found that besides *MYCN* amplified tumors, subtypes of *MYCN* single-copy tumors, namely stage 4-NA and, to a lesser extent, stage 4s-NA, also showed increased MYCN/c-MYC target gene activation compared to localized-NA tumors. In general, low *MYCN* mRNA and protein levels are found in most stage 4-NA tumors [14-16], which does not explain the high mRNA levels of MYCN/c-MYC target genes in this sub-

Table 2

Association of MYCN/c-MYC target gene clusters with overall survival in primary neuroblastomas (n = 251)

Cluster	Number of genes	Rank OS*	*p*-value OS†	Rank OS with CV*	*p*-value OS with CV†
168 (I)	19	3	<0.0001	1	0.0004
140 (I)	38	4	<0.0001	2	0.0006
195 (I)	21	31	<0.0001	12	0.0060
308 (II)	33	18	<0.0001	26	0.0161
280 (II)	32	29	<0.0001	37	0.0232
336 (II)	26	51	<0.0001	45	0.0280

*Rank of all 504 clusters tested for association with overall survival (OS) using the Global test without and with adjustment for co-variables (CV; amplified *MYCN*, stages 1, 2, 3, 4s versus 4, age at diagnosis ≥1.5 years). †*p*-value from Global test adjusted for multiple testing. In the Cluster column, I or II gives the cluster group as defined by the SOM analysis using SH-EP^{MYCN} cells.

type. Here, we describe an inverse correlation of *MYCN* and c-*MYC* expression levels in stage 4-NA and stage 4s-NA tumors. From experiments in neuroblastoma cell lines, it is known that MYCN and c-MYC control their expression via autoregulatory loops and via repressing each other at defined promoter sites [20]. Neuroblastoma cell lines with high expression of MYCN as a result of amplification lack c-MYC expression. Whenever MYCN and c-MYC are co-expressed in neuroblastoma cell lines, c-MYC expression predominates. Together, this suggests that increased activity of c-MYC represses MYCN in a substantial number of stage 4-NA tumors. In contrast, an inverse regulation, namely the repression of c-*MYC* by MYCN, is found in *MYCN* amplified and, to a lesser extent, in stage 4s-NA tumors. It is important to note that localized-NA tumors also express *MYCN* as well as c-*MYC* and it is likely that they are active because these tumors frequently show high tumor cell proliferation indices [14]. Nevertheless, in localized-NA tumors, we did not observe that one MYC transcription factor dominates over the other, such as in the other neuroblastoma subtypes.

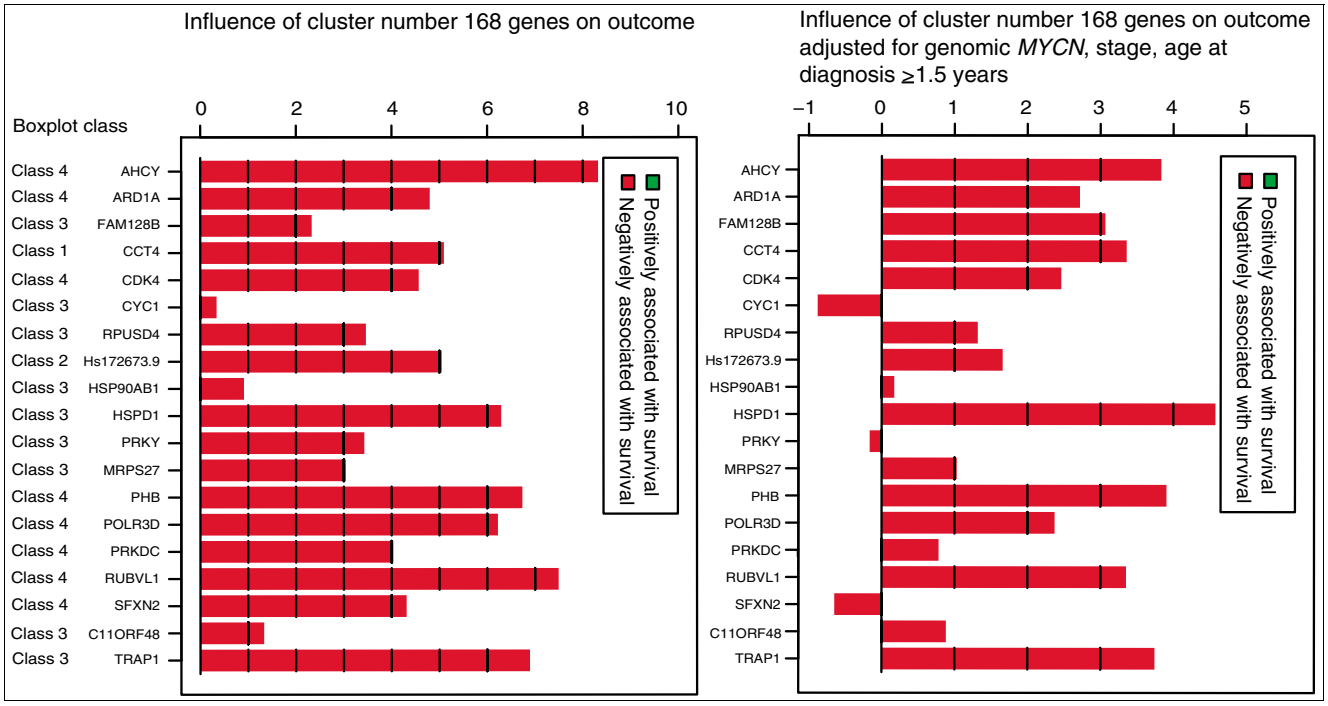
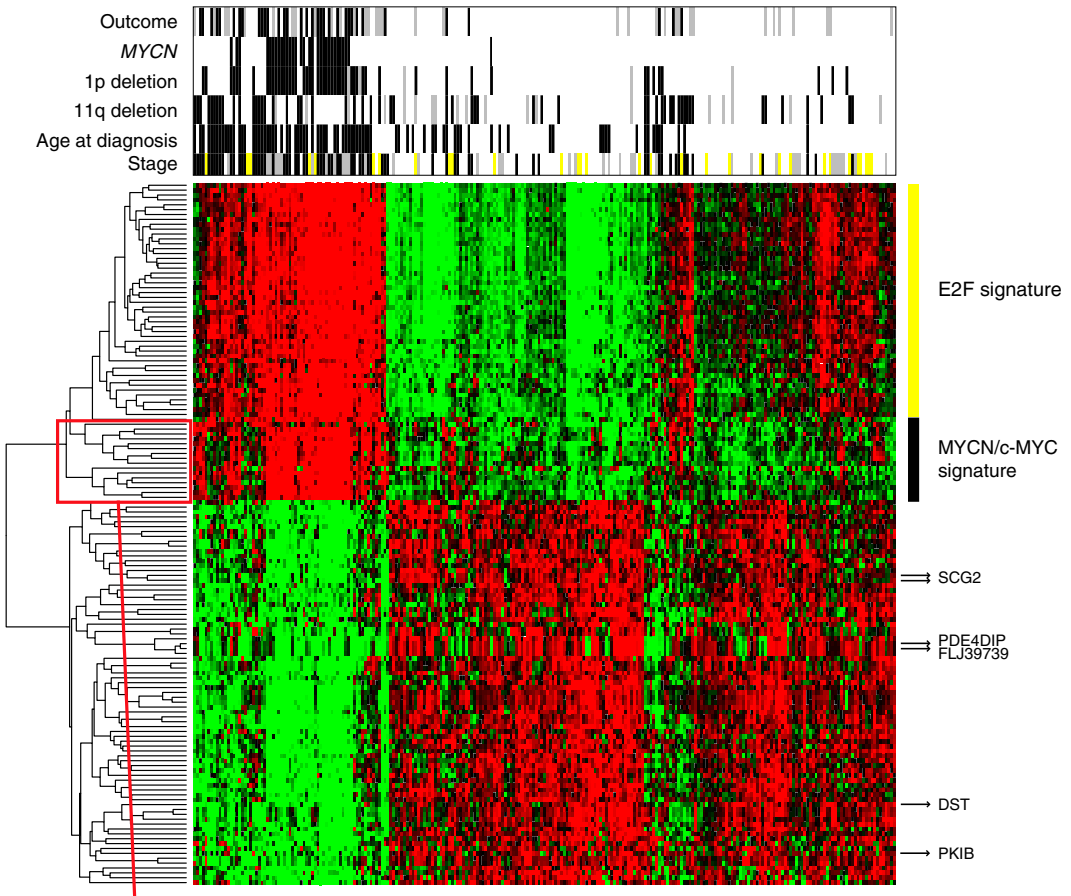


Figure 4
Association of cluster 168 genes with overall survival. The two gene plots illustrate the influence on overall survival of each gene from cluster 168. The gene plot gives the influence on overall survival without (left) and with (right) adjustment for the variables genomic *MYCN* status, age at diagnosis (≥1.5 years), and disease stage (stages 1, 2, 3, 4s versus stage 4). The gene plot shows a bar and a reference line for each gene tested. In a survival model, the expected height is zero under the null hypothesis that the gene is not associated with the clinical outcome (= reference line). Marks in the bars indicate by how many standard deviations the bar exceeds the reference line. The bars are colored to indicate a negative (red) association of a gene's expression with overall survival. In addition, the boxplot class is given for each gene.



Rank classifier	Symbol	Cluster	Ebox	MYCN/c-MYC-regulated	c-MYC TGDB
130	VRK1	447		2 up	*UP
108	FLJ10151	40		2.4 up	
137	HIST1H1C	482		down/up	
3	SLC25A5	383		1.8 up	
90	PRDX4	221	346	1.6 up	*UP
58	HSPC163	247		down/up	
139	MRPL3	336	81	2.2 up	
57	SNRPE	336		1.8 up	
107	ZNF525	25		1.7 up	
105	RPL36A	1		1.4 up	*UP
110	FBL	223/195	-4	2.6 up	*UP
100	BX119435	466		5.6 up	
129	NOLA1	307	284	1.5 up	*UP
4	GMPS	445		1.8 up	
95	LOC200916	364		2.5 up	
1	PAICS	364	(-283/18/617/685)	2.2 up	*UP
5	AHCY	196		1.6 up	*UP
13	AHCY	168		2.2 up	*UP

Figure 5
Representation of MYCN/c-MYC target genes in a gene expression-based neuroblastoma risk stratification system. Two-way hierarchical cluster analysis using 144 oligonucleotide probes from the gene expression-based classifier and the 251 patients from the entire cohort. Clinical characteristics (outcome, white = no event, gray = relapse/progression, black = death due to neuroblastoma; genomic *MYCN* status, white = NA, black = amplified; chromosome 1p status, white = normal, black = 1p deleted, gray = not available; chromosome 11q status, white = normal, black = 11q deleted, gray = not available; age at diagnosis, white <1.5 years, black ≥1.5 years; disease stage, white = stage 1, 2, gray = stage 3, yellow = stage 4s, black = stage 4) are added to the heatmap of gene expression. The gene expression cluster with direct MYCN/c-MYC target genes is highlighted. The Rank Classifier column gives the classifier rank found by the Prediction Analysis for Microarrays algorithm and a complete 10-times-repeated 10-fold cross validation. The Cluster column gives the results from the SOM analysis using gene expression profiles from SH-EP^{MYCN} cells. The MYCN/c-MYC regulated column gives the fold changes after MYCN induction. The Ebox column gives the position of a canonical E-box in the promoter. The c-MYC TGDB column gives the entries in the public c-MYC target gene database. *UP, upregulated.

Our findings further indicate that MYCN/c-MYC target gene activation gradually increases from stage 4s-NA through stage 4-NA to *MYCN* amplified tumors. High expression of a large number of MYCN/c-MYC target genes was found in stage 4-NA and *MYCN* amplified tumors, but not in stage 4s-NA tumors, which is probably involved in the divergent clinical outcome of these subtypes. This also suggests that MYCN in stage 4s tumors is a weaker transactivator than c-MYC in stage 4-NA tumors. Whether this effect is due to the cellular context in which they are expressed and/or due to different functions of the two MYC proteins in neuroblastoma cells is unclear. In favor of a cellular context factor, we observed that promoter constructs from the *PTMA* gene, which is highly expressed in stage 4s NA and MYCN amplified tumors, showed a strong activation in N-type but not S-type neuroblastoma cell lines despite similar MYCN protein levels (unpublished data). In favor of different functions of the two MYC proteins, our analyses in SH-EP^{MYCN} cells suggest that a large number of MYCN/c-MYC target genes (subgroup I genes) are less responsive to MYCN than to c-MYC. Another unsolved question is which molecular mechanisms induce elevated MYCN activity in stage 4s-NA tumors or elevated c-MYC activity in stage 4-NA tumors. Candidate pathways involved in differential regulation of MYC proteins are the Sonic hedgehog pathway (Shh) for MYCN activation [30] and the Wnt/beta-catenin pathway for c-MYC activation [31,32]. However, we observed that c-*MYC* mRNA levels are not significantly higher in stage 4-NA than in localized-NA tumors. This suggests that molecular mechanisms that increase c-MYC protein abundance/stability or simply c-MYC activity are involved in MYCN/c-MYC target gene activation in stage 4 tumors.

Our data are in line with a model where stage 4s-NA tumors exhibit a moderate MYCN function gain compared to localized-NA tumors. Both subtypes usually have favorable outcome. Most localized-NA tumors are cured by surgery alone or even regress spontaneously. Stage 4s-NA tumors frequently regress spontaneously but regression can also be induced by a 'mild' chemotherapy. We found that stage 4s-NA tumors express, on average, the highest *MYCN* mRNA levels of all non-amplified tumors [14]. From the experimentally defined direct MYCN target genes, only a restricted set of 25 genes, including *CCT4*, *FBL*, *MDM2*, *NCL*, *NPM1*, *PTMA*, and *TP53*, was overexpressed in stage 4s-NA compared to localized-NA tumors, indicating that elevated *MYCN* in stage 4s-NA tumors only partially activates its downstream target genes. On the one hand, this suggests that moderate MYCN function gain in stage 4s-NA tumors is involved in the metastatic phenotype. On the other hand, moderate MYCN function gain in this subtype is still compatible with, or might even favor, spontaneous regression. From the list of MYCN target genes overexpressed in stage 4s-NA tumors, *TP53* as a pro-apoptotic gene, and *MDM2*, coding for the direct inhibitor of p53 and mediating pro-tumorigenic activities, are strong candidates to be involved in the unique phenotype of stage 4s-NA tumors. However, it is important to note that *TP53* and *MDM2* are co-expressed at higher levels also in stage 4-NA and *MYCN* amplified tumors. Both subtypes initially respond to therapy, but rapidly acquire resistance and frequently show progression/relapse, suggesting that additional conditions activating MDM2 and/or suppressing TP53 functions are acquired. In line with this, alterations disrupting the p14-MDM2-p53 pathway, such as *MDM2* amplification, *p14* methylation/deletion, and *TP53* mutations are found in neuroblastoma cell lines that were established from relapsed patients [33]. In this context, it remains to be shown whether small compounds that selectively inhibit MDM2, such as nutlin-3, and that induce proliferation arrest and apoptosis in neuroblastoma cell lines [34,35] represent a new therapeutic option for high-risk neuroblastomas.

Conclusions

High expression of a defined subset of direct MYCN/c-MYC target genes turned out to be a robust marker for poor overall survival independent of the established markers, amplified *MYCN*, disease stage (stage 4 versus stages 1, 2, 3, and 4s) and age at diagnosis (≥1.5 years). Recently, several gene expression-based neuroblastoma risk stratification systems have been developed that predict outcome more accurately than established risk markers [24,28,29]. Unfortunately, the classifier gene lists emerging from these studies hardly overlap, which has been ascribed to the different composition of the investigated cohorts and the different high-throughput gene expression platforms used. Our data show that markers of increased MYCN/c-MYC activity are consistently represented in these classifier gene lists, indicating that a gene expression-based classifier that reflects MYCN/c-MYC function should make an attractive tool for neuroblastoma classification and risk prediction.

Materials and methods

Patients

All patients from this study (n = 251) were enrolled in the German Neuroblastoma Trials NB90-NB2004 with informed consent and diagnosed between 1989 and 2004 (patient characteristics are in Additional data files 2 and 12). Tumor samples were collected prior to any cytoreductive treatment. The only criterion for patient selection was availability of sufficient amounts of tumor material. Tumor specimens were checked for at least 60% tumor content.

Neuroblastoma sample preparation and gene expression analysis

Gene expression profiles from the tumors were generated as dye-flipped dual-color replicates using customized 11K oligonucleotide microarrays as previously described [24]. The 11K Agilent microarray was constructed in our laboratory based on extensive neuroblastoma transcriptome information from different whole-genome analyses from primary tumors and

neuroblastoma cell lines. These also include comparative transcriptome analysis of *MYCN* amplified versus not amplified tumors as well as of neuroblastoma cell lines with variable/conditional MYCN/c-MYC expression that allowed the enrichment with MYCN/c-MYC-regulated genes [14,24] (unpublished data). The reference for each tumor RNA was an RNA pool of 100 neuroblastoma tumor samples. Data normalization and quality control is described in Additional data file 2. All raw and normalized microarray data are available at the ArrayExpress database (Accession: E-TABM-38) [36].

Neuroblastoma cell line experiments and SOM analysis

The SH-EP^{MYCN} cell line, previously also denoted as TET21N [23], expressing a *MYCN* transgene under the control of a tetracycline-repressible element was used to generate gene expression profiles from different time points after *MYCN* induction showing variable MYCN and c-MYC levels. RNA isolation from SH-EP^{MYCN} cells was performed as previously described [14]. Gene expression profiles were generated as dye-flipped dual-color replicates using the same customized 11K oligonucleotide microarray platform used for the tumor samples. The reference for RNA from SH-EP^{MYCN} cells after *MYCN* induction was RNA from SH-EP^{MYCN} cells cultured in parallel that lack *MYCN* expression. Gene expression profiles from SH-EP^{MYCN} cells with variable MYCN and c-MYC levels were taken for a SOM analysis (Additional data file 2). Protein expression was assessed by immunoblotting using 50 μg of total cell lysates from the cell line experiments as previously described [37]. Blots were probed with antibodies directed against MYCN (SantaCruz, sc-53993, Santa Cruz, CA, USA) and c-MYC (SantaCruz, sc-764, Santa Cruz, CA, USA).

