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

Presented by
Alexandra Maria Poos, M.Sc.
born in Mönchengladbach, Germany
Oral examination: 17.05.2019

Mixed Integer Linear Programming based approaches to study telomere maintenance mechanisms

Referees: Prof. Dr. Karsten Rippe

Prof. Dr. Rainer König

Declaration

I hereby declare that I have written the submitted dissertation “Mixed Integer Linear Programming based approaches to study telomere maintenance mechanisms” myself and in this process have used no other sources or materials than those explicitly indicated. I hereby declare that I have not applied to be examined at any other institution, nor have I used the dissertation in this or any other form at any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

_____ _____

(Place, Date) Alexandra Poos

List of publications

During this thesis, I contributed to the following publications:

Poos AM, Maicher A, Dieckmann AK, Oswald M, Eils R, Kupiec M, Luke B, König R (2016). Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. *Nucleic Acids Res*, 44(10):e93.

Gietzelt M, Höfer T, Knaup-Gregori P, König R, Löpprich M, **Poos A**, Ganzinger M (2016). The Use of Tools, Modelling Methods, Data Types, and Endpoints in Systems Medicine: A Survey on Projects of the German e:Med-Programme, *Stud Health Technol Inform*. 2016; 228:670-4.

Deeg KI, Chung I, **Poos AM**, Braun DM, Korshunov A, Oswald M, Kepper N, Bender S, Castel D, Lichter P, Grill J, Pfister SM, König R, Jones DTW, Rippe K (2017). Dissecting the alternative lengthening of telomeres pathway in pediatric glioblastoma. *bioRxiv* 129106; doi: https://doi.org/10.1101/129106.

Minner S, Lutz J, Hube-Magg C, Kluth M, Simon R, Höflmayer D, Burandt E, Tsourlakis MC, Sauter G, Büscheck F, Wilczak W, Steurer S, Schlomm T, Huland H, Graefen M, Haese A, Heinzer H, Jabobsen F, Hinsch A, **Poos AM**, Oswald M, Rippe K, König R, Schroeder C (2019) Loss of CCAAT/Enhancer binding protein alpha (CEBPA) is linked to poor prognosis in PTEN deleted and TMPRSS2:ERG fusion type prostate cancers. *Prostate*, 79: 302-11.

Poos AM, Kordaß T, Kolte A, Ast V, Oswald M, Rippe K, König R (2019). Modelling *TERT* regulation across 19 different cancer types based on the MIPRIP gene regulatory network approach. *bioRxiv* 513259; doi: https://doi.org/10.1101/513259. Under review at BMC Bioinformatics.

Mallm JP, Murat I, Ishaque N, Kugler SJ, Muino JM, Teif VB, Klett LC, **Poos AM**, Großmann S, Erdel F, Tavernari D, Koser SD, Schumacher S, Brors B, König R, Remondini D, Vingron M, Stilgenbauer S, Lichter P, Zapatka M, Mertens D, Rippe K (2019). Linking aberrant chromatin features in chronic lymphocytic leukemia to transcription factor networks. *Mol Syst Biol*, in revision.

Höflmayer D, Steinhoff A, Hube-Magg C, Kluth M, Simon R, Burandt E, Tsourlakis MC, Minner S, Sauter G, Büscheck F, Wilczak W, Steurer S, Huland H, Graefen M, Haese A, Heinzer H, Schlomm T, Jacobsen F, Hinsch A, **Poos AM**, Oswald M, Rippe K, König R, Schroeder C (2019). High expression of CCCTC-binding factor (CTCF) is linked to poor prognosis in prostate cancer, *Mol Oncol*, submitted.

In preparation:

Poos AM, Schröder C, Meiners J, Braun DM, Oswald M, Bauer C, Gunkel M, Wollmann T, Spilger R, Makrypidi-Fraune G, Röll D, Jaishankar N, Stenzel A, Chung I, Frank L, Rohr K, Erfle H, Baniahmad A, Simon R, Sauter G, Rippe K, König R (2019) Modelling of *TERT* regulation leads to new prognostic markers in prostate cancer, in preparation.