ChIP, ChIP-chip and protein analysis

Chromatin immunoprecipitation was performed as described previously [38,39] using 10 μg of MYCN (SantaCruz, sc-53993), c-MYC (SantaCruz, sc-764) [40,41] and normal mouse IgG (SantaCruz, sc-2025) antibodies and Dynabeads ProteinG (Invitrogen, Carlsbad, CA, USA). Eluted and purified MYCN-ChIP-DNA (1 μl) of *IMR5/75* and *SH-EP*^{MYCN} was used as a template in PCR reactions running for 35 cycles. The primer sequences are given in Additional data file 8. In addition, ChIP-DNA templates from *SH-EP*^{MYCN}, *SH-EP*, *Kelly*, *IMR5/75*, *SJNB-12* and *SY5Y* cells using MYCN and c-MYC antibodies were amplified for DNA microarray analysis (Agilent Human Promoter ChIP-chip Set 244K) using the WGA (Sigma-Aldrich, St. Louis, MO, USA) method [42]. DNA labeling, array hybridization and measurement were performed according to Agilent mammalian ChIP-chip protocols. For the visualization of ChIP-chip results, the cureos package v0.2 for R was used (available upon request). The *in silico* promoter analysis for the identification of putative MYC binding sites (canonical and non-canonical E-boxes) is described in Additional data file 2.

Differential gene expression and survival analysis

Differential gene expression of *MYCN*/c-*MYC* and their target genes in neuroblastoma tumors was evaluated for stage 4s-NA, stage 4-NA and *MYCN* amplified using localized-NA tumors (stages 1, 2, 3) as reference using Goeman's Global test and the Wilcoxon rank sum test. A result was judged as 'statistically significant' at a p-value of 0.05 or smaller. Differential expression of *MYCN* was evaluated in two partially overlapping cohorts, one measured by quantitative PCR [14] and the other by oligo microarray (the overlap was 101 patients). To test the association of *MYCN in vitro* clusters with overall survival (death due to neuroblastoma disease), Goeman's Global test was used [27]. To evaluate the influence of gene expression on outcome independent of established markers, the Global test was adjusted for the following covariables: genomic *MYCN* status, stage of the disease (stage 4 versus stages 1, 2, 3, and 4s), and age at diagnosis (≥1.5 years versus <1.5 years). Because of multiple testing of probably dependent gene clusters, p-values were adjusted according to Benjamini and Yekutieli [43] to control the false discovery rate of 5%.

Abbreviations

ChIP, PCR-based chromatin immunoprecipitation; ChIP-chip, array-based chromatin immunoprecipitation; NA, non-amplified; SOM, self-organizing map.

Authors' contributions

FW designed and coordinated the study. FW and DM interpreted results and drafted the manuscript. AO, MF, AB, BB and FW carried out array-based expression profiling and data analyses of neuroblastoma tumor samples and cell lines. BH was responsible for clinical data management. TB and RK performed *in silico* promoter analyses. JV and FP contributed samples and performed literature searches of MYCN/c-MYC target genes. DM performed chromatin immunoprecipitation experiments. DM, TB and FW analyzed ChIP-chip data. AB, BH and FW carried out global test and survival analyses. FW, DM, KOH, JV, FP and MS contributed to the manuscript. All authors read and approved the final manuscript.

Additional data files

The following additional data are available. Additional data file 1 is a figure showing a Cluster map of genetic programs regulated by conditional expression of c-MYC and MYCN proteins in SH-EP^{MYCN} cells. Additional data file 2 is a document describing in more detail the methods and materials. Additional data files 3 and 4 are sets of figures showing ChIP-chip results of MYCN/c-MYC target genes in the Kelly and SJ-NB12 cell lines. Additional data files 5, 6 and 7 are tables listing MYCN/c-MYC target genes overexpressed in stage 4s-NA, stage 4-NA and *MYCN* amplified tumors, respectively, compared to localized-NA tumors. Additional data file 8 is a table

of genes and primers selected to confirm ChIP-chip results. Additional data files 9, 10 and 11 are figures showing the association of MYCN/c-MYC induced genes with neuroblastoma subtypes using the Global test. Additional data file 12 is a table providing patient data.

Acknowledgements

We thank Steffen Bannert and Yvonne Kahlert for technical assistance. We thank the German Neuroblastoma Tumor Bank for providing tumor samples, the German Neuroblastoma Study Group (study chair Frank Berthold) for providing clinical data and the reference laboratories for providing molecular data. This work was supported by program project grants from the Krebshilfe, BMBF (NGFN2 and Kompetenznetz Pediatric Oncology/Hematology) and the EU. The platform iCHIP (Integration Center of HIgh throughPut experiments) has been used for the annotation of this study. Jo Vandesompele is a postdoctoral researcher of the Research Foundation - Flanders (FWO-Vlaanderen). Filip Pattyn is supported by a grant of the Ghent University Special Research Fund (BOF).

References

1. Schwab M, Westermann F, Hero B, Berthold F: **Neuroblastoma: biology and molecular and chromosomal pathology.** *Lancet Oncol* 2003, **4:**472-480.
2. Vandesompele J, Baudis M, De Preter K, Van Roy N, Ambros P, Bown N, Brinkschmidt C, Christiansen H, Combaret V, Lastowska M, Nicholson J, O'Meara A, Plantaz D, Stallings R, Brichard B, Broecke C Van den, De Bie S, De Paepe A, Laureys G, Speleman F: **Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma.** *J Clin Oncol* 2005, **23:**2280-2299.
3. Seeger RC, Brodeur GM, Sather H, Dalton A, Siegel SE, Wong KY, Hammond D: **Association of multiple copies of the N-myc oncogene with rapid progression of neuroblastomas.** *N Engl J Med* 1985, **313:**1111-1116.
4. Weiss WA, Aldape K, Mohapatra G, Feuerstein BG, Bishop JM: **Targeted expression of MYCN causes neuroblastoma in transgenic mice.** *EMBO J* 1997, **16:**2985-2995.
5. Schwab M, Varmus HE, Bishop JM: **Human N-myc gene contributes to neoplastic transformation of mammalian cells in culture.** *Nature* 1985, **316:**160-162.
6. Adhikary S, Eilers M: **Transcriptional regulation and transformation by Myc proteins.** *Nat Rev Mol Cell Biol* 2005, **6:**635-645.
7. Kleine-Kohlbrecher D, Adhikary S, Eilers M: **Mechanisms of transcriptional repression by Myc.** *Curr Top Microbiol Immunol* 2006, **302:**51-62.
8. Prochownik EV, Li Y: **The ever expanding role for c-Myc in promoting genomic instability.** *Cell Cycle* 2007, **6:**1024-1029.
9. Fulda S, Lutz W, Schwab M, Debatin KM: **MycN sensitizes neuroblastoma cells for drug-induced apoptosis.** *Oncogene* 1999, **18:**1479-1486.
10. Pritchard J, Hickman JA: **Why does stage 4s neuroblastoma regress spontaneously?** *Lancet* 1994, **344:**869-870.
11. Sawada T, Hirayama M, Nakata T, Takeda T, Takasugi N, Mori T, Maeda K, Koide R, Hanawa Y, Tsunoda A, *et al.*: **Mass screening for neuroblastoma in infants in Japan. Interim report of a mass screening study group.** *Lancet* 1984, **2:**271-273.
12. Woods WG, Tuchman M, Robison LL, Bernstein M, Leclerc JM, Brisson LC, Brossard J, Hill G, Shuster J, Luepker R, Byrne T, Weitzman S, Bunin G, Lemieux B: **A population-based study of the usefulness of screening for neuroblastoma.** *Lancet* 1996, **348:**1682-1687.
13. Schilling FH, Spix C, Berthold F, Erttmann R, Sander J, Treuner J, Michaelis J: **Children may not benefit from neuroblastoma screening at 1 year of age. Updated results of the population based controlled trial in Germany.** *Cancer Lett* 2003, **197:**19-28.
14. Westermann F, Henrich KO, Wei JS, Lutz W, Fischer M, Konig R, Wiedemeyer R, Ehemann V, Brors B, Ernestus K, Leuschner I, Benner A, Khan J, Schwab M: **High Skp2 expression characterizes high-risk neuroblastomas independent of MYCN status.** *Clin Cancer Res* 2007, **13:**4695-4703.
15. Cohn SL, London WB, Huang D, Katzenstein HM, Salwen HR, Reinhart T, Madafiglio J, Marshall GM, Norris MD, Haber M: **MYCN expression is not prognostic of adverse outcome in advanced-stage neuroblastoma with nonamplified MYCN.** *J Clin Oncol* 2000, **18:**3604-3613.
16. Tang XX, Zhao H, Kung B, Kim DY, Hicks SL, Cohn SL, Cheung NK, Seeger RC, Evans AE, Ikegaki N: **The MYCN enigma: significance of MYCN expression in neuroblastoma.** *Cancer Res* 2006, **66:**2826-2833.
17. Lutz W, Fulda S, Jeremias I, Debatin KM, Schwab M: **MycN and IFN-gamma cooperate in apoptosis of human neuroblastoma cells.** *Oncogene* 1998, **17:**339-346.
18. Edsjö A, Nilsson H, Vandesompele J, Karlsson J, Pattyn F, Culp LA, Speleman F, Påhlman S: **Neuroblastoma cells with overexpressed MYCN retain their capacity to undergo neuronal differentiation.** *Lab Invest* 2004, **84:**406-417.
19. Sadée W, Yu VC, Richards ML, Preis PN, Schwab MR, Brodsky FM, Biedler JL: **Expression of neurotransmitter receptors and myc protooncogenes in subclones of a human neuroblastoma cell line.** *Cancer Res* 1987, **47:**5207-5212.
20. Breit S, Schwab M: **Suppression of MYC by high expression of NMYC in human neuroblastoma cells.** *J Neurosci Res* 1989, **24:**21-28.
21. Slack A, Chen Z, Tonelli R, Pule M, Hunt L, Pession A, Shohet JM: **The p53 regulatory gene MDM2 is a direct transcriptional target of MYCN in neuroblastoma.** *Proc Natl Acad Sci USA* 2005, **102:**731-736.
22. Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P, Roobeek I, Weis I, Voute PA, Schwab M, Versteeg R: **N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis.** *EMBO J* 2001, **20:**1383-1393.
23. Lutz W, Stöhr M, Schürmann J, Wenzel A, Löhr A, Schwab M: **Conditional expression of N-myc in human neuroblastoma cells increases expression of alpha-prothymosin and ornithine decarboxylase and accelerates progression into S-phase early after mitogenic stimulation of quiescent cells.** *Oncogene* 1996, **13:**803-812.
24. Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R, Ernestus K, König R, Haas S, Eils R, Schwab M, Brors B, Westermann F, Fischer M: **Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification.** *J Clin Oncol* 2006, **24:**5070-5078.
25. Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV: **An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets.** *Genome Biol* 2003, **4:**R69.
26. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34(Database issue):**D108-D110.
27. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC: **Testing association of a pathway with survival using gene expression data.** *Bioinformatics* 2005, **21:**1950-1957.
28. Schramm A, Schulte JH, Klein-Hitpass L, Havers W, Sieverts H, Berwanger B, Christiansen H, Warnat P, Brors B, Eils J, Eils R, Eggert A: **Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling.** *Oncogene* 2005, **24:**7902-7912.
29. Ohira M, Oba S, Nakamura Y, Isogai E, Kaneko S, Nakagawa A, Hirata T, Kubo H, Goto T, Yamada S, Yoshida Y, Fuchioka M, Ishii S, Nakagawara A: **Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas.** *Cancer Cell* 2005, **7:**337-350.
30. Hatton BA, Knoepfler PS, Kenney AM, Rowitch DH, de Alboran IM, Olson JM, Eisenman RN: **N-myc is an essential downstream effector of Shh signaling during both normal and neoplastic cerebellar growth.** *Cancer Res* 2006, **66:**8655-8661.
31. Liu X, Mazanek P, Dam V, Wang Q, Zhao H, Guo R, Jagannathan J, Cnaan A, Maris JM, Hogarty MD: **Deregulated Wnt/beta-catenin program in high-risk neuroblastomas without MYCN amplification.** *Oncogene* 2008, **27:**1478-1488.
32. van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, Hurlstone A, van der Horn K, Batlle E, Coudreuse D, Haramis AP, Tjon-Pon-Fong M, Moerer P, van den Born M, Soete G, Pals S, Eilers M, Medema R, Clevers H: **The beta-catenin/TCF-4 complex**

imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* 2002, **111:**241-250.
33. Carr J, Bell E, Pearson AD, Kees UR, Beris H, Lunec J, Tweddle DA: **Increased frequency of aberrations in the p53/MDM2/p14(ARF) pathway in neuroblastoma cell lines established at relapse.** *Cancer Res* 2006, **66:**2138-2145.
34. Van Maerken T, Speleman F, Vermeulen J, Lambertz I, De Clercq S, De Smet E, Yigit N, Coppens V, Philippé J, De Paepe A, Marine JC, Vandesompele J: **Small-molecule MDM2 antagonists as a new therapy concept for neuroblastoma.** *Cancer Res* 2006, **66:**9646-9655.
35. Barbieri E, Mehta P, Chen Z, Zhang L, Slack A, Berg S, Shohet JM: **MDM2 inhibition sensitizes neuroblastoma to chemotherapy-induced apoptotic cell death.** *Mol Cancer Ther* 2006, **5:**2358-2365.
36. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress—a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31:**68-71.
37. Wiedemeyer R, Westermann F, Wittke I, Nowock J, Schwab M: **Ataxin-2 promotes apoptosis of human neuroblastoma cells.** *Oncogene* 2003, **22:**401-411.
38. Strieder V, Lutz W: **E2F proteins regulate MYCN expression in neuroblastomas.** *J Biol Chem* 2003, **278:**2983-2989.
39. Lee TI, Johnstone SE, Young RA: **Chromatin immunoprecipitation and microarray-based analysis of protein location.** *Nat Protoc* 2006, **1:**729-748.
40. Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN: **Myc influences global chromatin structure.** *EMBO J* 2006, **25:**2723-2734.
41. Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B: **Myc-binding-site recognition in the human genome is determined by chromatin context.** *Nat Cell Biol* 2006, **8:**764-770.
42. O'Geen H, Nicolet CM, Blahnik K, Green R, Farnham PJ: **Comparison of sample preparation methods for ChIP-chip assays.** *Biotechniques* 2006, **41:**577-580.
43. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Stat* 2001, **29:**1165-1188.
44. Brodeur GM, Pritchard J, Berthold F, Carlsen NL, Castel V, Castelberry RP, De Bernardi B, Evans AE, Favrot M, Hedborg F, *et al.*: **Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment.** *J Clin Oncol* 1993, **11:**1466-1477.
45. Ambros PF, Ambros IM, SIOP Europe Neuroblastoma Pathology, Biology, and Bone Marrow Group: **Pathology and biology guidelines for resectable and unresectable neuroblastic tumors and bone marrow examination guidelines.** *Med Pediatr Oncol* 2001, **37:**492-504.
46. **CRAN** [http://www.r-project.org]
47. **Bioconductor** [http://www.bioconductor.org]
48. Buness A, Huber W, Steiner K, Sultmann H, Poustka A: **arrayMagic: two-colour cDNA microarray quality control and preprocessing.** *Bioinformatics* 2005, **21:**554-556.
49. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl 1):**S96-S104.
50. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, *et al.*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35(Database issue):**D610-D617.
51. **MYCNot** [http://medgen.ugent.be/MYCNot/]
52. Kohonen T: *Self-organizing Maps* Second edition. Heidelberg: Springer; 1997. [*Springer Series in Information Sciences*, volume 30]
53. **SOM Toolbox** [http://www.cis.hut.fi/projects/somtoolbox/]

Riehl et al. BMC Genomics 2010, 11:537
http://www.biomedcentral.com/1471-2164/11/537

BMC Genomics

RESEARCH ARTICLE Open Access

Identification of the Rage-dependent gene regulatory network in a mouse model of skin inflammation

Astrid Riehl[1], Tobias Bauer[2], Benedikt Brors[2], Hauke Busch[2,3,4], Regina Mark[1,8], Julia Németh[1], Christoffer Gebhardt[6], Angelika Bierhaus[7], Peter Nawroth[7], Roland Eils[2,5], Rainer König[2,5], Peter Angel[1*], Jochen Hess[1,8]

Abstract

Background: In the past, molecular mechanisms that drive the initiation of an inflammatory response have been studied intensively. However, corresponding mechanisms that sustain the expression of inflammatory response genes and hence contribute to the establishment of chronic disorders remain poorly understood. Recently, we provided genetic evidence that signaling via the receptor for advanced glycation end products (Rage) drives the strength and maintenance of an inflammatory reaction. In order to decipher the mode of Rage function on gene transcription levels during inflammation, we applied global gene expression profiling on time-resolved samples of mouse back skin, which had been treated with the phorbol ester TPA, a potent inducer of skin inflammation.

Results: Ranking of TPA-regulated genes according to their time average mean and peak expression and superimposition of data sets from wild-type (*wt*) and *Rage*-deficient mice revealed that Rage signaling is not essential for initial changes in TPA-induced transcription, but absolutely required for sustained alterations in transcript levels. Next, we used a data set of differentially expressed genes between TPA-treated *wt* and *Rage*-deficient skin and performed computational analysis of their proximal promoter regions. We found a highly significant enrichment for several transcription factor binding sites (TFBS) leading to the prediction that corresponding transcription factors, such as Sp1, Tcfap2, E2f, Myc and Egr, are regulated by Rage signaling. Accordingly, we could confirm aberrant expression and regulation of members of the E2f protein family in epidermal keratinocytes of Rage-deficient mice.

Conclusions: In summary, our data support the model that engagement of Rage converts a transient cellular stimulation into sustained cellular dysfunction and highlight a novel role of the Rb-E2f pathway in Rage-dependent inflammation during pathological conditions.

Background

A striking feature of many human cancers is an underlying and unresolved inflammation, which often predates the disease and orchestrates a tumor supporting microenvironment. Indeed, several lines of evidence, including population-based epidemiological and clinical studies as well as experimental animal model systems, highlighted chronic infection and persistent inflammation as major risk factors for various types of cancer [1,2]. Thus, molecular mechanisms converting a transient inflammatory tissue reaction into a tumor promoting microenvironment as well as signaling and gene regulatory networks implicated in cellular communication between tumor and immune cells will be auspicious targets for innovative strategies of translational cancer research.

Recently, we could show that the receptor for advanced glycation end products (Rage) drives the strength and maintenance of inflammation during tumor promotion in a mouse model of inflammation-associated skin carcinogenesis [3]. Accordingly, tumor formation in mutant mice with *Rage* deletion (*Rage*$^{-/-}$) was impaired in this model, but also in a tumor model of colitis-induced colon cancer [3,4].

* Correspondence: p.angel@dkfz-heidelberg.de
[1]Signal Transduction and Growth Control, German Cancer Research Center (DKFZ), DKFZ-ZMBH Alliance, Heidelberg, Germany
Full list of author information is available at the end of the article

Rage is a multi-ligand as well as pattern recognition receptor of the immunoglobulin super-family with low expression levels in most adult tissues. However, Rage expression increases at sites of inflammation, mainly on inflammatory cells, endothelial cells and epithelial cells, and propagates cellular dysfunction in numerous inflammation-related pathological states, such as diabetes, vascular disease, neurodegeneration, chronic inflammation, and cancer [5-7].

With respect to Rage signaling, several target genes have been identified in the past, including pro-inflammatory mediators, matrix metalloproteinases, and adhesion proteins, however, their expression critically depends on the cell type, its microenvironment, and quality of the stimulus [8]. In the process of neoplastic transformation and malignant progression, activation of Rage by its ligands, such as advanced glycation end products (AGEs), high mobility group box-1 (Hmgb1), and members of the S100 protein family, can stimulate tumor cell proliferation, invasion, chemoresistance, and metastasis [9-11]. Rage ligands derived from cancer cells can also support the establishment of a pro-tumorigenic microenvironment by activation of leukocytes, vascular cells, fibroblasts, and modulation of immune tolerance [11]. Although multiple intracellular signaling pathways, including MAP kinases, Rho GTPases, PI3K, JAK/STAT, and NF-κB, have been found to be altered following Rage stimulation, the molecular mechanisms how Rage triggers intracellular signaling to regulate cellular decisions remain largely elusive, and the identity of direct signaling molecules downstream of the receptor are still unknown [5,12-14].

In order to elucidate how Rage receptor signaling converts a transient stimulus into a long lasting response, global gene expression kinetics were recorded with skin samples of *wt* and *Rage*$^{-/-}$ mice upon TPA stimulation. We applied a recently published computational analysis tool that enables a global, holistic view on cellular responses over a time frame of hours based on dynamic transcription level data [15], and identified the characteristic duration and temporal order of transient and Rage-dependent events upon TPA stimulation. Subsequently, a computational approach was applied to predict transcription factors that are implicated in the Rage-dependent regulation of pro-inflammatory gene expression, and thus, to identify novel key molecules as putative targets for innovative strategies of anti-inflammatory therapy.

Results

Identification of Rage-dependent gene expression upon TPA treatment of mouse back skin

In order to identify alterations in the gene expression profile during the process of skin inflammation we applied TPA on the back skin of *wt* and *Rage*$^{-/-}$ mice and prepared total RNA at consecutive time points after treatment (6, 12, 24, and 48 hours following TPA application in three individual animal experiments). The RNA was hybridized on whole mouse genome oligonucleotide microarrays followed by feature extraction and quantile normalization procedure (Figure 1A). The gene fold expression was calculated with respect to non-treated controls (0 h), and TPA-responsive genes in samples of *wt* back skin were ranked according to their combined averaged mean and peak expression within the experimental time window of 48 hours for each individual kinetic series. Subsequently, we identified a common subset of 341 genes among the 1,000 highest ranked genes in all three experiments with a small variance between the experiments (Figure 1B and see Additional file 1). These genes were further separated into six expression profile sets according to k-means clustering (see Additional file 2). Most candidate genes were found in cluster 3 (n = 125) or in cluster 6 (n = 84), representing genes that were either TPA-repressed or TPA-induced within 6 hours and maintained altered expression for at least 24 hours (Figure 1C). Interestingly, when we considered the transcript levels of these genes in *Rage*$^{-/-}$ back skin and superimposed both *wt* and *Rage*$^{-/-}$ data sets we found a comparable response in both genotypes at 6 hours. However, initial transcript level responses ceased to basal levels in *Rage*$^{-/-}$ skin between 12 and 24 hours upon stimulation, whereas the response was sustained in *wt* animals. Our data suggest the existence of two phases of the TPA response: an initial Rage-independent response that is followed by a second Rage-dependent maintenance of the altered transcript levels (Figure 1C).

Next, linear models with empirical Bayesian correction were applied to identify differentially expressed genes between *wt* and *Rage*$^{-/-}$ back skin at the investigated time points after TPA administration. In line with preceding analyses and previous results [3], genes (n = 122) that differ significantly between both genotypes were only found 24 hours after TPA stimulation (see Additional file 3). According to their temporal expression pattern, differentially expressed genes were further divided by unsupervised hierarchical clustering of their correlation distance (one minus the Pearson correlation coefficient) into three sub-clusters. While cluster 1 (n = 52) and cluster 2 (n = 25) shared TPA-induced genes, cluster 3 (n = 45) was composed of TPA-repressed genes (see Additional file 3). We selected several differentially expressed genes (*Tgfb1, Tnf, Fosl1, Mmp2, Irf7, Hmgb2*, and *Hdac2*) and could confirm altered transcript levels by quantitative real-time PCR using cDNA from back skin of *wt* and *Rage*$^{-/-}$ mice 24 hours after TPA treatment (see Additional file 4). With regard to

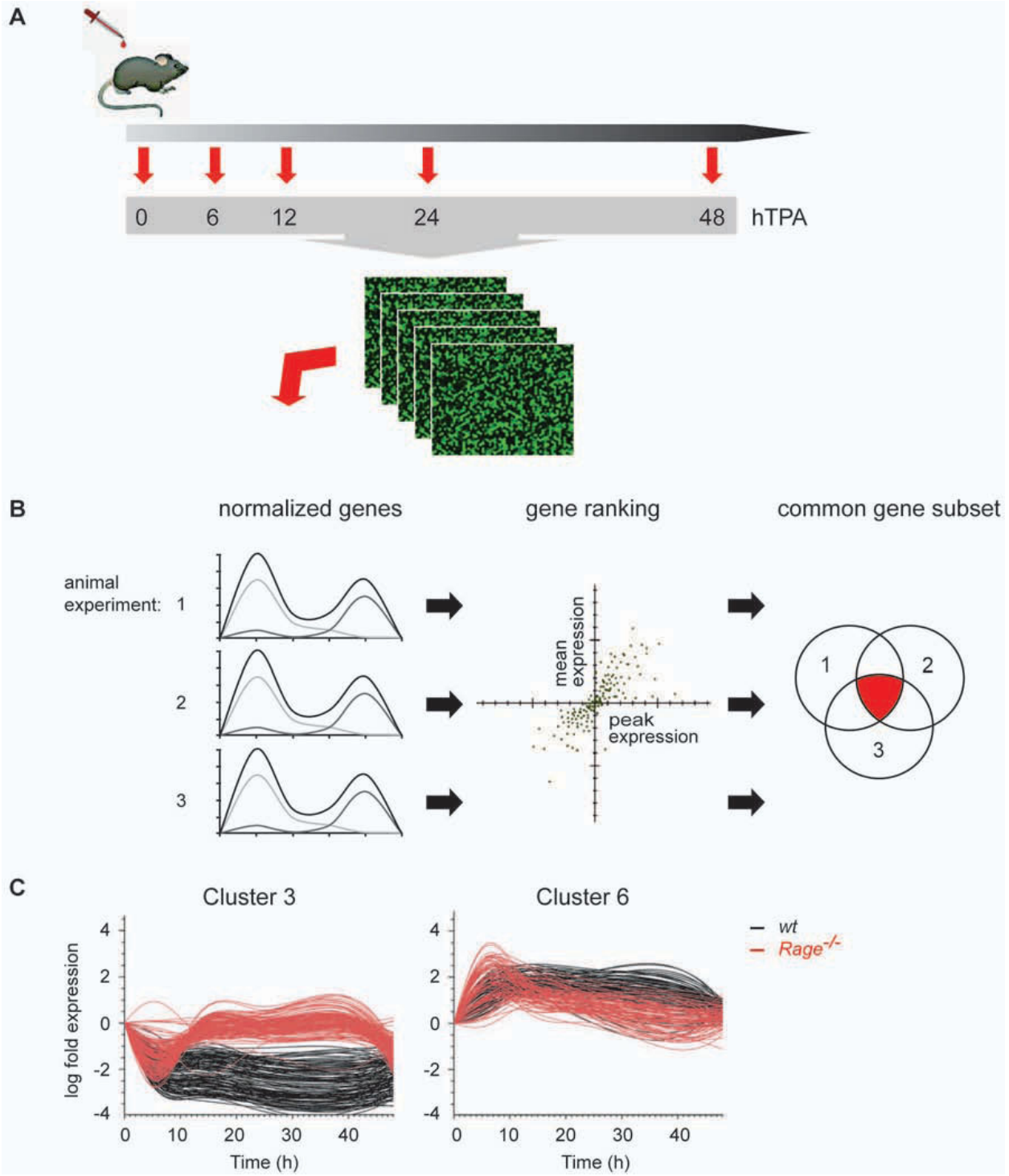


Figure 1 Global gene expression analysis of *wt* and *Rage*$^{-/-}$ skin following single TPA treatment. (A) Mice of respective phenotypes were treated once with TPA, and back skin was isolated 6, 12, 24, or 48 hours after stimulation. Non-treated and acetone-treated mice served as control (0). Global gene expression analysis of RNA samples was performed on whole mouse genome oligonucleotide microarrays (n = 3 for each genotype and time point). (B) Following quantile normalization, *wt* genes of each kinetic from three independent animal experiments (1, 2, and 3) were ranked according to high mean and peak expression separately to filter for TPA-responsive genes. A common subset of 341 genes was identified out of the top 1000 ranked genes within each kinetic. (C) K-means clustering revealed 6 clusters of which cluster 3 and 6 shared most genes (for cluster 1, 2, 4, and 5 see Additional file 2). Black lines represent transcript levels of TPA-regulated genes in *wt* samples and red lines represent the corresponding genes in *Rage*$^{-/-}$ samples.

their functional annotation, genes of cluster 1 were correlated with immune effector process, tissue remodeling and cell signaling, while genes of cluster 3 showed evident connection to histone and chromatin modifications as well as metabolic processes (data not shown).

Taken together, time-resolved global gene expression analysis of *wt* and *Rage*$^{-/-}$ skin upon TPA application disclosed expression patterns that subdivide the TPA-induced response into an initial Rage-independent phase and a second Rage-dependent maintenance of the established signal. Differentially expressed genes 24 hours after TPA stimulation revealed three gene clusters characterized by distinct functions.

Prediction of transcription factors implicated in the Rage-dependent gene regulatory network

In order to identify relevant transcription factors implicated in the regulation of Rage-dependent genes we performed an *in silico* promoter analysis. We used the probes that were differentially expressed between *wt* and *Rage*$^{-/-}$ mice at the time point 24 hours after TPA application and selected those that mapped unambiguously to one Entrezgene-ID and for which the promoter sequence was available (n = 97). These probes were clustered by their correlation distance within the samples from t = 24 hours into three clusters (see Additional file 5). We analyzed 2 kb upstream and downstream sequences of the annotated transcriptional start site and calculated the enrichment of transcription factor binding sites (TFBS) compared to all other available genes represented on the microarray by Fisher's exact tests. The analysis revealed several highly enriched TFBS for Specificity protein 1 and 4 (Sp1 and 4), Activator protein 2 (Ap2/Tcfap2), E2-promoter-binding factor (E2f), Myc-associated zinc-finger protein and Myc-associated zinc-finger protein-related protein (Mazr), Early growth response factor (Egr), CAC-binding protein (CAC-bp), v-Myc myelocytomatosis viral oncogene homolog (Myc), Nuclear receptor subfamily 2 group F members (Nr2f/COUP-TF), and Wilms tumor 1 homolog (Wt1) (Table 1). The enrichment tests were also

Table 1 *In silico* promoter analysis of differentially expressed genes 24 hours after TPA stimulation

Genes	BF Name	Fischertest P.Val	Corrected P.Val	With PWM cluster	Without PWM cluster
all	Sp1	5.33E-07	1.06E-04	94	3
	Sp1 isoform 1	5.33E-07	1.06E-04	94	3
	Sp4	1.72E-06	2.27E-04	83	14
	AP-2beta	1.24E-06	1.22E-03	77	20
	AP-2alpha	1.60E-05	1.27E-03	79	18
	AP-2gamma	2.13E-05	1.41E-03	79	18
	MAZR	6.03E-05	3.41E-03	74	23
	CAC-binding protein	1.32E-04	6.52E-03	81	16
	Egr-1	3.56E-04	1.56E-02	85	12
	Egr-3	4.38E-04	1.73E-02	78	19
	E2F	4.85E-04	1.75E-02	58	39
	c-Myc	7.31E-04	2.41E-02	67	30
	Egr-2	9.69E-04	2.94E-02	80	17
	COUP-TF1	1.06E-03	2.94E-02	89	8
	WT1	1.19E-03	2.94E-02	67	30
	WT1-isoform1	1.19E-03	2.94E-02	67	30
	COUP-TF2	1.45E-03	3.38E-02	48	49
Cluster 1	Sp4	2.01E-06	7.93E-04	40	2
	Sp1	6.13E-05	6.86E-03	42	0
	Sp1 isoform 1	6.13E-05	6.86E-03	42	0
	MAZR	6.93E-05	6.86E-03	36	6
	HNF-4alpha7	1.84E-04	1,46E-02	29	13
	CAC-binding protein	3.12E-04	2.06E-02	38	4
Cluster 2	MAZR	2.14E-04	7.64E-02	19	1
	WT1	5.79E-04	7.64E-02	18	2
	WT1-isoform1	5.79E-04	7.64E-02	18	2
Cluster 3	E2F	1.88E-05	7.45E-03	28	8
	E2F-1	8.50E-05	1.68E-02	28	8

applied on each of the three clusters separately to address the question, whether specific TFBS were significantly associated with differentially expressed genes in distinct clusters. While TFBS for Mazr were enriched in promoters of genes of at least 2 of 3 clusters, a significant correlation of TFBS for Sp1, Sp4, Hnf4, and CAC-bp were only found for promoters of genes in cluster 1 (Figure 2A). Similarly, significant enrichment of TFBS for Wt1 was restricted for gene promoters in cluster 2, and TFBS for E2f were limited to gene promoters in cluster 3 (Figure 2A).

In summary, the enrichment analyses highlighted the putative involvement of several transcription factors, such as E2f and Wt1, that were previously not associated with Rage signaling, and therefore, represent an exciting starting point for further investigation.

Impact of Rage on the regulation of E2f proteins in TPA-treated skin

To confirm our prediction on transcription factors implicated in Rage-dependent gene transcription, we investigated the expression and regulation of the E2f transcription factor that is well known to determine cellular responses to growth factors, stress and differentiation signals, as well as DNA damage [16]. E2f represents a family of eight transcription factors that are further subdivided into a group of potent transcriptional

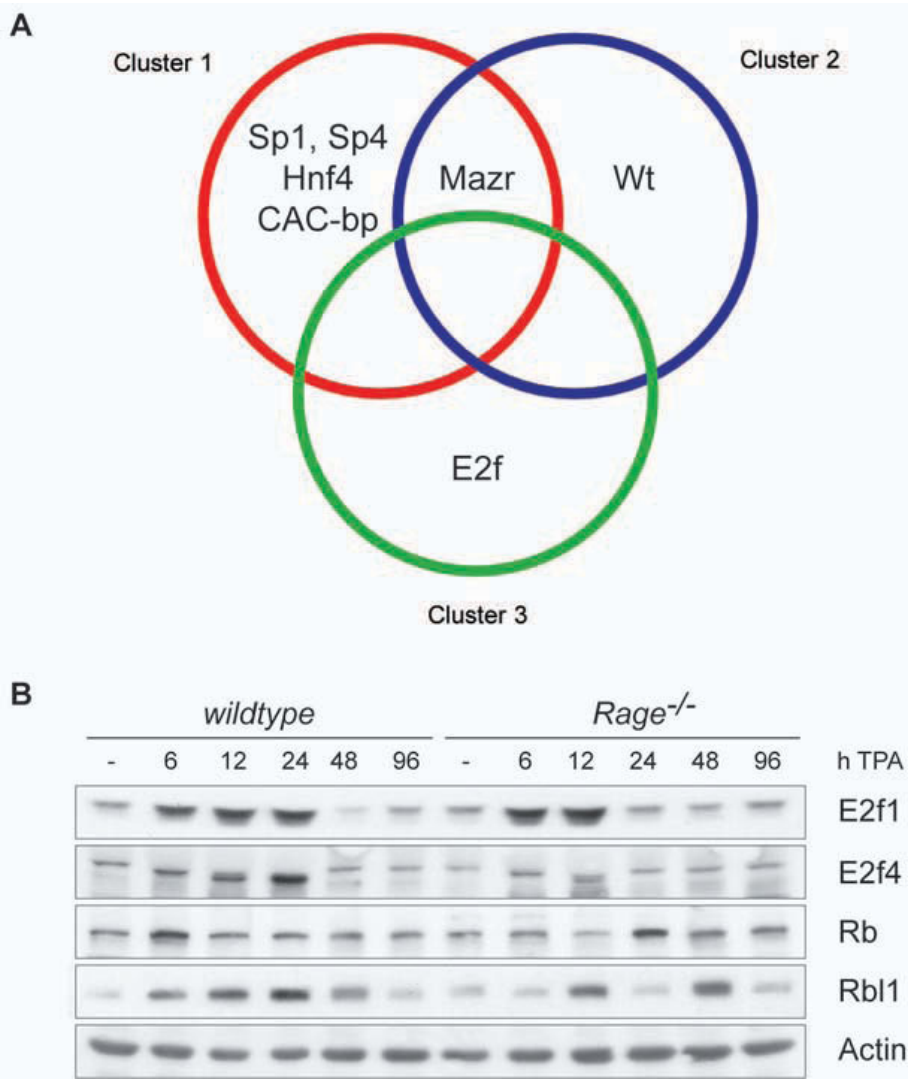


Figure 2 *In silico* promoter analysis of differentially expressed genes 24 hours after TPA treatment. (A) Annotated promoter sequences of 97 differentially expressed genes between *wt* and *Rage*$^{-/-}$ skin were screened for significant enrichment of transcription factor binding sites (TFBS). TFBS were grouped according to their enrichment within clusters of differentially expressed genes between *wt* and *Rage*$^{-/-}$ skin 24 hours after TPA treatment. (B) Protein expression of E2f1, E2f4, Rb, and Rbl1 was investigated by Western blot analysis with whole cell lysates of TPA- and control-treated back skin of *wt* and *Rage*$^{-/-}$ mice. Actin protein levels served as control for protein quality and quantity.

activators (E2f1-3a) and a group of preferential transcriptional repressors (E2f3b, E2f4-8) [17]. Analysis of our global gene expression data revealed no major alteration in transcript levels of most E2f family members between both genotypes, suggesting an impact of Rage on posttranslational regulation of E2f proteins. Interestingly, we found strong induction in protein levels of E2f1, a representative for the group of transcriptional activators, 6 and 12 hours after TPA application in skin lysates of both genotypes. However, enhanced protein levels 24 hours after TPA stimulation were only detected in *wt* lysates (Figure 2B). E2f4, a representative for the group of transcriptional repressors, gradually increased upon TPA treatment, showing a peak at 24 hours in *wt* skin, while no alterations in protein level were detected in *Rage*$^{-/-}$ samples throughout the kinetic (Figure 2B). Retinoblastoma (Rb) and retinoblastoma-like (Rbl) proteins regulate the activity of E2f transcription factors [16]. We found changes for Rb protein expression with highest levels 6 hours after stimulation in *wt* skin samples and 24 hours in *Rage*$^{-/-}$ skin samples (Figure 2B). Rbl1 protein expression was induced in *wt* and *Rage*$^{-/-}$ skin samples following TPA stimulation, but a concerted increase over time was only detected for *wt* animals. Together, these data support the conception that the Rb-E2f pathway is downstream of Rage signaling and critically contributes to altered gene transcription during TPA-induced skin inflammation.