Summary

Gene regulation is mainly controlled by a complex network of transcription factors (TFs). TFs activate or repress the expression of a gene by binding to its promoter or enhancer regions. Mutations of TFs or their binding sites are linked to several diseases including cancer. Here, I present a new approach to identify the regulatory relationships between TFs and their target genes based on Mixed Integer Linear Programming (MILP) and machine learning called '**M**ixed **I**nteger linear **P**rogramming based **R**egulatory **I**nteraction **P**redictor' (MIPRIP). Compared to other approaches, MILP has the advantage of using the L1 norm for regression to avoid overestimating outliers and to implement constraints to get sparse models. MIPRIP predicts the expression of a gene of interest by a linear model with all TFs potentially binding to the gene's promoter as covariates. MIPRIP was first enhanced with a statistical analysis pipeline to compare the regulatory processes of a particular gene between two or multiple conditions (MIPRIP-Comparison). The second enhancement of MIPRIP enables a modularity-based approach to analyze the gene regulatory network of the TFs regulating the gene of interest (MIPRIP-Network).

MIPRIP was applied to study the regulation of telomere maintenance, which is crucial for cancer cells to proliferate unlimitedly. The majority of cancer cells maintain their telomeres by re-expressing the reverse transcriptase telomerase, while a minor fraction uses the alternative lengthening of telomeres (ALT) pathway. Firstly, MIPRIP was used to study the regulation of telomerase expression in *Saccharomyces cerevisiae*. *S. cerevisiae* is a well suited model system to study telomere maintenance because of its active telomerase and high structure-function homology to humans. In yeast, I uncovered novel regulators of telomerase expression, several of which affect histone levels or modifications, e.g. Sum1 and Hst1. Secondly, I performed a pan-cancer MIPRIP analysis to identify the most common regulators of the human telomerase reverse transcriptase (*TERT*) gene across 19 different cancer entities and also the specific *TERT* regulators in each cancer type. For prostate cancer, the modularity-based analysis using MIPRIP predicted a subnetwork of 20 regulators, in which PITX1, CTCF, IRF1, TFAP2D, MITF and BHLHE40 were the most important regulators of *TERT* expression. Four

out of these six *TERT* regulators could be validated as novel prognostic markers with elevated protein expression levels in patient samples.

Thirdly, I constructed a classifier to predict the active telomere maintenance mechanism (ALT or non-ALT) of pediatric glioblastoma (pedGBM) patients based on typical telomere features extracted from next-generation sequencing, cytological and molecular assays. After the patient classification several regulators could be identified which were differentially expressed between ALT and non-ALT pedGBM patients using the MIPRIP framework.

In summary, the newly developed MIPRIP framework extends the methodological toolbox to study gene regulation. The application on telomere maintenance provided novel insights about the regulatory processes underlying (i) telomerase expression and (ii) the ALT pathway. Furthermore, I identified new prognostic markers for prostate cancer.

Zusammenfassung

Genregulation wird hauptsächlich von einem komplexen Netzwerk an Transkriptionsfaktoren (TFs) kontrolliert. TFs beeinflussen die Expression der Gene durch Bindung an Promotor- oder Enhancer-Regionen, positiv oder negativ. Mutationen von TFs oder ihren Bindestellen sind mit verschiedenen Tumorerkrankungen assoziiert. Um den Zusammenhang der TFs und ihrer Zielgene besser zu verstehen, habe ich einen neuen Ansatz basierend auf gemischt-ganzzahliger linearer Programmierung und Maschinenlernverfahren entwickelt, den „Mixed Integer linear Programming based Regulatory Interaction Predictor“ (MIPRIP). Dieser Ansatz hat im Vergleich zu anderen Algorithmen den Vorteil, dass eine L1 Norm für die Regression verwendet wird, um zu hoch eingeschätzte Werte (Ausreißer) zu vermeiden, und durch die Implementierung von Beschränkungen dünnbesetzte Modelle entstehen zu lassen. MIPRIP verwendet ein lineares Modell mit allen an den Promotor des Genes bindenden TFs als Kovariaten, um die Genexpression eines Genes vorherzusagen. MIPRIP wurde um eine statistische Analysepipeline erweitert, um die regulatorischen Prozesse eines Genes zwischen zwei oder mehreren Datensätzen zu vergleichen (MIPRIP-Comparison). Des Weiteren wurde MIPRIP mit einem Modularitäts-basierten Ansatz kombiniert, um ein Genregulatorisches Netzwerk von TFs zu identifizieren, welches für die Regulation eines bestimmten Genes verantwortlich ist (MIPRIP-Network).

MIPRIP wurde angewandt, um die Regulation der Telomererhaltung zu untersuchen. Telomererhaltungsmechanismen sind für Tumorzellen essentiell, um sich unbegrenzt teilen zu können. Die meisten Tumorzellen verlängern ihre Telomere, indem sie die reverse Transkriptase Telomerase exprimieren, während andere Tumorzellen einen alternativen Mechanismus verwenden, „alternative lengthening of telomeres“ (ALT). Mit Hilfe von MIPRIP wurde zuerst die Regulation der Telomerase-Gene in *Saccharomyces cerevisiae* untersucht, da sich Hefe auf Grund ihrer aktiven Telomerase und der großen Homologie zu menschlichen Telomer-Proteinen als idealer Modellorganismus für Telomererhaltungsstudien eignet. Hierbei wurden neue Regulatoren der Telomeraseexpression identifiziert, wovon einige, wie z.B. Sum1 und Hst1, die Level oder Modifikationen der Histone beeinflussen. Weiterhin wurde MIPRIP verwendet, um die Regulatoren des

menschlichen Telomerase-Reverse-Transkriptase (*TERT*) Genes in 19 verschiedenen Tumorentitäten zu identifizieren. Dabei lag der Fokus auf TFs, die *TERT* in allen Tumorentitäten regulieren, sowie auf den spezifischen TFs der einzelnen Tumorentitäten. Für Prostatakrebs-Daten wurde mit Hilfe der Modularitäts-basierten Erweiterung von MIPRIP ein Netzwerk von 20 Regulatoren vorhergesagt, wovon PITX1, CTCF, IRF1, TFAP2D, MITF und BHLHE40 die wichtigsten Regulatoren von *TERT* waren. Vier von diesen sechs Regulatoren konnten anhand deren Proteinexpression auf Gewebeschnitten als klinisch neue prognostische Marker validiert werden.

Außerdem habe ich einen Klassifikator konstruiert, der den aktiven Telomererhaltungsmechanismus in pädiatrischen Glioblastomen (pedGBM) vorhersagt. Dafür wurden Ergebnisse für typische ALT-Merkmale aus Sequenzierdaten, sowie zytologischen und molekularen Analysen kombiniert. Damit konnten alle pedGBM Patienten in ALT-positiv und ALT-negativ unterteilt werden. Mit Hilfe von MIPRIP konnten Regulatoren identifiziert werden, die eine unterschiedliche Aktivität zwischen ALT positiven und ALT negativen Patienten zeigten.

Zusammenfassend habe ich mit MIPRIP einen neuen Ansatz entwickelt, um die Regulation von Genen zu untersuchen. Die Anwendung auf Telomererhaltungsmechanismen hat zu neuen Erkenntnissen über die Regulation der Telomerase, sowie den ALT-Mechanismus geführt. Außerdem konnten neue prognostische Marker für Prostatakrebs identifiziert werden.

Contents

List of Figures

List of Tables

List of Abbreviations

General abbreviations

Acc	accuracy
ALT	alternative lengthening of telomeres
AML	acute myeloid leukemia
APB	ALT-associated PML nuclear body
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
aREA	analytic rank-based enrichment analysis
AUC	area under the curve
BH	Benjamini-Hochberg
biRte	Bayesian inference of context-specific regulator activities and transcriptional networks
bp	base pair
C228T	C>T transitions at position 124 bp
C250T	C>T transitions at position 146 bp
C-circles	circular C-rich extrachromosomal telomeric repeats
ChEA	ChIP Enrichment Analysis database
ChIP	chromatin immunoprecipitation
ChIP-on-ChIP	ChIP on a DNA-microarray
ChIP-seq	ChIP followed by sequencing
CLL	chronic lymphocytic leukemia
CNV	copy number variation
D-loop	displacement loop
DNA	deoxyribonucleic acid
ECTRs	extrachromosomal telomeric repeats
EMSA	electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA Elements
FISH	fluorescence in situ hybridization
FPKM	fragments per kilobase per million
GBM	glioblastoma multiforme
GDAC	Genome Data Analysis Center
GC	Guanine-cytosine