Immunohistochemical staining was performed on tissue sections of *wt* and *Rage*$^{-/-}$ back skin upon single TPA treatment in order to investigate whether keratinocytes were the cellular origin of altered E2f protein levels. While slight staining for E2f1 protein was detected in keratinocytes of control-treated *wt* back skin, intense nuclear staining was found in kerationcytes upon TPA stimulation (Figure 3A-E). A similar staining pattern for E2f1 protein was observed in control-treated *Rage*$^{-/-}$ back skin and 6 hours after TPA administration, however, less intense staining was determined at later time points (Figure 3F-J). Immunohistochemical analysis of E2f4 protein revealed a strong but transient induction in the cytoplasm of keratinocytes of *wt* back skin 12 hours after TPA stimulation, followed by translocation of E2f4 protein into the nucleus by 24 hours after TPA application (Figure 3K-O). Again, an obvious change in E2f4 protein expression was detectable in *Rage*$^{-/-}$ back skin (Figure 3P-T). These data are in clear accordance with our immunoblot data and demonstrate a direct correlation between RAGE signaling and E2f-dependent gene expression in epidermal keratinocytes upon TPA-induced skin inflammation. Finally, we also determined Rb, Rbl1, and Rbl2 protein levels by immunochistochemistry. While no major alteration in Rbl2 protein levels was observed following TPA stimulation or between both genotypes (see Additional file 6), we found a stronger staining for Rb and Rbl1 proteins in keratinocytes of TPA-treated *wt* compared to *Rage*$^{-/-}$ back skin (Figure 4).

Discussion

The aim of our study was to highlight the molecular mechanism how Rage signaling contributes to the dynamic long-term gene regulatory response under physiological and pathological conditions of inflammation. We selected the model of TPA-induced inflammation on mouse back skin since it allows a highly reproducible and temporal analysis of altered gene expression during acute phase inflammation, including the initiation as well as the resolution phase. Furthermore, we recently provided experimental evidence that *Rage*$^{-/-}$ mice are defective in the establishment and maintenance of dermal inflammation upon TPA stimulation accompanied by impaired tumor formation in a chemically induced skin tumor model [3]. Finally, it is worthwhile to note that Rage and its ligands are expressed or induced in numerous cell types, including keratinocytes, immune cells and endothelial cells [5]. With regard to this complex autocrine and paracrine signaling, *in vitro* approaches to identify the role of Rage signaling on gene regulatory networks under pathological conditions are almost impossible.

Global profiling of gene expression kinetics with samples from TPA-treated *wt* mice revealed a comprehensive list of differentially expressed genes that were strongly induced or repressed within 6 hours and maintained altered transcript levels for at least 24 hours. Intriguingly, most of these genes exhibited similar changes in expression at early time points, but rapidly returned to basal levels in the absence of Rage, providing experimental evidence that Rage is not necessarily required to initiate gene regulation in TPA-induced skin inflammation. However, Rage is absolutely required to maintain altered expression of genes implicated in immune effector processes, cell signaling, as well as histone and chromatin organization, and thereby sustains the tissue response.

Our data fit into the model that engagement of Rage converts a transient cellular stimulation into sustained cellular dysfunction. An important driver of this conversion is the long-term activation of pro-inflammatory transcription factors, especially NF-κB, which represents a key feature of most intracellular signaling pathways that have been described downstream of Rage stimulation [6,8,18]. However, we hypothesized that in addition to NF-κB other transcription factors contribute to Rage-dependent modulation of the gene regulatory network. Obviously, most genes with a significant difference in

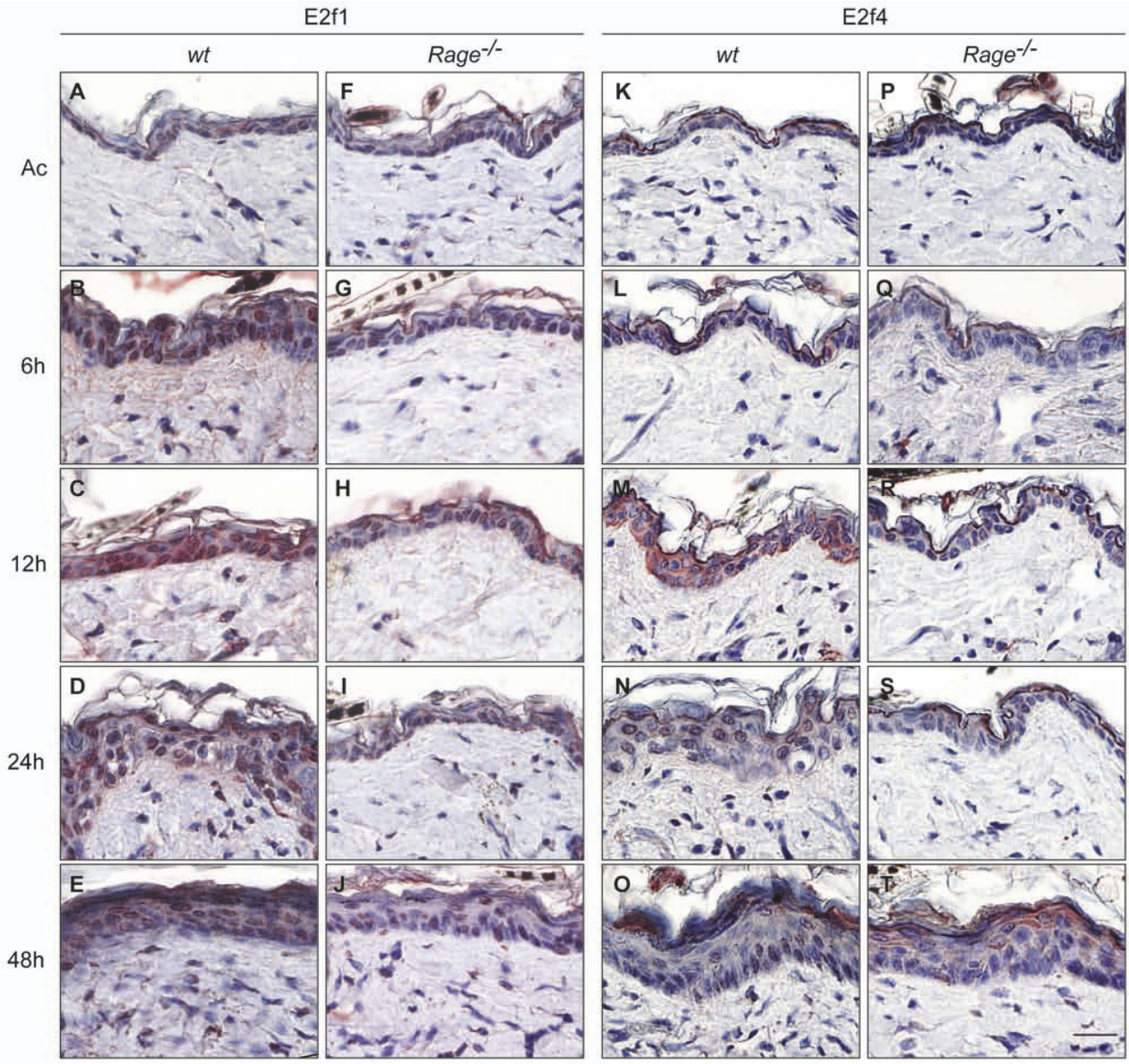


Figure 3 E2f1 and E2f4 protein expression in skin following a single TPA stimulus. Tissue sections of acetone- (Ac) or TPA-treated (6, 12, 24, and 48 hours) back skin from *wt* and *Rage*$^{-/-}$ mice were analyzed by immunohistochemical staining using E2f1- and E2f4-specific antibodies. Representative images of at least 2 animals of each genotype and time point are shown with red staining for E2f1 or E2f4 and counterstaining with hematoxylin. Scale bar = 25 μm.

transcript levels between *wt* and *Rage*$^{-/-}$ back skin were detectable 24 hours after TPA administration due to the effect of Rage on the temporal expression pattern. Computational analysis of TFBS in proximal promoter regions of differentially expressed genes allowed prediction of specific transcription factors that act downstream of Rage signaling during TPA-induced inflammation. In line with our data, a couple of recent studies describe a direct link between Rage signaling and Egr-1 activation in endothelial cells and liver cells [19-22]. In this context, Egr-1 was found in the physiological response to hypoxia and stress signals by direct up-regulation of inflammatory and pro-thrombotic genes. Moreover, a systems biology approach with human monocytes treated with the immunomodulatory peptide LL-37 revealed an involvement of Ap2, Sp1, E2f and Egr in gene regulation during conditions based on innate immunity [23]. Our data also predict that these transcription factors seem to be co-activated by various conditions of inflammation and synergize with well-known pro-inflammatory transcription factors such as NF-κB and AP-1 in a Rage-dependent manner. It is worth to note

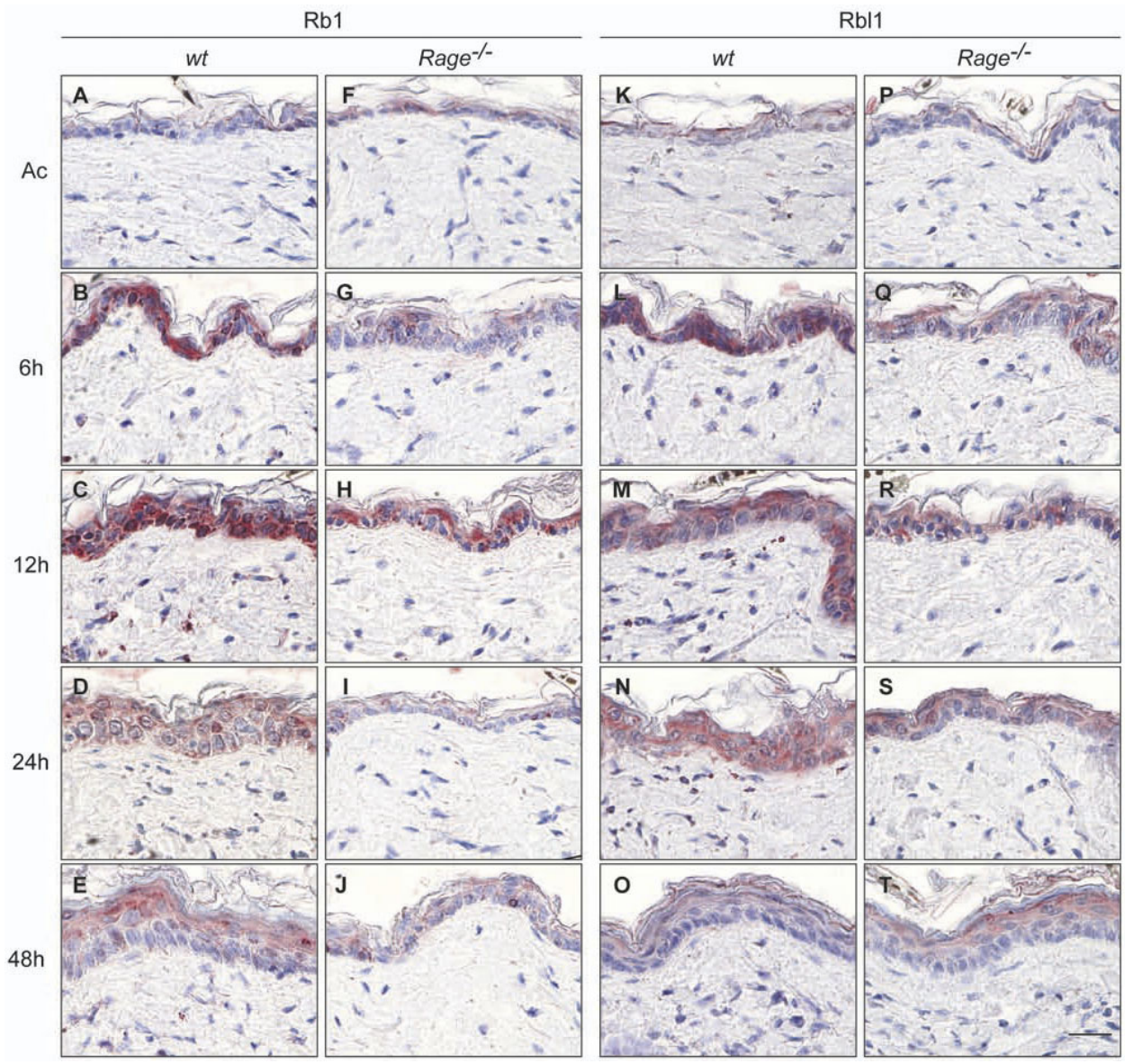


Figure 4 Rb and Rbl1 protein expression in skin following a single TPA stimulus. Tissue sections of acetone- (Ac) or TPA-treated (6, 12, 24, and 48 hours) back skin from *wt* and *Rage*$^{-/-}$ mice were analyzed by immunohistochemical staining using Rb- and Rbl1-specific antibodies. Representative images of at least 2 animals of each genotype and time point are shown with red staining for Rb or Rbl1 and counterstaining with hematoxylin. Scale bar = 25 μm.

that TFBS for Sp1, Ap2, and Egr are also present in the promoter of *Rage* [24,25], suggesting a positive feedback loop by up-regulation of the receptor, which ensures maintenance and amplification of cellular activation in settings where ligands of Rage accumulate. A similar scenario has been described for NF-κB, which also represents a target of Rage signaling and activator of Rage expression [5,6].

Interestingly, our analysis revealed a significant enrichment of TFBS for the transcription factor E2f. There are a number of findings demonstrating diverse transcriptional regulation of E2f-responsive genes, suggesting that expression of these genes is regulated by different sets of Rb-E2f protein complexes [16]. However, it is currently uncertain how individual E2f members recognize a particular E2f-binding site during cell cycle progression or differentiation. One possibility is that the DNA-binding specificity of E2f members is influenced by other transcriptional regulatory factors, such as Sp1, that bind to sites contiguous to the E2f-binding site [26,27]. We found TPA-induced protein levels for E2f1 and E2f4 in epidermal keratinocytes of *wt*

mice. At the same time, we observed a prominent up-regulation of Rb1 and Rbl1, suggesting the formation of Rb-E2f protein complexes. In contrast to *wt* mice, induction of E2f and Rb family proteins was impaired or only transient in keratinocytes of TPA-treated *Rage*$^{-/-}$ back skin. Rb family proteins associate with a wide range of chromatin remodeling proteins forming transcriptional repressor complexes [28]. Thus, the existence of Rb-E2f complexes in keratinocytes after TPA stimulation could explain the enrichment of TFBS for E2f in the gene set characterized by strong and sustained repression. In addition to control of gene expression and binding to inhibitory Rb proteins, the activity of E2F proteins is tightly regulated by post-translational modification and regulation of protein turnover [29]. As an example, free E2F1 and E2F4 proteins are unstable due to ubiquitination and proteasomal degradation. Numerous cellular proteins have been described to regulate E2F protein ubiquitination, such as the CK1-MDM2 complex [30,31], ARF proteins (p14ARF in human and p19ARF in the mouse; [32]), Set9 and LSD1 [33]. However, our analysis did not reveal major changes in expression of any of these regulators upon TPA application or between RAGE-deficient mice and controls, and a functional link between RAGE signaling with one of these proteins has not been documented to the best of our knowledge. Ivanova and colleagues reported that in differentiating keratinocytes calcium-induced protein kinase C (PKC) activation reduces E2F1 protein level, which requires activation of novel PKC isoforms by the MAP kinase p38 [34]. Again E2F1 down-regulation in differentiating keratinocytes involves its ubiquitination and proteosomal degradation subsequent to CRM1-dependent nuclear export and degradation of E2F1 during differentiation [35]. Indeed, we observed strong cytoplasmic staining for E2F1 protein 12 hours after TPA treatment in keratinocytes of *Rage*$^{-/-}$ mice and *wt* controls. However, in contrast to *wt* controls, which show obvious nuclear staining for E2F1 until 48 hours after treatment, nuclear staining in keratinocytes of *Rage*$^{-/-}$ mice was hardly visible at any time point, suggesting that RAGE signaling might regulate nuclear-cytoplasmic shuttling of E2F proteins.