GO	Gene Ontology
GRN	gene regulatory network
GSEA	gene set enrichment analysis
HTRIdb	Human Transcriptional Regulation Interaction database
ICGC	International Cancer Genome Consortium
IHC	immunohistochemistry
ILP	integer linear programming
ISMARA	Integrated System for Motif Activity Response Analysis
IQ-Gleason	quantitative Gleason
kb	kilo base
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LP	linear programming
MI	mutual information
MILP	mixed integer linear programming
MIP	mixed integer program
MIPRIP	Mixed Integer linear Programming based Regulatory Interaction Predictor
MIPRIP-Comp	MIPRIP-Comparison
MIPRIP-Net	MIPRIP-Network
NGS	next generation sequencing
NIRVANA	NGE inference via a network-based approach
ORF	open reading frame
PAINT	Predicting ALT IN Tumors
pedGBM	pediatric GBM
PPI	protein-protein interaction
PRAD	prostate cancer
p-value	probability value
PWM	position weight matrix
RABIT	Regression Analysis with Background Integration
RACER	Regression Analysis of Combined Expression Regulation
RMA	robust multi-array average
RNA	ribonucleic acid
RPKM	reads per kilobase per million mapped reads

RSEM	RNA-Seq by Expectation Maximization
SELEX	systematic evolution of ligands by exponential enrichment
SEM	standard error
siRNA	small/ short interfering RNA
SKCM	cutaneous melanoma skin cancer
TBA	total binding affinity
TCGA	The Cancer Genome Atlas
TF	transcription factor
TFBS	TF binding site
TLM	telomere length maintenance
tlm	TLM gene mutants
t-loop	telomeric loop
TMA	tissue microarray
TMM	telomere maintenance mechanism
TRAP	telomeric repeat amplification protocol
TRF	telomere restriction fragment
TRS	telomere length regulating subnetwork
TSS	transcription start site
VIPER	Virtual Inference of Protein activity by Enriched Regulon analysis
VSN	variance stabilization normalization
WGS	whole genome sequencing
wt	wild-type
YEASTRACT	YEAst Search for Transcriptional Regulators And Consensus Tracking

Genes, proteins and transcripts

AR	androgen receptor
ASF1	anti-silencing function 1 histone chaperone
ATM	ataxia telangiectasia mutated protein
ATR	ataxia telangiectasia and Rad3-related protein
ATRX	alpha thalassemia/ mental retardation syndrome X-linked
BATF	basic leucine zipper ATF-like TF
BHLHE40	basic helix-loop-helix family member e40
CEBPA	CCAAT/Enhancer binding protein alpha
CHK1,2	Checkpoint kinase 1,2
CTCF	CCCTC-binding factor
DAXX	death-associated protein 6
ERG	ETS-TF
EST1-3	Ever Shorter Telomeres 1-3
ETS	E-twenty six
GABPA	GA binding protein TF subunit alpha
GLN3	nitrogen responsive transcriptional regulator
H3F3A	H3 histone family member 3A
HMGA2	high mobility group AT-hook 2
HP1	heterochromatin protein 1
HST1	histone deacetylase
IRF1	interferon regulatory factor 1
MITF	melanocyte inducing TF
MXI1	MAX interactor 1
PAX	paired box
PITX1	paired-like homeodomain 1
PML	promyelocytic leukemia
POLR2A	RNA polymerase II subunit A
POT1	protection of telomeres protein 1
PSA	prostate specific antigen
PTEN	phosphatase and tensin homolog
RAD50,51	double strand break repair protein 50,51
RAP1	repressor/ activator protein 1

SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
SRB2	mediator of RNA-polymerase II transcription
SUM1	suppressor of mar 1
TAF1	TATA-box binding protein associated factor 1
TERC	telomerase RNA Component
TERRA	telomeric repeat-containing RNA
TERT	Telomerase Reverse Transcriptase
TFAP2D	AP-2, delta
TIN2	TRF1-interacting protein II
TLC1	template RNA of the RNA in S. cerevisiae
TMPRSS2	transmembrane serine protease 2
TP53	tumor protein p53
TPP1	TIN2- and POT1-interacting protein
TRF1	telomere repeat binding factor 1
TRF2	telomere repeat binding factor 2