Finally, our data predict that the Rb-E2f pathway and its target genes not only act downstream of Rage signaling, but also might be pivotal for the process of skin inflammation upon TPA treatment. Indeed, CDK activity, which is up-stream of Rb-E2f, was recently correlated with roles in inflammatory cell differentiation, adhesion and recruitment as well as cytokine production and inflammatory signaling [36]. Intriguingly, CDK inhibitor drugs that are well-known to impair cell cycle progression in tumor cells have emerged recently as potential anti-inflammatory, pharmacological agents by influencing the resolution of inflammation [36,37].

Conclusions

In summary, our approach to combine gene expression profiling with computational analysis did not only highlight the topology of Rage-dependent gene regulation in skin inflammation, but also allowed the prediction of novel transcription factors downstream of Rage signaling (Figure 5). A major challenge in the future will be the integration of known and newly identified transcription factors in a common model of Rage-dependent signaling network and to predict a dynamic program of inflammation in settings of physiological as well as pathological conditions.

Methods

Animal work and sample preparation

Rage$^{-/-}$ animals were described previously [38], and *wt* controls were obtained from Charles River Laboratories. Mice were housed and treated with TPA as described previously [3]. In short, 10 nmol TPA/100 μl Acetone was applied to the shaved dorsal back skin and mice were sacrificed at indicated time points. Mice receiving acetone or no treatment served as controls. The procedures for performing animal experiments were in accordance with the principles and guidelines of the 'Arbeitsgemeinschaft der Tierschutzbeauftragten in Baden-Württemberg' and were approved by the 'Regierungspräsidium Karlsruhe', Germany (AZ 129/02).

Skin necropsies for RNA or protein preparation were immediately frozen in liquid nitrogen after isolation. For histological analysis, tissues were fixed with 4% paraformaldehyde (PFA) in PBS pH 7.4, paraffin embedded, and cut into 6 μm sections as described previously [3]. Tissue sections were stained with hematoxylin-eosin and were examined by several experienced experimenters.

RNA preparation

Total RNA extraction from mouse back skin of untreated, 24 hours acetone- and 6, 12, 24, 48 hours TPA-treated *wt* and *Rage*$^{-/-}$ mice was performed according to the manufacturer's instructions using peqGOLD RNAPure™ Reagent (Peq Lab, Erlangen, Germany). For RNA integrity and degradation analysis, the 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA) with RNA 6000 Nano LabChip Kit was used according to the manufacturer's instructions. Only RNA preparations with a RNA Integrity Number (RIN) of at least seven were used for microarray analysis.

Microarray analysis

Global gene expression analysis was performed on 4x44K whole mouse genome one-color 60-mer oligonucleotide microarrays (Agilent Technologies) containing

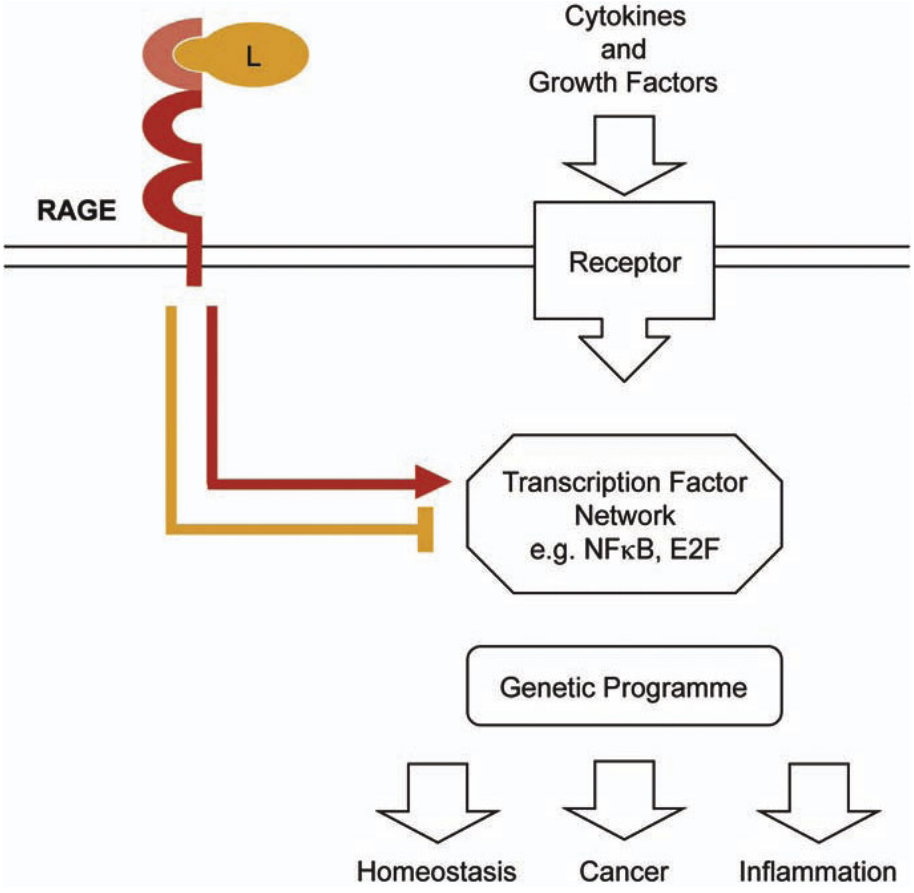


Figure 5 Model of physiological and pathological functions of RAGE signaling. Combination of gene expression profiling and computational approaches revealed that RAGE is not necessary for the initial response after stimulation, but absolute required for sustaining altered gene transcription. This is mainly due to the impact of RAGE-mediated signaling on expression and activity of transcription factors such as E2F. Thereby, RAGE modulates the kinetic of transcript levels of genes implicated in tissue homeostasis, inflammation, and cancer.

41,174 unique probes. For amplification and single-color labeling of 1 μg RNA, the Low RNA Input Linear Amplification Kit and the One-Color RNA Spike-in Kit (Agilent Technologies, Santa Clara, CA) were used according to the manufacturer's protocol. Upon hybridization, microarray read out was accomplished in the Agilent Scanner G25505B (Agilent Technologies, Santa Clara, CA) with 5 μm resolution and automatically adjusting PMT voltages according to manufacturer's specification. Data processing was performed using Feature Extraction FE V9.5 Image Analysis software (Agilent Technologies, Santa Clara, CA) as recommended by the manufacturer.

Quantitative real-time polymerase chain reaction analysis

Quantitative real-time polymerase chain reaction analysis was performed as described previously [3]. Primers used for RQ-PCR analysis are listed in Additional file 7. Target gene cycle of threshold values were normalized to the corresponding cycle of threshold values of using the change in cycle of threshold method.

Statistical analysis

Array data were normalized by the quantile method [39], in combination with the "normexp" background correction implemented in Limma [40], and log2-transformed. Differentially expressed genes were identified by applying a linear model with the factors "time point" and "genotype", and subsequent empirical Bayesian correction [40]. For each time point t, the following contrast was calculated: $(I_{r,t} - I_{r\text{-}0}) - (I_{w,t} - w_{w,0})$, where I is the logarithm of the vector of intensities, indices r and w refer to *Rage*$^{-/-}$ and *wt* mice, respectively, and 0 is the control condition ($t = 0$). P-values from the F test of the linear model were adjusted for multiple testing by the method of Benjamini and Hochberg [41]. All adjusted P-values < 0.05 were considered significant. All calculations were carried out in R version 2.6.2 http://www.R-project.org and Limma

version 2.12.0. A list of differentially expressed genes at 24 hours after TPA stimulation is given in Additional file 3.

In order to further assess the dynamic response of the gene regulation upon TPA stimulation, we ranked the gene expression kinetics according to the peak and mean fold expression of each genes within the experimental time window of 48 hours [15]. Fold expression of each gene was calculated with respect to the control condition of *wt* and *Rage*$^{-/-}$ animals. Next, a rank score s was defined for every gene as $s = \sqrt{FE_t^2 + FE_p^2}$, where $FE_i = \langle g_i(t) \rangle_T$ and, FE_p denote the time-averaged mean and peak gene fold expression (FE) of the gene's time series, respectively, normalized to the maximal peak and maximal mean fold expression of all measured genes. Gene ranking was performed separately for the three biological replicates of the TPA stimulation. Taking the 1,000 genes having the highest rank score s in each replicate, a set of 341 genes common to all rank lists was identified for further analysis (Additional file 1). Taking the top 500, 2,000 or 3,000 genes did not change the quality of the results. Subsequent k-means clustering of these 341 *wt* expression profiles was performed and expression levels of the same genes in *Rage*$^{-/-}$ samples were superimposed. Pathway analysis was accomplished by using web-based DAVID [42,43].

In silico promoter analysis

All available promoter sequences of murine genes represented on the whole mouse genome microarray were extracted from NCBI Entrezgene database ([44]; download: June 19th, 2008). Putative transcription factor binding sites (TFBS) were scanned within 2 kb upstream and downstream of the annotated transcriptional start site utilizing a position-weight matrix (PWM) scan as implemented in cureos package v0.3 ([45]; http://www.bepress.com/sagmb/vol2/iss1/art7) and described in Westermann et al. [46] for R open-source software http://www.R-project.org. PWMs were taken from the TRANSFAC database ([47]; release: January 12th, 2008). *P*-values for each PWM were obtained by comparing their scores to those of 1,000 random 4 kb sequence permutations. A general *P*-value cut-off of $p < 0.1$ was set as a reasonable compromise between false-positives and false-negatives. The genes that were differentially expressed between *wt* and *Rage*$^{-/-}$ mice at the time point 24 hours after TPA application were divided into three sub-clusters with different gene expression profiles by unsupervised hierarchical clustering within samples at the time point t = 24 hours (complete linkage, the distance was calculated by one minus the Pearson correlation coefficient). Fisher's exact tests were performed to determine enrichments of PWM hits (counting genes with ≥ 1 significant score as a hit) for each cluster as compared to all other genes represented on the microarray. All *P*-values were Benjamini-Hochberg-corrected for multiple testing.

Western Blot analysis

Western Blot analyses were performed with whole cell lysates from mouse back skin with antibodies listed in Additional file 8 according to the manufacturer's instructions. Whole cell lysates were prepared with RIPA buffer (50 mM Tris-HCL pH 8, 150 mM NaCl, 0.1% SDS, 0.5% deoxylacid Na^+-salt, 1% NP-40).

Immunohistochemistry analysis

Immunohistochemistry staining was performed on back skin sections from *wt* and *Rage*$^{-/-}$ mice with the Immunodetection Kit (Vector Laboratories; Burlingame, CA) according to the manufacturer's instructions. Primary and secondary antibodies used are listed in Additional file 8.

Additional material

Additional file 1: Table with 341 TPA-responsive genes in back skin of *wt mice*.

Additional file 2: K-means clustering of TPA-responsive genes. K-means clustering of common TPA-responsive genes in the kinetics of three independent experiments with *wt* animals revealed 6 clusters. Cluster 1 (n = 5), cluster 2 (n = 45), cluster 3 (n = 125), cluster 4 (n = 71), cluster 5 (n = 11), and cluster 6 (n = 84). Black lines represent transcript levels of genes in *wt* skin samples. Red lines represent transcript levels in *Rage*$^{-/-}$ skin samples.

Additional file 3: Table of 122 genes differentially expressed 24 hours after TPA stimulation.

Additional file 4: Quantitative real-time PCR of differentially expressed genes 24 hours after TPA treatment. Relative transcript levels of differentially expressed genes were determined by quantitative real-time PCR with cDNA derived from *wt* and *Rage*$^{-/-}$ back skin 24 hours upon TPA treatment. Transcript levels for genes of interest were determined in triplicates with *wt* and *Rage*$^{-/-}$ cDNA samples and normalized to *Hprt* transcript levels. Next, expression values of genes of interest derived from *wt* cDNA were set to one and bars represent relative transcript levels for *Rage*$^{-/-}$ cDNA samples. Similar data were obtained for two independent biological replicates (data not shown).

Additional file 5: Cluster dendrogram of genes differentially expressed at t = 24 hours. Clustering was done only over samples from t = 24 hours via Person correlation distance, complete linkage algorithm. Three clusters were defined from the dendrogram.

Additional file 6: Rbl2 protein expression in skin following a single TPA stimulus. Tissue sections of acetone- (Ac) or TPA-treated (6, 12, 24, and 48 hours) back skin from *wt* and *Rage*$^{-/-}$ mice were analyzed by immunohistochemical staining using Rbl2-specific antibodies. Representative images of at least 2 animals of each genotype and time point are shown with red staining for Rbl2 and counterstaining with hematoxylin. Scale bar = 25 μm.

Additional file 7: Table of primer sequences used for quantitative real-time PCR analysis.

Additional file 8: Table of primary antibodies used for Western Blot (WB) and immunohistochemistry analysis (IHC).

Acknowledgements
We gratefully acknowledge Angelika Krischke and Ingeborg Vogt for excellent technical assistance and Axel Szabowski for helpful discussion. Our work was supported by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (to P.A., J.H., T. B., and R.K.), the Cooperation in Cancer Research of the Deutsche Krebsforschungszentrum and Israeli's Ministry of Science, Culture and Sport (to P.A. and J.H.), the Excellence Initiative of the German Federal and State Governments (to H.B.), the Federal Ministry of Education and Research through the National Genome Research Network (grant 01GS0883 to B.B. and R.E.), and the Dietmar Hopp Foundation (to J.H.).

List of abbreviations
AGE: advanced glycation end products; AP1: Activator protein 1; AP2: Activator protein 2; CAC-bp: CAC-binding protein; E2f: E2-promoter-binding factor; Egr: Early growth response factor; Hmgb1: High mobility group box-1; JAK: Janus kinase; MAP kinases: mitogen-activated protein kinases; Maz: Myc-associated zinc-finger protein; Mazr: Myc-associated zinc-finger protein-related protein; Myc: v-myc meylocytomatosis viral oncogene homolog; NF-κB: Nuclear factor kappa B; Nr2f: nuclear receptor subfamily 2 group F members, PI3K: Phosphoinositide-3-kinase; PKC: protein kinase C; Rage: receptor for advanced glycation end products; Rb: Retinoblastoma protein; Rbl1: Retinoblastoma-like protein 1; Rbl2: Retinoblastoma-like protein 2; Sp1: Specificity protein 1; Sp4: Specificity protein 4; STAT: Signal transducer and activator of transcription; Tcfap2: transcription factor AP-2, alpha; TFBS: transcription factor binding sites; TPA: 12-O-tetradecanoylphorbol-13-acetate; Wt: wild-type; Wt1: Wilms tumor 1 homolog.

Author details
[1]Signal Transduction and Growth Control, German Cancer Research Center (DKFZ), DKFZ-ZMBH Alliance, Heidelberg, Germany. [2]Theoretical Bioinformatics, German Cancer Research Center, Heidelberg, Germany. [3]Freiburg Institute for Advanced Studies - FRIAS School of Life Sciences - LIFENET Albert-Ludwigs-University Freiburg, Germany. [4]Center for Biosystems Analysis, Albert-Ludwigs-University Freiburg, Germany. [5]Institute of Pharmacy and Molecular Biology and Bioquant Center, University of Heidelberg, Germany. [6]Department of Dermatology, University Hospital Heidelberg, Germany. [7]Department of Medicine I and Clinical Chemistry, University Hospital Heidelberg, Germany. [8]Experimental Head and Neck Oncology, Department of Otolaryngology, Head and Neck Surgery, University Hospital Heidelberg, Germany.

Authors' contributions
AR, CG, PA and JH design and analysis of experimental research; AR, RM, JN and CG performed experimental research; TB, BB, HB, RK and RE designed and performed computational and statistical analysis; AB and PN provision of animal model system and analytic tools; AR and JH wrote the paper; TB, BB, HB, RM, JN, RK, CG, AB, PN, RE and PA critical review and editing of the manuscript.
None of the authors had any personal or financial conflicts of interest.