1 Introduction

1.1 Transcriptional regulation

The regulation of genes is a central process in living cells. A complex network of transcription factors (TFs), coregulators and chromatin modifier controls the transcription of DNA into RNA. TFs bind to specific DNA sequences in the gene's promoter or enhancer regions activating or repressing the recruitment of RNA polymerase and therefore regulate the transcription of the gene (Fulton *et al.*, 2009; Spitz and Furlong, 2012; Vaquerizas *et al.*, 2009). In humans, there are over 1,600 TFs known which is around 8 % of all human genes (Lambert *et al.*, 2018). TFs can be grouped into families based on their binding domains. In eukaryotes the most prominent TF families are the families of C2H2-zinc finger, homeodomain, basic helix-loop-helix, basic leucine zipper and nuclear hormone receptor (Weirauch and Hughes, 2011). The DNA binding sites of the TFs are typically only 6-12 bases long and are called binding motifs. In some cases, the TF can directly recruit RNA polymerase while in other cases some accessory factors are needed (Frietze and Farnham, 2011). Gene regulation is a complex process. Most TFs act cooperatively with other TFs or co-regulators to induce transcription (Vaquerizas *et al.*, 2009). Several TFs can regulate different genes depending on the cell type (Gertz *et al.*, 2012) and they can bind to more than one binding site in the promoter region of the gene (Wunderlich and Mirny, 2009). Furthermore, the regulation by TF can be controlled through posttranslational modifications, e.g. phosphorylation, ubiquitination or methylation. These modifications substantially influence the regulation of the TFs on their target genes (Filtz *et al.*, 2014). For instance, phosphorylation often leads to dimerization or binding of the TF on the target gene's promoter. Furthermore, TFs can regulate other TFs directly by binding to their promoters or indirectly by influencing the expression of signaling molecules, e.g. kinases, which regulated the other TF. Therefore, the activity of the TFs can often not be determined from their gene expression levels. A deregulation of transcriptional regulators can cause disease, including cancer. This can happen e.g. through mutations of the TF or in their binding sites (Lambert *et al.*, 2018). Therefore, transcriptional regulators are interesting putative drug targets, although

for several TFs the exact mechanism is not known. Therefore, it is important to identify TF-mediated gene-regulatory mechanisms in more detail.

1.2 Identification of TF to target gene interactions

TF-target gene interactions can be studied by chromatin-immunoprecipitation (ChIP) on DNA-microarrays (ChIP-on-ChIP) or followed by sequencing (ChIP-seq), electrophoretic mobility shift assay (EMSA) or systematic evolution of ligands by exponential enrichment (SELEX) (Wilkinson *et al.*, 2017). Particularly ChIP-seq and ChIP-on-ChIP allow large-scale studies of TFs to identify their interactions (Wilkinson *et al.*, 2017). Large repositories of data from ChIP-experiments are e.g. the Encyclopedia of DNA Elements (ENCODE) (Consortium *et al.*, 2012) or the ChIP Enrichment Analysis (ChEA) database (Lachmann *et al.*, 2010). Besides experimental data, there are large collections of computational predictions of TF binding sites (TFBS) (Tompa *et al.*, 2005), e.g. TRANSFAC (Matys *et al.*, 2006) and JASPAR database (Khan *et al.*, 2018), or methods to identify the binding sites of the TFs based on a motif search. These motifs can be represented by position weight matrices (PWMs) (Stormo and Zhao, 2010) and depicted by sequence logos (Schneider and Stephens, 1990). A PWM is a probabilistic description of the binding affinity of the TF showing which nucleotide is preferred at each position (Stormo and Zhao, 2010). PWMs were used to identify TFBS in the genome independent of the cellular function (Kranz *et al.*, 2011). The identified TFBS were only probable predictions and did not include if the TF is present resulting in a high number of false positives (Stormo, 2000). Another computational based method to predict TF binding sites is the 'Total Binding Affinity' (TBA). TBA estimates the binding probability of a TF to the whole promoter region of the gene (Grassi *et al.*, 2015; Molineris *et al.*, 2011). But ChIP-experiments and computational binding site predictions have a high rate of false positives (Pickrell *et al.*, 2011). Furthermore, each ChIP-assay is restricted to only one TF in one condition limiting the characterization of a TF in many different cell types (Trescher *et al.*, 2017). All the available TF-target gene interactions together are a great resource to study gene regulation. Various modeling approaches have been developed to identify the functionally active TF-target gene interactions dependent on the biological context.

Gene regulatory network (GRN) models are well suited to study the transcriptional regulation.