Received: 21 April 2010 Accepted: 5 October 2010
Published: 5 October 2010

References
1. Coussens LM, Werb Z: **Inflammation and cancer.** *Nature* 2002, **420**:860-867.
2. Balkwill F, Mantovani A: **Inflammation and cancer: back to Virchow?** *Lancet* 2001, **357**:539-545.
3. Gebhardt C, Riehl A, Durchdewald M, Nemeth J, Furstenberger G, Muller-Decker K, Enk A, Arnold B, Bierhaus A, Nawroth PP, *et al*: **RAGE signaling sustains inflammation and promotes tumor development.** *J Exp Med* 2008, **205**:275-285.
4. Turovskaya O, Foell D, Sinha P, Vogl T, Newlin R, Nayak J, Nguyen M, Olsson A, Nawroth PP, Bierhaus A, *et al*: **RAGE, carboxylated glycans and S100A8/A9 play essential roles in colitis-associated carcinogenesis.** *Carcinogenesis* 2008, **29**:2035-2043.
5. Riehl A, Nemeth J, Angel P, Hess J: **The receptor RAGE: Bridging inflammation and cancer.** *Cell Commun Signal* 2009, **7**:12.
6. Bierhaus A, Nawroth PP: **Multiple levels of regulation determine the role of the receptor for AGE (RAGE) as common soil in inflammation, immune responses and diabetes mellitus and its complications.** *Diabetologia* 2009, **52**:2251-2263.
7. Stern D, Yan SD, Yan SF, Schmidt AM: **Receptor for advanced glycation endproducts: a multiligand receptor magnifying cell stress in diverse pathologic settings.** *Adv Drug Deliv Rev* 2002, **54**:1615-1625.
8. Clynes R, Moser B, Yan SF, Ramasamy R, Herold K, Schmidt AM: **Receptor for AGE (RAGE): weaving tangled webs within the inflammatory response.** *Curr Mol Med* 2007, **7**:743-751.
9. Taguchi A, Blood DC, del Toro G, Canet A, Lee DC, Qu W, Tanji N, Lu Y, Lalla E, Fu C, *et al*: **Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases.** *Nature* 2000, **405**:354-360.
10. Kang R, Tang D, Schapiro NE, Livesey KM, Farkas A, Loughran P, Bierhaus A, Lotze MT, Zeh HJ: **The receptor for advanced glycation end products (RAGE) sustains autophagy and limits apoptosis, promoting pancreatic tumor cell survival.** *Cell Death Differ* 2009.
11. Logsdon CD, Fuentes MK, Huang EH, Arumugam T: **RAGE and RAGE ligands in cancer.** *Curr Mol Med* 2007, **7**:777-789.
12. Stern DM, Yan SD, Yan SF, Schmidt AM: **Receptor for advanced glycation endproducts (RAGE) and the complications of diabetes.** *Ageing Res Rev* 2002, **1**:1-15.
13. Lin L: **RAGE on the Toll Road?** *Cell Mol Immunol* 2006, **3**:351-358.
14. van Beijnum JR, Buurman WA, Griffioen AW: **Convergence and amplification of toll-like receptor (TLR) and receptor for advanced glycation end products (RAGE) signaling pathways via high mobility group B1 (HMGB1).** *Angiogenesis* 2008, **11**:91-99.
15. Busch H, Camacho-Trullio D, Rogon Z, Breuhahn K, Angel P, Eils R, Szabowski A: **Gene network dynamics controlling keratinocyte migration.** *Mol Syst Biol* 2008, **4**:199.
16. DeGregori J, Johnson DG: **Distinct and Overlapping Roles for E2F Family Members in Transcription, Proliferation and Apoptosis.** *Curr Mol Med* 2006, **6**:739-748.
17. Polager S, Ginsberg D: **p53 and E2f: partners in life and death.** *Nat Rev Cancer* 2009, **9**:738-748.
18. Bierhaus A, Schiekofer S, Schwaninger M, Andrassy M, Humpert PM, Chen J, Hong M, Luther T, Henle T, Kloting I, *et al*: **Diabetes-associated sustained activation of the transcription factor nuclear factor-kappaB.** *Diabetes* 2001, **50**:2792-2808.
19. Lv B, Wang H, Tang Y, Fan Z, Xiao X, Chen F: **High-mobility group box 1 protein induces tissue factor expression in vascular endothelial cells via activation of NF-kappaB and Egr-1.** *Thromb Haemost* 2009, **102**:352-359.
20. Li M, Shang DS, Zhao WD, Tian L, Li B, Fang WG, Zhu L, Man SM, Chen YH: **Amyloid beta interaction with receptor for advanced glycation end products up-regulates brain endothelial CCR5 expression and promotes T cells crossing the blood-brain barrier.** *J Immunol* 2009, **182**:5778-5788.
21. Zeng S, Dun H, Ippagunta N, Rosario R, Zhang QY, Lefkowitch J, Yan SF, Schmidt AM, Emond JC: **Receptor for advanced glycation end product (RAGE)-dependent modulation of early growth response-1 in hepatic ischemia/reperfusion injury.** *J Hepatol* 2009, **50**:929-936.
22. Chang JS, Wendt T, Qu W, Kong L, Zou YS, Schmidt AM, Yan SF: **Oxygen deprivation triggers upregulation of early growth response-1 by the receptor for advanced glycation end products.** *Circ Res* 2008, **102**:905-913.
23. Mookherjee N, Hamill P, Gardy J, Blimkie D, Falsafi R, Chikatamarla A, Arenillas DJ, Doria S, Kollmann TR, Hancock RE: **Systems biology evaluation of immune responses induced by human host defence peptide LL-37 in mononuclear cells.** *Mol Biosyst* 2009, **5**:483-496.
24. Li J, Schmidt AM: **Characterization and functional analysis of the promoter of RAGE, the receptor for advanced glycation end products.** *J Biol Chem* 1997, **272**:16498-16506.
25. Reynolds PR, Kasteler SD, Cosio MG, Sturrock A, Huecksteadt T, Hoidal JR: **RAGE: developmental expression and positive feedback regulation by Egr-1 during cigarette smoke exposure in pulmonary epithelial cells.** *Am J Physiol Lung Cell Mol Physiol* 2008, **294**:L1094-1101.
26. Dimova DK, Dyson NJ: **The E2F transcriptional network: old acquaintances with new faces.** *Oncogene* 2005, **24**:2810-2826.
27. Polager S, Ginsberg D: **E2F - at the crossroads of life and death.** *Trends Cell Biol* 2008, **18**:528-535.
28. Macaluso M, Montanari M, Giordano A: **Rb family proteins as modulators of gene expression and new aspects regarding the interaction with chromatin remodeling enzymes.** *Oncogene* 2006, **25**:5263-5267.

29. Stevaux O, Dyson NJ: **A revised picture of the E2F transcriptional network and RB function.** *Curr Opin Cell Biol* 2002, **14**:684-691.
30. Huart AS, MacLaine NJ, Meek DW, Hupp TR: **CK1alpha plays a central role in mediating MDM2 control of p53 and E2F-1 protein stability.** *J Biol Chem* 2009, **284**:32384-32394.
31. Zhang Z, Wang H, Li M, Rayburn ER, Agrawal S, Zhang R: **Stabilization of E2F1 protein by MDM2 through the E2F1 ubiquitination pathway.** *Oncogene* 2005, **24**:7238-7247.
32. Martelli F, Hamilton T, Silver DP, Sharpless NE, Bardeesy N, Rokas M, DePinho RA, Livingston DM, Grossman SR: **p19ARF targets certain E2F species for degradation.** *Proc Natl Acad Sci USA* 2001, **98**:4455-4460.
33. Kontaki H, Talianidis I: **Lysine methylation regulates E2F1-induced cell death.** *Mol Cell* 2010, **39**:152-160.
34. Ivanova IA, D'Souza SJ, Dagnino L: **E2F1 stability is regulated by a novel-PKC/p38beta MAP kinase signaling pathway during keratinocyte differentiation.** *Oncogene* 2006, **25**:430-437.
35. Ivanova IA, Dagnino L: **Activation of p38- and CRM1-dependent nuclear export promotes E2F1 degradation during keratinocyte differentiation.** *Oncogene* 2007, **26**:1147-1154.
36. Leitch AE, Haslett C, Rossi AG: **Cyclin-dependent kinase inhibitor drugs as potential novel anti-inflammatory and pro-resolution agents.** *Br J Pharmacol* 2009, **158**:1004-1016.
37. Hallett JM, Leitch AE, Riley NA, Duffin R, Haslett C, Rossi AG: **Novel pharmacological strategies for driving inflammatory cell apoptosis and enhancing the resolution of inflammation.** *Trends Pharmacol Sci* 2008, **29**:250-257.
38. Constien R, Forde A, Liliensiek B, Grone HJ, Nawroth P, Hammerling G, Arnold B: **Characterization of a novel EGFP reporter mouse to monitor Cre recombination as demonstrated by a Tie2 Cre mouse line.** *Genesis* 2001, **30**:36-44.
39. Smyth GK, Speed T: **Normalization of cDNA microarray data.** *Methods* 2003, **31**:265-273.
40. Smyth GK: *Limma: linear models for microarray data* New York: Springer 2005.
41. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Royal Stat Soc B* 1995, **57**:289-300.
42. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**:P3.
43. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44-57.
44. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2005, **33**:D54-58.
45. Rahmann S, Muller T, Vingron M: **On the power of profiles for transcription factor binding site detection.** *Stat Appl Genet Mol Biol* 2003, **2**:Article7.
46. Westermann F, Muth D, Benner A, Bauer T, Henrich KO, Oberthuer A, Brors B, Beissbarth T, Vandesompele J, Pattyn F, *et al*: **Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas.** *Genome Biol* 2008, **9**:R150.
47. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, *et al*: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**:D108-110.

doi:10.1186/1471-2164-11-537
Cite this article as: Riehl *et al.*: Identification of the Rage-dependent gene regulatory network in a mouse model of skin inflammation. *BMC Genomics* 2010 **11**:537.

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 16 2011, pages 2239–2247
doi:10.1093/bioinformatics/btr366

Gene expression

Advance Access publication June 20, 2011

RIP: the regulatory interaction predictor—a machine learning-based approach for predicting target genes of transcription factors

Tobias Bauer[1,2], Roland Eils[1,2,*] and Rainer König[1,2,*]

[1]Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), INF 280, 69120 Heidelberg and [2]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, INF 267, University of Heidelberg, 69120 Heidelberg, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Understanding transcriptional gene regulation is essential for studying cellular systems. Identifying genome-wide targets of transcription factors (TFs) provides the basis to discover the involvement of TFs and TF cooperativeness in cellular systems and pathogenesis.

Results: We present the regulatory interaction predictor (RIP), a machine learning approach that inferred 73 923 regulatory interactions (RIs) for 301 human TFs and 11 263 target genes with considerably good quality and 4516 RIs with very high quality. The inference of RIs is independent of any specific condition. Our approach employs support vector machines (SVMs) trained on a set of experimentally proven RIs from a public repository (TRANSFAC). Features of RIs for the learning process are based on a correlation meta-analysis of 4064 gene expression profiles from 76 studies, *in silico* predictions of transcription factor binding sites (TFBSs) and combinations of these employing knowledge about co-regulation of genes by a common TF (TF-module). The trained SVMs were applied to infer new RIs for a large set of TFs and genes. In a case study, we employed the inferred RIs to analyze an independent microarray dataset. We identified key TFs regulating the transcriptional response upon interferon alpha stimulation of monocytes, most prominently interferon-stimulated gene factor 3 (ISGF3). Furthermore, predicted TF-modules were highly associated to their functionally related pathways.

Conclusion: Descriptors of gene expression, TFBS predictions, experimentally verified binding information and statistical combination of this enabled inferring RIs on a genome-wide scale for human genes with considerably good precision serving as a good basis for expression profiling studies.

Contact: r.koenig@dkfz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 10, 2010; revised on May 4, 2011; accepted on June 13, 2011

1 INTRODUCTION

Human gene regulation involves numerous mechanisms comprising protein–protein interaction, DNA binding and transcription, epigenetic DNA modifications, RNA interference and translation. Central to this is the specific binding of transcription factors (TFs) to promoters of genes to regulate their transcription, and the discovery of such regulatory interactions (RIs) to reconstruct large-scale regulatory networks is a main focus of systems biology research. So far, several hundreds of TFs have been identified for many species (Matys *et al.*, 2006). Some TFs bind exclusively to distinct DNA sequence motifs at specific conditions, whereas others are ubiquitously active (Farnham, 2009). Chromatin immunoprecipitation (ChIP) assays have been used to infer TF binding to the promoter of the investigated gene. This was scaled up by ChIP-on-chip technology to obtain the location of specific TF binding genome wide. However, results from such investigations strongly depend on the studied cellular system and treatment. Besides this, computational methods were developed and applied to predict transcription factor binding sites (TFBSs) independent from the samples under study (Stormo, 2000; Valen *et al.*, 2009). These predictions were mainly based on motif searches with position weight matrices (PWMs). PWMs are probabilistic representations of a frequency distribution of nucleotides at each position of a binding site. In contrast to ChIP-on-chip assays, genome-wide PWM searches detect potential TFBSs for any TF (for which a PWM has been assembled from experimentally discovered binding sites) independent of conditional restrictions and thus provide information about TFBSs in an unbiased manner. Predictions with PWMs have been effectively applied to identify the relevant TFs and their sets of regulated genes in gene expression data (Segal *et al.*, 2003a; Sinha, 2006). However, TFBS predictions are rather unspecific and therefore come along with high false positive rates (Stormo, 2000).

Several methods have been developed to construct large-scale regulatory networks using gene expression data, genome-wide ChIP profiles, PWM-scans and a combination of these (Bonneau, 2008). The availability of abundant experimental data for various model organisms enabled to infer regulatory networks for microorganisms explaining and predicting gene expression, e.g. for *Escherichia coli* (Faith *et al.*, 2007), *Saccharomyces cerevisiae* (Bar-Joseph *et al.*, 2003; Joshi *et al.*, 2009; Segal *et al.*, 2003b) and the *Halobacterium* NRC-I (Bonneau *et al.*, 2006). Furthermore, methods were designed to infer significant RIs between TFs and genes using Pearson's correlation and mutual information of gene expression, e.g. the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE; Margolin *et al.*, 2006) and the Context Likelihood of Relatedness (CLR; Faith *et al.*, 2007). ARACNE and CLR were jointly applied to identify target genes for Nrf2 (nuclear factor

*To whom correspondence should be addressed.

erythroid 2-related factor) in the lung of mice in response to oxidative stress (Taylor *et al*., 2008). Recently, the third Dialogue on Reverse Engineering Assessment and Methods (DREAM3) compendium has set up a synthetic data compendium to benchmark several methods inferring RIs. The top five performing algorithms integrated both the provided steady-state (unperturbed, knockdown and knockout) and time-series data (multifactorial perturbations). The performance of the best method apparently depended mainly on predictions reconstructed from steady-state levels of the gene knockout datasets (Marbach *et al.*, 2010) and prediction algorithms for steady-state conditions are mainly required for predicting regulation in tumor samples. A modified version of CLR was among the top five performers and its optimal performance was achieved when using comprehensive knock-out data alone (Madar *et al.*, 2010). The compendium data resembled small subnets of *E.coli* and yeast and provides synthetic transcriptional data. However, the data—like many of the prediction methods—neglect post-transcriptional regulation of TFs. Besides this, the underlying presumption that expression of the target genes depends mainly on the mRNA gradients of their regulating TFs is often violated, specifically in higher eukaryotes. In turn, regulation of TFs on the protein level plays a substantial role, e.g. for hypoxia-inducible factors (HIFs; Kaelin *et al.*, 2002), p53 (Harris and Levine, 2005) and retinoblastoma 1 (RB1; Chen *et al.*, 2009). Additionally, some prediction algorithms are tailored to infer condition-specific RIs for a single or a few TFs rather than predicting RIs for numerous TFs and a wide range of cellular systems and conditions.

To address these limitations, we developed the regulatory interaction predictor (RIP), a supervised machine learning approach that predicts RIs between a large number of human TFs and genes, independent of any specific condition. Our approach distinguishes between TFs and genes and does not presume any dependency of target genes on the gene expression gradients of their regulating TFs. It bases on the knowledge of experimentally derived regulatory interactions in human. For deriving RIs for regulation studies of human cells, RIP has been implemented in a package for the statistical software R and is available for download at http://www.ichip.de/software/RIP.html.

2 METHODS

2.1 Gene expression analysis

Gene expression data was taken from the CAMDA 2007 dataset containing 5896 gene expression profiles collected from a wide range of human cancer types comprising normal and disease tissue samples which were performed with Affymetrix HG-U133A microarrays (ArrayExpress, www.ebi.ac.uk/arrayexpress, accession E-TABM-185). To get unbiased datasets, we disregarded cell line experiments and experiments with <10 samples. Finally, we used gene expression data from 4064 primary human tissue samples of 76 experimental subsets (=conditions) for our meta-analysis. Microarray probe-sets were included in the analysis if they mapped to exactly one gene from the EntrezGene database (Maglott *et al.*, 2007) according to Affymetrix annotations. For each of these probe-sets, the raw expression values were used from the probes located at the 3′ end of their target sequence to minimize RNA degradation effects and reverse transcriptase errors. With this, we obtained expression levels of 13 069 genes for all 76 conditions. Only these genes were considered. Each subset (microarrays of one condition) was normalized using the Robust multi-array average (RMA) method as implemented in the affy R-package (Bioconductor release 2.4, www.bioconductor.org). Pearson's correlation coefficients were computed for each gene pair and condition by a correlation meta-analysis. To account for anticorrelation due to inhibitory signaling propagation, the absolute correlation values were used. We employed a filtering approach adapted from Zhou *et al.* (2005) to select highly correlated gene pairs. The filter consisted of two parameters: The filter consisted of two parameters: correlation coefficient (CC) and fraction of conditions (FoC). When co-applied, they select gene pairs that exceed a defined minimum correlation (absolute value) in a defined minimum percentage of the 76 conditions. CC and FoC each take values between 0 and 1. Applying CC = 0.6 and FoC = 0.25 therefore selected those gene pairs that correlated >0.6 (or <-0.6) in >19 conditions (25%).

2.2 Identification of functionally related gene pairs using Gene Ontology

To estimate the functional relatedness between genes in a gene pair, we compared their Gene Ontology (GO) terms. The mapping of GO terms of biological processes was downloaded from EntrezGene (http://www.ncbi.nlm.nih.gov/gene). The GO term hierarchy was taken from the GO.db R-package in Bioconductor release 2.4. Following an approach described elsewhere (Zhou *et al.*, 2005), we constructed 81 functional categories. A GO term was selected as a functional category if its annotation contained $\geq$150 genes of our analyzed genes and if each of its children contained $<$150 genes, resulting in 81 midrange GO terms. These 81 GO terms described functional categories that were used to estimate the functional relation of gene pairs from the correlation meta-analysis. We assessed a pair of genes as functionally related if they shared at least one of these functional categories.

2.3 The gold standard

The TRANSFAC database v2009.2 (Wingender *et al.*, 1996) provided PWMs used for TFBS predictions as well as a collection of TFs and their target genes derived from published experiments. TRANSFAC contained redundant entries for a number of TFs. Therefore, we manually corrected this by pooling TF entries if all their subunits were encoded by the same genes (the same EntrezGene IDs). We further discarded TFs with less than two RIs. This was done because we could not define appropriate validation sets for such TFs (to estimate their performance). Additionally, 72% of these TFs did not have any PWM motif in TRANSFAC (which was needed for several machine learning features, see below). This yielded 303 TFs with 2896 RIs for 949 regulated genes. For all these genes, we had the respective probes on the Affymetrix microarrays and promoters for the TFBS predictions. For the machine learning approach, we defined a gold standard comprising true positive and true negative regulatory interactions (True RIs and True non-RIs). True RIs (2896) were extracted from TRANSFAC and based on published experiments. The remaining 284 641 possible combinations of the 303 TFs and 949 genes were defined as True non-RIs. This is based on the assumption that regulatory networks are sparse and therefore a vast majority of unknown TF–gene pairs are unlikely to interact. It is to note that a large number of interactions may not have been discovered yet. Still, even if one assumes that e.g. only 10% of interactions have yet been discovered, our 'True' non-interactions would comprise ∼26 000 wrongly labeled interactions. This would still be acceptable compared to the much larger amount of remaining ∼258 000 real True non-RIs (out of 284 641 non-RIs).