1.2.1 Basic principle of linear regression models

Linear regression models describe the relationship between a dependent (response) variable and one or more independent (predictor) variables. A simple linear regression model with one dependent variable *y* and one independent variable *x* and *n* observations is given by:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{with } i = 1,2,\ldots,n \quad . \tag{1}$$

Equation 1 describes the dependency of *y* on x_i. There can be either a positive, a negative or none-relationship between both variables. ε_i indicates the error term, which is the difference between the given and the estimated value of *y*. The regression line is depicted as the straight red line that fits the *n* observations best. β_0 is the y-axis intercept and β_1 represents the slope of the regression line (Figure 1).

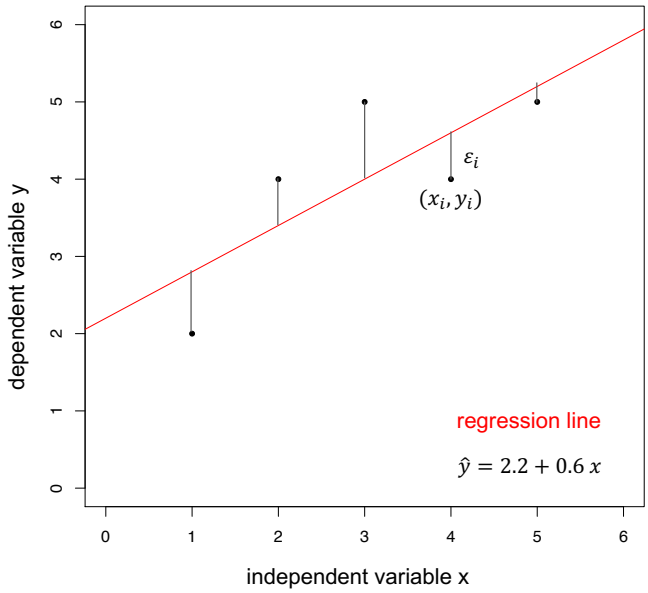

Figure 1. A simple linear regression model.
The red line indicates a straight regression line which can best fit the observations. β_0 is the y-axis intercept, β_1 the slope of the regression line and $\hat{y}$ the estimated *y*-value of variable *x*.

The β-parameters are optimized for all *n* observations during the modeling process by minimizing the difference between the estimated and the actual value which is equal to a minimization of the errors in equation (1). This optimization problem can be solved by using L1 norm which uses the sum of the absolute differences between the estimated and the actual value. To find the best solution all possible combinations have to be tested. L1 optimization can be solved by specific algorithms, in our case by linear programming (LP).
Alternatively, the β-parameters can be estimated by the least-squares method which is based on L2 norm. Compared to L1 norm, L2 norm minimizes the sum of squares of the differences between estimated and the actual value.
A multiple linear regression model describes the relationship of a dependent variable and multiple independent variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \text{ with } i = 1,2,\ldots,n \ , \qquad (2)$$

where *k* is the number of independent variables and *i* is the number of observations. The β-parameters are the regression coefficients. ε_i describes the variations of *y* that are not known or cannot be described by the independent variables *x*. The computation of multiple regression models is very complex. Since the error is not known only estimations of *y* are possible. As some variables are redundant or do not lead to an improvement of the model, they have to be eliminated. Feature selection is a powerful method to exclude redundant or irrelevant variables. Furthermore, it helps to avoid overfitting. One approach performing variable selection and regularization is the 'Least Absolute Shrinkage and Selection Operator' (LASSO) regression analysis which is based on L2 norm.

1.2.2 Modeling approaches to study TF-target gene interactions

Several modeling approaches to predict TF-target gene interactions are based on linear regression models and use binding data from ChIP-experiments or computational predictions as background. The most important regulators are identified by predicting the gene expression of a particular gene by the activities of the TFs. The activity of a TF is influenced more by posttranslational modifications and protein stability than by the gene expression value of the TF itself. Therefore,

several approaches infer the activity of a TF based on the expression of its target genes (Balwierz *et al.*, 2014; Schacht *et al.*, 2014). For several methods, the objective is to minimize the sum of errors between the measured and the predicted gene expression value over all samples. Schacht *et al.* uses a multivariate linear regression model to predict regulators of MITF in human melanoma cells using a Mixed Integer Linear Programming (MILP) based approach (Schacht et al. 2014, Kordaß et al. 2016). The 'Regression Analysis of Combined Expression Regulation' (RACER) method uses a two-stage regression model by integrating mRNA and miRNA expression data, copy number variations as well as DNA methylation data. In the first regression step, activity values for each sample are calculated and in the second step the regression coefficients from the first model are used to identify the TF/miRNA-target gene interactions by using a sparse LASSO approach. Applying this model to gene expression data of acute myeloid leukemia patients, Li and coworkers identified a pre-dominant list of 18 regulators that are linked to leukemogenesis (Li *et al.*, 2014). Another approach is the 'Regression Analysis with Background Integration' (RABIT). Comparable to RACER, RABIT uses gene expression data, somatic mutations, CNVs and DNA methylation data to identify regulators with differential expressed target gene in cancer (Jiang *et al.*, 2015). The 'Integrated System for Motif Activity Response Analysis' (ISMARA) infers the activity of TFs or miRNAs from motif binding information. The active TFs or miRNAs of a certain promoter are identified by combining the motif binding information with gene expression data using a linear model similar as Schacht *et al.* did (Balwierz *et al.*, 2014). Setty and coworkers used the above described linear model to identify the subtype specific regulators in glioblastoma multiforme (Setty et al. 2012), while Dong *et al.* focused on the relationship between chromatin features and expression levels using random forests as well as linear regression (Dong et al. 2012). Besides linear models there are also other approaches to predict TF-target gene interactions like e.g. Bayesian models. The 'Bayesian inference of context-specific regulator activities and transcriptional networks' (biRte) approach uses a probabilistic framework to integrate TF-target gene predictions and gene expression data. The active TFs are then identified by a maximum likelihood model (Frohlich, 2015). A distinctively different approach without any background knowledge is the 'Algorithm for the Reconstruction of Accurate Cellular Networks' (ARACNE). ARACNE uses pairwise mutual information (MI) to determine the relationships between a

predefined list of regulators and their target genes. In a bootstrapping process, MI networks are calculated from randomly sampled gene expression profiles and a consensus network is constructed considering Poisson distributions to estimate if a specific edge is detected significantly often over all runs (Lachmann *et al.*, 2016; Margolin *et al.*, 2006) without using any kind of binding information. The 'Virtual Inference of Protein activity by Enriched Regulon analysis' (VIPER) calculates the activity of regulators including TFs, co-factors and signaling molecules by using 'analytic rank-based enrichment analysis' (aREA). The target genes of the regulators are extracted from the ARACNE network and are ranked based on their relative and their absolute gene expression profile per sample. aREA tests if there is a global shift between the ranks in the relative compared to the absolute gene expression profile of the target genes (Alvarez *et al.*, 2016).

1.2.3 Identification of regulatory subnetworks

Most of the above described approaches address the co-operativity of the TFs and aim to predict TF target gene interactions, but they ignore TF-TF interactions as well as feedback loops. Ideker and coworkers showed that an integration of protein-protein interaction (PPI) or protein-DNA interaction networks and gene expression data can lead to active subnetworks. The regulators of these subnetworks significantly change their expression between subsets or conditions. For yeast, 5 significant subnetworks were identified which could explain more than half of all significant expression changes (Ideker *et al.*, 2002). Chuang and co-workers constructed subnetworks to identify new marker genes to distinguish between metastatic and non-metastatic breast cancer samples (Chuang *et al.*, 2007) and to predict the disease progression in chronic lymphocytic leukemia (CLL) patients (Chuang *et al.*, 2012). In both studies they combined a PPI network with gene expression profiles to map the gene to the corresponding protein. The activity of each subnetwork was defined as the average of the gene expression values of each patient. Differentially expressed subnetworks were identified based on mutual information between the activity values of the subnetwork and the disease state across all patients (Chuang *et al.*, 2007). For breast cancer and CLL new subnetwork markers could be identified, which improve the classification into metastatic or non-metastatic (Chuang *et al.*, 2007) as well as prediction of disease progression (Chuang *et al.*, 2012). Yosef and coworkers constructed regulatory

subnetworks from a PPI network based on a Steiner tree and a shortest path approach. Altogether, this approach was applied to link the genes that lead to telomere shortening or elongation when mutated (*TLM* genes) with the telomerase complex (anchor genes). The PPI network consisted of edges weighted based on their reliability. The goal was to identify a connected subnetwork which links the root (telomerase genes) and the terminals (*TLM* genes). First, the likelihood of the subnetwork was maximized by minimizing the sum of edge weights in the graph (Steiner tree problem). Second, the sum of the edge weights of the shortest paths were optimized. They discovered that the proteasome is an important feature in the regulation of the telomere length and is associated with transcription and DNA repair (Yosef *et al.*, 2009). These approaches are only based on PPIs and neglect the TF-target gene interactions, which are essential for regulation processes.