2.4 TFBS predictions with PWM motifs

Promoter sequences were extracted from EntrezGene using the biomaRt package for R (Bioconductor release 2.4). Sequences from the annotated transcriptional start site up to 1 kb upstream were considered. To detect putative TFBS, PWMs were taken from TRANSFAC v2009.2 and PWM-scans were performed as described previously using the cureos R-package v0.3 (Westermann *et al.*, 2008). *P*-values for each prediction were obtained

by comparing its score to 10 000 randomly generated sequences (for details, see Westermann *et al.*, 2008). A motif was considered to be significant if $P < 0.1$. Hits with a $P \geq 0.1$ were discarded.

2.5 Defining the features for the classifier

Ten features were calculated to describe discriminating properties of a pair of a TF and a gene (putative RI) based on TFBS predictions from PWM-scans, the correlation meta-analysis and information about co-regulation of genes from the gold standard. The correlation was used to (i) identify all genes that correlate at high levels with a given gene (defining sets of *correlation neighbors*, see below) and to (ii) compare the average correlation of a candidate target gene to known TF targets (or non-targets). The features, therefore, do not presume any dependency between the expression profiles of a target gene and the expression of TF encoding genes.

All possible gene pairs from the gene expression analysis were filtered by applying the two filters CC = 0.6 and FoC = 0.25. For each remaining gene pair, we defined the two genes to be *linked by correlation*. The set of genes with correlation links to a given gene were then designated to be its *correlation neighbors*. Six features for a putative RI were based on correlation neighbors:

(1) Feature one was the number of correlation neighbors of the corresponding gene.

(2) Feature two was the number of correlation neighbors (including the corresponding gene) which were known to be regulated by the corresponding TF (True RIs, taken from the gold standard).

(3) Feature three was $-\log 10(P)$ in which P was the estimated significance for the enrichment of known regulated genes (True RIs in the training sets, taken from the gold standard) in the correlation neighbors (including the corresponding gene) in comparison to all other genes (P was calculated by a Fisher's exact test).

(4) Feature four was the number of correlation neighbors with a significant PWM hit of the corresponding TF.

(5) Feature five was $-\log 10(P)$ in which P was the estimated significance for enrichment of PWM-hits of the TF within the correlation neighbors including the corresponding gene (P was calculated by a Fisher's exact test).

(6) Feature six was the number of correlation neighbors that were known to be regulated (True RIs in the training sets, taken from the gold standard) and which had a significant PWM hit of the TF.

Two features were added describing TFBS predictions and knowledge about co-regulation:

(7) Feature seven was $-\log 10(P)$ in which P was the significance of the PWM hit of the corresponding TF.

(8) Feature eight was the number of genes known to be regulated by the TF (True RIs in the training sets, taken from the gold standard). This feature was added to enable differentiation between universal and specific TFs.

Furthermore, two features were defined considering the correlation of known co-regulated or non-co-regulated genes:

(9) We selected all genes known to be regulated by the corresponding TF (True RIs in the training sets, taken from the gold standard). For each of these target genes, we calculated the median correlation to the corresponding gene of the RI over all conditions (76 conditions). Feature nine was the average of these medians.

(10) We selected all genes which were not known to be regulated by the corresponding TF (True non-RIs in the training sets, taken from the gold standard). Feature 10 was then calculated in analogy to feature nine.

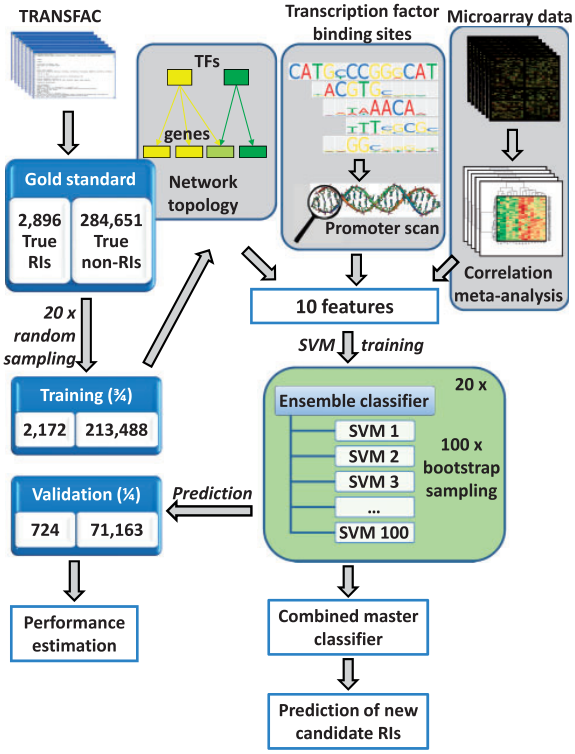


Fig. 1. General workflow. Features for inferring RIs between TFs and genes were derived from three different aspects: co-regulation of genes derived from a meta-analysis of gene expression profiles, TF binding site predictions and statistics about a combination of both including experimentally validated binding information from the training set (gold standard). The information of the gold standard was also used to define True RIs and True non-RIs. SVMs were trained with True RIs and True non-RIs and then used to predict new RIs. For training, True RIs and True non-RIs were divided into a training set and a validation set. An equal number of True RIs and True non-RIs were randomly drawn (bootstrapping approach) 100 times and used to train 100 different SVMs yielding one ensemble classifier. Each ensemble classifier was evaluated with its validation set. This procedure was repeated 20 times yielding an averaged estimate about their performances. The classifiers were combined to one master RIP classifier containing 2000 SVMs and applied to predict new RIs.

During training the classifiers, information of known RIs from the validation sets was not used, and all RIs from the validation sets were considered as True non-RIs.

2.6 Training, validating and applying the classifier

We performed a 20× 100-fold stratified cross-validation (overview of the workflow is given in Fig. 1). The classifications were performed using support vector machines (SVMs) from the R-package MCRestimate of Bioconductor release 2.4. SVMs with Gaussian kernels were employed. For the training sets, we randomly selected 75% of all True RIs and True non-RIs (2172 True RIs, 213 488 True non-RIs). The remaining of 25% of RIs served for validation to estimate the performance of the classifiers. To optimize the parameters (kernel width and cost function), 75% of True RIs of the training set were drawn with replacement from the 2172 True

RIs and the same amount from the 213 488 True non-RIs of the training set. One SVM classifier was trained with these samples and kernel width γ and cost function c were optimized by a grid search employing $\gamma = 2^i$, $i \in \{-10, -8, -6, -4, -2, 0, 2, 4\}$ and $c = 2^j$, $j \in \{-6, -4, -2, 0, 2, 4, 6, 8, 10\}$ using the rest of the training set for validation. This was done by a 10-fold inner cross-validation of the MCR estimate package. To obtain variability and to account for the high amount of True non-RIs, we performed this procedure 100 times yielding 100 SVM classifiers by using different sets of randomly selected True RIs and True non-RIs from the training set. All 100 trained machines were combined and used as one ensemble classifier. The ensemble classifier was validated with the validation set by a voting scheme. Each SVM classifier voted for an RI of the validation set and the minimum number of positive votes was set to define the stringency for the ensemble classifier. During training, True RIs from the validation sets were set as True non-RIs to leave any class-label information of the validation sets untouched (for feature 10, this was done *vice versa* with True non-RIs from the validation sets). The whole process was repeated 20 times using different randomly selected training and validation sets. The overall performance was then estimated from the average of all 20 cross-validations for each stringency threshold. To predict new RIs, all information from the gold standard was employed to train the machines. All trained machines (2000) were taken as one combined master classifier (RIP master classifier) using again the voting scheme in which each SVM contributed one vote. Confidence values of the predictions were calculated from the number of positive votes according to the averaged precision values of the cross-validation.

2.7 Enrichment tests for differentially expressed genes of the case study

In the case study, interferon α (IFNα) induction was examined using human monocytes and Affymetrix HG-U95Av2 microarrays (Tassiulas *et al.*, 2004). Data were downloaded from the NCBI Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo/, accession GSE1740) and normalized as described above. Differentially expressed genes were determined between samples with and without IFNα stimulation (six samples each) using the significance analysis of microarrays (SAM) and a false discovery rate (FDR) of <0.01 (Tusher *et al.*, 2001). One-sided Fisher's exact tests were performed to identify overrepresented differentially expressed genes among the predicted TF-modules (vote cutoff 1600). *P*-values were corrected for multiple testing by the (Benjamini and Hochberg, 1995).

2.8 Associating TF-modules to pathways

Predicted TF-modules were associated with pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000). For each TF-module, genes were selected from predicted RIs with ≥ 1600 votes from the RIP master classifier. One-sided Fisher's exact tests were performed to identify pathways being significantly enriched in genes of the predicted TF-modules. To focus on major signatures, pathways and TF-modules with less than three genes as well as pathways assigned to diseases were not considered. This resulted in 176 pathways of signaling and metabolism for the analysis. All *P*-values were corrected for multiple testing using the Benjamini–Hochberg method.

3 RESULTS AND DISCUSSION

3.1 Predicting regulated genes of TFs with a machine learning approach

Figure 1 gives an overview of the workflow. True TF–gene interactions (True RIs) were derived from TRANSFAC (Wingender *et al.*, 1996) database (the gold standard) comprising 2896 experimentally well-studied RIs of 303 TFs and 949 genes. All other 284 651 pairs of TFs and genes from these sets were considered as True non-Regulatory Interactions (True non-RIs). For each RI, 10 features were calculated to separate True RIs from True non-RIs. These features were combined from results of a correlation meta-analysis of gene expression (4064 microarrays covering 76 conditions), PWM-scans and information about co-regulation of genes by common TFs (TF-modules). We used these features for training and validation of classifiers (SVMs). As the number of True RIs was sparse compared with the number of True non-RIs, we performed a stratified 20× 100-fold cross-validation for training the classifiers and estimating their performances. Finally, one master classifier was used combining 20 ensemble classifiers (consisting of 100 SVMs each) to predict new RIs.

3.2 Genes with correlated gene expression share biological processes

Our prediction method is based on the assumption that genes regulated by the same TF share common cellular processes and thus correlate in their gene expression. To get a quantitative estimate for this assumption, we performed a correlation meta-analysis and compared the correlation of gene pairs with similar function to randomly selected gene pairs. For the correlation meta-analysis, we calculated Pearson's correlation for all possible gene pairs for all 76 conditions. We selected gene pairs at different levels of correlation (adjusting CC and FoC, see Section 2) and estimated their functional relation using GO terms (Ashburner *et al.*, 2000) for biological processes. We considered 81 GO terms for the analysis adapting an approach described elsewhere (Zhou *et al.*, 2005). The 81 GO terms were retrieved by selecting terms that contained at least 150 annotated genes and that had only children with <150 annotated genes. This cross-section through the GO graph yielded a well-balanced selection of GO terms which were sufficiently descriptive and specific to describe particular biological functions on a broad range. To quantify the functional relatedness of gene pairs, we defined the functional similarity score (FS-score) as the percentage of gene pairs sharing at least one selected GO term out of all gene pairs. Figure 2 shows the results for different correlation stringencies. Notably, for a wide range of cutoffs (selecting ≤ 5000 genes, see Fig. 2), the FS-score of our inferred gene pairs was higher than for gene pairs of the gold standard (pairs of genes known to be regulated by the same TF). Our inferred pairs showed FS-scores between 14.8% and 58.3% (stringency parameters CC = 0.6–0.9, FoC = 0.25–0.5, see Section 2) while the gold standard had an FS-score of 35.3%. The FS-score increased with higher stringency (up to CC = 0.8, FoC = 0.35) from 14.8% to 57.3%.

Interestingly, increasing the stringency further reduced the FS-score (at cutoffs for which the number of selected gene pairs was <300). We found that this behavior was due to an increased fraction of constitutively expressed gene families (e.g. hemoglobins, histones, immunoglobulins) that show high correlation of expression between each other without sharing any common biological processes. We compared these results to 100 000 randomly selected gene pairs sampled from the same gene-pool. For these random samples, the FS-score was distinctively lower (11.2%). These results demonstrate that genes with correlated expression were considerably often involved in similar cellular processes, which was comparable to gene pairs known to be regulated by the same TFs.

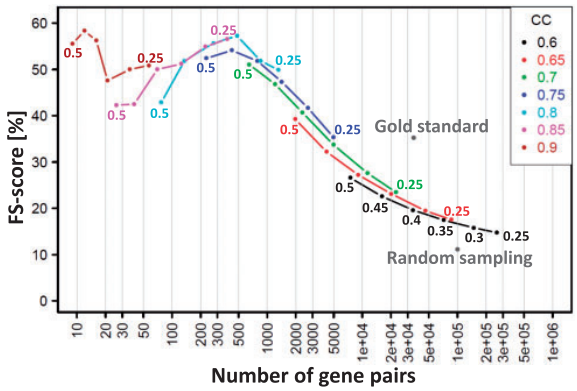


Fig. 2. Gene pairs with high expression correlation are functionally related. The graph shows the FS-score which is the percentage of gene pairs sharing at least one functional category for a variety of different stringency criteria [Pearson correlation (CC) and fraction of classes (FoC)]. For example, setting the threshold for CC to 0.85 in >25% (FoC = 0.25) of the datasets yielded 380 annotated gene pairs of which 56.6% (=215) shared the same functional GO category. For comparison, the gold standard comprised 35.3% (12 176 out of 34 538) pairs having at least one functional GO category in common and only 11.2% of 100 000 randomly selected gene pairs had common functional GO categories.

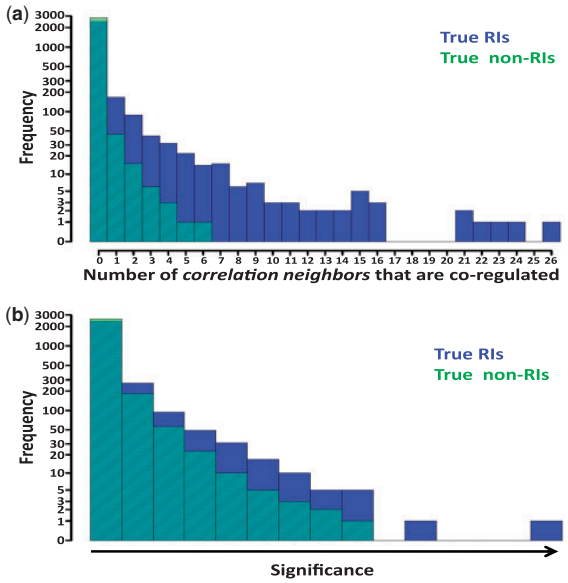


Fig. 3. Distributions for True RIs and True non-RIs of two selected features. (**a**) Frequency distributions are given for feature two of True RIs (blue bars) and True non-RIs (green bars, overlap with blue appears dark green). Feature two was the number of correlation neighbors which were known to be regulated by the corresponding TF (True RIs, taken from the gold standard). The higher the feature value, the more correlating genes are known to be regulated by the TF. Genes of True RIs had more correlation neighbors that were regulated by the same TF than genes of True non-RIs. (**b**) Frequency distribution of feature five: $-\log 10(P)$ and P was the significance of enrichment of correlation neighbors with TFBSs of the TF. True RIs showed more significant enrichment of correlation neighbors with TFBSs than non True RIs. For comparability, counts for True non-RIs were stratified to the total number of True RIs in this figure.

3.3 Groups of correlated genes are frequently regulated by common TFs

The correlation of gene pairs served as a good basis for deducing features of RIs for a machine learning approach to distinguish True RIs from True non-RIs. For this, we defined correlation links for gene pairs with sufficiently high expression correlation. We chose CC = 0.6 and FoC = 25% as a robust cutoff yielding the most correlation links for the highest number of genes, while having an FS-score that was still sufficiently higher than the FS-score of random gene pairs. For a gene of interest (of an RI), we defined genes as its correlation neighbors if they had a correlation link to that gene. We calculated six features for RIs based on correlation neighbors (features 1–6) and two additional features containing the averaged correlation over all conditions (features 9 and 10). For example, we considered the number of correlation neighbors that were known to have a True RI to the TF of interest. As expected, if there was a True RI between a gene and a TF, the number of correlation neighbors with True RIs of the same TF was higher compared with True non-RIs (Fig. 3a). The other features based on correlation neighbors yielded similar discriminative power between True RIs and True non-RIs. Table 1 contains the results for all features.

3.4 Predicted TFBSs discriminate known RIs from the bulk

To find putative TFBSs, the promoter of each gene was scanned for known binding motifs (using PWMs). We found that True RIs had more often a putative TFBS of the particular TF than True non-RIs ($P < 4.6E$-86 using a Wilcoxon test). Moreover, within the group of correlation neighbors of a gene with a True RI, putative TFBSs of the corresponding TF were more significantly enriched in comparison to True non-RIs (Fig. 3b). In total, we derived four features from TFBS predictions (features 4–7). All four features showed highly discriminative power distinguishing True RIs from True non-RIs (Table 1).

Table 1. *P*-values of Wilcoxon rank sum tests for all features

Feature	1	2	3	4	5	6	7	8	9	10
P-value	5.1 E-03	< 4.6 E-86	< 4.6 E-86	1.5 E-50	4.6 E-86	< 4.6 E-86	< 4.6 E-86	1.5 E-50	7.2 E-55	1.1 E-02

3.5 Classifier performance

The 20 training sets of True RIs and True non-RIs were assembled. For each training set, all features were calculated and 100 SVMs (denoted as one ensemble classifier in the following) were trained by a cross-validation. Each of the 20 ensemble classifiers predicted a separate validation set from the gold standard. The results of all 20 ensemble classifiers were averaged (as an estimate of their performance) and compared with the performance of conventional PWM-scans. We were specifically interested in correctly predicting RIs which was estimated by the precision (rate of true positives

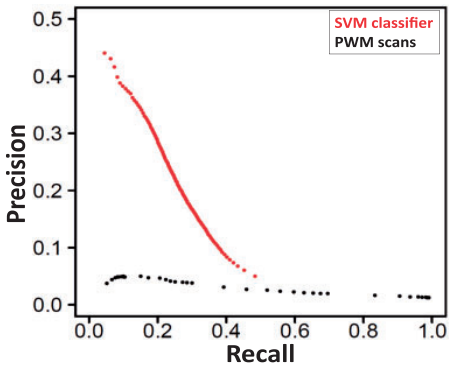


Fig. 4. Recall–precision curve of the ensemble classifiers (red) and the PWM-scans (black). Mean values of all 20 ensemble classifiers were calculated to estimate precision and recall for the master classifier. We obtained recall levels between 4.5% and 48.4% and precision levels between 44.0% and 5.0% by decreasing the threshold of required positive votes from 100 to 1. In contrast, PWM scans only yielded precisions of <5% at all stringencies.

out of all positively classified). Notably, the classifiers showed very good precision. At the most stringent cutoff (all 100 SVMs voted positively), the classifiers reached a precision of 44.0% (recall: 4.5%, accuracy: 99%, specificity: 99.9%). This level of precision is considerably good regarding that regulatory networks are sparse, i.e. the number of True RIs is substantially lower than the number of True non-RIs (∼1:99 in the gold standard). For the lowest stringency (only one classifier needed to vote positively), we got the best recall of 48.4% (precision: 5%; accuracy: 90.1%; specificity: 90.6%). The recall–precision curve for all stringencies is shown in Figure 4. We analyzed the features of True RIs that were never classified positively and compared them to True RIs that were always correctly classified. More than 50% of these misclassified True RIs did not have any significant PWM hits within the promoter regions, whereas >99% of the correctly classified True RIs had significant PWM hits. In contrast to our ensemble classifiers, PWM-scans alone were below 5% precision at all stringency settings, and were therefore considerably outperformed by the ensemble classifiers. It is to note that we estimated the performance of our classifiers rather conservatively. The actual prevalence of True RIs is likely to be much higher than the one obtained from the gold standard as many RIs may not have been discovered so far. Our approach was designed to discover such new RIs and the considerably good precision of the ensemble classifiers implied that combining them to a master classifier suited well to infer new RIs (Section 3.6).

3.6 Inferring new regulatory interactions

To infer new RIs on a large scale, we combined all trained 20 ensemble classifiers to one RIP master classifier. We investigated all combinations of the set of 303 TFs and all 13 069 investigated genes yielding 3 959 907 candidate RIs. Features were calculated with the entire training and validation set. The RIP master classifier was applied providing 2000 votes (one vote from each SVM machine) for each candidate RI. Confidence values were assessed from the precisions of the validation set (averages of 20 ensemble classifiers, see Section 4). With the most stringent cutoff, we predicted 6073 RIs with 44.0% confidence. With ≥31.5% confidence (cutoff of ≥1600 votes), we yielded 73 923 RIs for 301 TFs and 11 263 genes. Supplementary Table S1 contains all predictions with this cutoff. Supplementary Table S2 provides an overview of the numbers of predicted RIs for different cutoffs. The chosen cutoff of 31.5% confidence and 17.7% recall yielded a sufficient number of genes from newly predicted RIs while potentially avoiding too many false positives, and these predicted RIs were used for further investigations. We compared the performance of RIP to the established methods CLR (Faith *et al.*, 2007) and ARACNE (Margolin *et al.*, 2006) using the same expression data as input. However, neither of the algorithms reached acceptable precision levels at any stringency (details of the method and results can be found in Supplementary Table S3.) We were interested why they performed comparably poorly. These methods were based on direct relationships between the expression profiles of TFs and their target genes. This assumption may well hold in lower organisms, but it was not appropriate for our human expression data. Only 2.2% of all true RIs showed an absolute overall correlation (computed over all 4064 microarray experiments) >0.6 between the expression profiles of the TF coding genes and their targets (see Supplementary Table S3). These findings support the utility of RIP, in particular, for regulation analysis of human cells.

Table 2. Association of predicted TF-modules with differentially expressed genes upon IFNα induction in monocytes

Transcription factor	Differentially expressed	Module size	Percentage	Corrected *P*-value
STAT1:STAT2:IRF9	20	28	71.4	6.95e-23
IRF1	58	1187	4.9	5.72e-03
IRF2	15	169	8.9	1.07e-02
STAT1	67	1513	4.4	1.15e-02
GAF	3	5	60	1.15e-02
NFKB1	36	681	5.3	1.59e-02
STAT3	23	384	6	3.21e-02
IRF7	4	17	23.5	3.53e-02
ETS1	48	1065	4.5	3.53e-02
RELA	37	762	4.9	3.53e-02
IRF3	4	18	22.2	3.53e-02
ELF2	3	9	33.3	3.70e-02
SPI1	24	439	5.5	4.63e-02

3.7 Applying the inferred regulatory interactions to a microarray gene expression study: identifying TFs responsive to IFNα

A typical application of our predicted RIs is to investigate the association of TFs and their regulated genes (TF-modules) to a list of differentially expressed genes from a microarray study. A TF can be associated to this list of differentially expressed genes, if the genes of the TF-module are significantly enriched in the gene list. We used microarray data from a study investigating the effect of IFNα on monocytes (Tassiulas *et al.*, 2004). The dataset contained six samples treated with IFNα and six reference samples. We identified 241 significantly differentially expressed genes (FDR < 0.01). These genes were significantly enriched ($P < 0.05$ of a Fisher's exact test) in 13 of the predicted TF-modules (Table 2). On top of the resulting

list were TFs known to respond to IFNα. The most significant enrichment was found for the heterotrimeric TF-complex IFN-stimulated gene factor 3 (ISGF3, $P = 6E\text{-}23$) for which 20 out of 28 predicted target genes were differentially expressed. ISGF3 consists of the subunits STAT1, STAT2 and interferon regulatory factor (IRF) 9. It is activated by cytokines and inflammatory factors (Tassiulas *et al.*, 2004). ISGF3 mediates the transcriptional activation of IFN-inducible genes dependent on IFNα treatment (Fu *et al.*, 1990). Furthermore, we found enrichments for IRF1, IRF2, IRF3, IRF7 and STAT3. Together with ISGF3, their response to IFNα treatment is mediated by the JAK–STAT signaling pathway. Two nuclear factor kappa b (NFκB) subunits (NFKB1 and RELA) and three E-twenty six (ETS)-domain TFs (ETS1, ELF2, SPI1) completed the list of associated TF-modules. Specific roles for all these TF classes have been described in monocytes (Brach *et al.*, 1993; Friedman, 2007) and in IFN signaling of T-helper cells (Grenningloh *et al.*, 2005). SPI1 plays a central role in monocyte and granulocyte development. It interacts with IRF4 and IRF8 upon phosphorylation, and IRF8:SPI1 complexes bind to an ETS/IRF composite element containing an SPI1 binding site. NFκB family members are key regulators of the inflammatory response in monocytes, and AP1 family members cooperate with SPI1 in gene regulation in erythroid cells (Friedman, 2007). We compared our predictions of ISGF3 target genes with predictions employing only PWM-scans. The stringency of the PWM-scans was chosen ($P \leq 0.005$) to get a number of predicted target genes that was comparable to the master classifier. The PWM-scan yielded 29 target genes for ISGF3, only 4 out of which (=13.8%) were differentially expressed in the IFNα study. Our predictions from the machine learning approach were considerably more sensitive to infer genes and TFs involved in the gene regulatory processes of IFNα response (our predictions: 20 differentially expressed out of 28 predicted ISGF3 target genes = 71.4%).

Additionally, we analyzed the expression levels of the TF encoding genes. IRF7 from the listed 13 TFs showed strong differential expression at a significant level and was upregulated in the IFNα-induced cells. IRF1, IRF2, STAT2, STAT3 and RELA had slightly (but consistently) elevated expression levels, whereas IRF3 was slightly decreased in the IFNα-induced cells.

3.8 Genes were highly enriched in pathways which were associated with TFs regulating these genes

To investigate the functional relevance of our predicted RIs, we associated the predicted TF-modules with signaling and metabolic pathways from KEGG database. We selected RIs from regulated genes which were found in KEGG, yielding 220 TF-modules of 22 345 predicted RIs with 3276 genes. Each gene set of the TF-modules was tested to be enriched in each pathway of KEGG (Fisher's exact test). The results cover a variety of signaling and metabolic pathways (Table 3). Additionally, we performed the same analysis for 565 genes with known RIs in TRANSFAC (which was the overlap of KEGG and TRANSFAC) and provided the results in Supplementary Table S4. Most of the associations with our predicted RIs have been described previously and reflect the biology of the pathways. Potentially novel associations are marked in bold in Table 3. In the following, we describe the findings for cell cycle and proliferation-related signaling pathways (yellow in Table 3) in more detail (other pathways are described in Supplementary Table S5). We observed highly significant associations of E2F1 and E2F4 and their hetero-dimers with TFDP1 and TFDP2 to cell cycle and related pathways. Their function in cell cycle progression has been extensively reported in the literature (see e.g. Weinberg, 2006).

Table 3. Associations of predicted TF-modules with pathways (new associations are given in bold)

Transcription factors	Pathway
IRF1, IRF2, IRF3, IRF5, **IRF7**, STAT1, STAT3 **NFATC2**, **NF-GMa**, CD28RC, HMGA1 **NFκB**, **NFKB1**, **NFKB1:RELA**, **RELA** JUN, CEBPA, CEBPB	Cytokine–cytokine receptor interaction
IRF7, **STAT4**, STAT1:STAT2:IRF9 CD28RC, NFATC2, **NF-GMa**, POU1F1	Jak–STAT signaling pathway
IRF1, **IRF3**, **IRF7**, **NFκB**, **RBPJ**	Toll-like receptor signaling
NFATC2, **NF-GMa**	Fc epsilon RI signaling pathway
NF-AT, **NFATC2**, **NF-Gma**, SPI1	Hematopoietic cell lineage
IRF2, **NF-AT**, **NF-AT1**	T cell receptor signaling
ELF1	Natural killer cell-mediated cytotoxicity
IRF1, IRF2, LEF1, XBP1 **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK**	Antigen processing and presentation
IRF1, XBP1 **RFX2**, **RFX3**, **RFX5:RFXAP:RFXANK**	Cell adhesion molecules (CAMs)
ETS1, STAT1	MAPK signaling pathway
IRF2, **NFKB1:RELA**	Apoptosis
SP4	Calcium signaling pathway
TCF7L2	Wnt signaling pathway
p53, p73	p53 signaling pathway
E2F:DP, E2F4, NFYA	Cell cycle
E2F:DP, E2F1:TFDP1/TFDP2, E2F4	DNA replication
E2F:DP	Purine metabolism
E2F:DP, E2F4	Pyrimidine metabolism
E2F:DP, E2F1:TFDP1/TFDP2, E2F4	Nucleotide excision repair
E2F1:TFDP1/TFDP2, E2F4	Mismatch repair
GATA4, **NR5A1**	C21-steroid hormone metabolism
NR5A1	Androgen and estrogen metabolism
NR1H4, **PPARA:RXRA**, **RXRA**	PPAR signaling pathway
NR1I2, **RXRA:NR1I2**, **RXRA:NR1I3**	Retinol metabolism
NR1I2, **RXRA:NR1I2**	Linoleic acid metabolism
HNF1A, **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3**	Drug metabolism—cytochrome P450
HNF1A, NFE2 **NR1I2**, **RXRA:NR1I2**, **RXRA:NR1I3**	Metabolism of xenobiotics by cytochrome P450
FLI1, HNF1B, SMAD3	ECM–receptor interaction
HNF1B	Focal adhesion
NFE2L2	Cell junctions
NR1H3, SP4	Glutathione metabolism
	Neuroactive ligand–receptor interaction
	Non-homologous endjoining
RARB	Proteasome
	Protein export
	Oxodative phosphorylation

Nuclear transcription factor Y alpha (NFYA) was also associated with these pathways. The binding of E2F is dependent on an adjacent CCAAT site being occupied by NFY (Zhu *et al.*, 2004), and functionality of NFY has been shown for G2/M transition of the cell cycle in combination with p53 (Imbriano *et al.*, 2005). p53 and its family member p73 were associated with the pathway of p53 signaling. TCF7L2 was associated with the Wnt signaling pathway, which is in accordance with the functionality of TCF7L2 in that canonical pathway. In summary, these findings well support the functional relevance of the predicted RIs for these pathways.

3.9 Predicted RIs are supported by comparison with an independent database

To validate our predictions, we compared the predicted RIs ($\geq$1600 votes) to known RIs of 74 TFs from 25 TF families of the Transcriptional Regulatory Element Database (TRED). Details on the procedure and all results can be found in Supplementary Table S6. In brief, we tested for association (by means of overrepresentation) of sets of RIs from TFs in TRED (TRED TF-modules) with our predicted RIs for these TFs (our predicted TF-modules). In 85.4% of all tested TFs, the correct TF-family was among the top three TRED TF-modules, and in 73.5% the actual TF was assigned correctly (again considering the top three hits). These results further validated RIP using an independent source of RIs of a significant number of well-studied TFs.

4 CONCLUSIONS

We presented a novel machine learning approach (RIP) that predicts gene regulation on a genome-wide scale with considerably high precision. The predictions are based on a broad range of conditions and can be applied to more specific experiments as well. Presumably, only a minor fraction of all RIs has been discovered so far. Given knowledge of 2896 True RIs for 949 human genes, a number of 6073 RIs (at the most stringent cutoff) predicted out of 3 959 907 candidate RIs of 303 TFs and 13 069 genes seems reasonable. Also a number of 73 923 RIs (lower stringency, requiring only 80% of positive votes) yielded useful results when applied to the case study of IFN signaling. We employed descriptors for inferring gene regulation from three different aspects: (i) a meta-analysis was performed to obtain groups of genes with high correlation in different cell and tissue types from different experiments. (ii) TFBSs were predicted *in silico* using PWMs to scan promoters of every investigated gene. With these predictions, known transcription factor–gene regulations could be well distinguished from other transcription factor–gene combinations. (iii) Statistical descriptors were used to exploit the association of co-regulation, correlation and TFBS predictions. This information was integrated by a machine learning approach yielding a powerful tool to infer regulatory networks that can be adjusted for recall and precision at a high level of prediction performance. We applied RIP to infer RIs in human. RIP is intended to be an *in silico* approach to extend lists of known and experimentally derived RIs. It needs PWMs (not necessarily extracted from TRANSFAC) of known TFs with known binding sites for initial learning. Thus, other TFs can also be included into the analysis if their binding motifs have been identified in some target genes and a PWM motif could be generated. With that knowledge, it is possible to extend the gold standard and predict RIs for these TFs. If a comprehensive gold standard of experimentally validated True RIs is given (and PWMs for the corresponding TFs), the method can be readily applied also to other well-studied organisms. The presented RIP classifier offers a wide range of applications for gene expression analyses such as identification of key transcription factors and pathways involved in the pathology and changed function of the investigated cells.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Andrea Califano and his team for their cooperativeness and help with the ARACNE algorithm.

Funding: BMBF-FORSYS consortium Viroquant (#0313923), Helmholtz Alliance on Systems Biology (SBCancer), and the Nationales Genom-Forschungs-Netz (NGFN+) for the neuroblastoma project ENGINE (#01GS0898).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Bonneau,R. (2008) Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.*, **4**, 658–664.

Bonneau,R. *et al.* (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Brach,M.A. *et al.* (1993) Transcriptional activation of the macrophage colony-stimulating factor gene by IL-2 is associated with secretion of bioactive macrophage colony-stimulating factor protein by monocytes and involves activation of the transcription factor NF-kappa B. *J. Immunol.*, **150**, 5535–5543.

Chen,H.Z. *et al.* (2009) Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer*, **9**, 785–797.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

Friedman,A.D. (2007) Transcriptional control of granulocyte and monocyte development. *Oncogene*, **26**, 6816–6828.

Fu,X.Y. *et al.* (1990) ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains. *Proc. Natl Acad. Sci. USA*, **87**, 8555–8559.

Grenningloh,R. *et al.* (2005) Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. *J. Exp. Med.*, **201**, 615–626.

Harris,S.L. and Levine,A.J. (2005) The p53 pathway: positive and negative feedback loops. *Oncogene*, **24**, 2899–2908.

Imbriano,C. *et al.* (2005) Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters. *Mol. Cell Biol.*, **25**, 3737–3751.

Joshi,A. *et al.* (2009) Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, **25**, 490–496.

Kaelin,W.G. Jr (2002) Molecular basis of the VHL hereditary cancer syndrome. *Nat. Rev. Cancer*, **2**, 673–682.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Madar,A. *et al.* (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE*, **5**, e9803.

Maglott,D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Segal,E. *et al.* (2003a) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19** (Suppl. 1), i273–i282.

Segal,E. *et al.* (2003b) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Sinha,S. (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, **22**, e454–e463.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tassiulas,I. *et al.* (2004) Amplification of IFN-alpha-induced STAT1 activation and inflammatory function by Syk and ITAM-containing adaptors. *Nat. Immunol.*, **5**, 1181–1189.

Taylor,R.C. *et al.* (2008) Network inference algorithms elucidate Nrf2 regulation of mouse lung oxidative stress. *PLoS Comput. Biol.*, **4**, e1000166.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Valen,E. *et al.* (2009) Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput. Biol.*, **5**, e1000562.

Weinberg,R.A. (2006) *The Biology of Cancer*. Garland Science, New York.

Westermann,F. *et al.* (2008) Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol.*, **9**, R150.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Zhou,X.J. *et al.* (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.*, **23**, 238–243.

Zhu,W. *et al.* (2004) E2Fs link the control of G1/S and G2/M transcription. *EMBO J.*, **23**, 4615–4626.