1.2.4 Mixed Integer Linear Programming

Linear Programming (LP) is a powerful method to describe and solve several optimization problems. The linear model is optimized by using several decision variables, an objective function as well as linear constraints. An example of an LP is given here:

$$\text{minimize} \quad 3\,x_1 + 5\,x_2 \tag{3}$$

$$\text{subject to} \quad x_1 + 4\,x_2 \geq 4 \tag{4}$$

$$2\,x_1 + 2\,x_2 \geq 5 \tag{5}$$

$$2\,x_1 + 2\,x_2 \geq 3 \tag{6}$$

$$\text{and} \quad x_1, x_2 \geq 0, \tag{7}$$

where equation (3) is the objective function and equations (4-7) are the constraints. The feasible solution of this LP can be graphically represented as a convex polyhedron in the *n* - dimensional space where *n* indicates the number of variables. The normal vector of the hyper planes goes through equation (3) resulting in a fixed point (Figure 2).

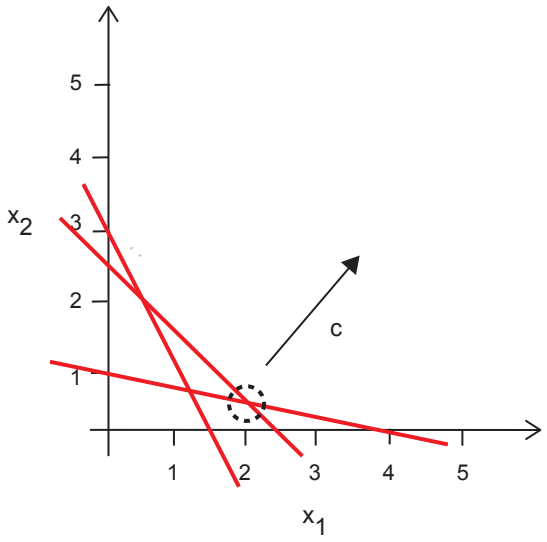

Figure 2. Graphical representation of the example LP

The convex polyhedron marks the solution space and is defined by the variables x_1 and x_2. Additional constraints lead to a further restriction of the solution space. The dashed circle indicates the solution at the extreme point. The optimal solution of this example is $x_1 = 2$ and $x_2 = 0.5$. This leads to an objective value of 8.5.

LP models can be generally written as:

$$\text{Objective function: } minimize\ c^T x \tag{8}$$

$$\text{Linear constraints: } Ax = b \tag{9}$$

$$\text{Boundaries: } I \leq x \leq u, \tag{10}$$

where *A* is a matrix, while *b* and *c* are vectors. $c^T x$ describes the linear combinations of parameters *c* and variables *x*. The objective is to minimize this linear combination during the optimization process (equation 8). Equation (9) described a matrix *A* with the linear constraints of *x*. The vector *I* indicates the lower bound and *u* the upper bound of *x* (equation 10). Compared to a pure LP, integer linear programming (ILP) describes a model in which all variables are integers and both the objective function, and the constraints are linear. In a Mixed Integer Linear Programming (MILP) binary, integer and continuous variables are allowed. MILP models need an additional integrality constraint (equation 11). The integrality variables x_i allow additional binary variables to capture discrete decisions.

$$\text{Integrality constraint: } some\ x_i\ must\ be\ integers \tag{11}$$

To solve MILP models, the branch-and-bound algorithm, the cutting-plane method or the branch-and-cut algorithm, which is a combination of the two other methods, can be used (Gurobi Optimization, 2018).

MILPs can have a high theoretical time complexity, but there exist very efficient solvers (e.g. Gurobi (http://www.gurobi.com/) and CPLEX (https://www.ibm.com/analytics/cplex-optimizer)). These solvers can find at least very accurate solutions within a given time limit by using the branch-and-cut algorithm.

MILP models can be applied to several different areas. The most prominent applications are the traveling salesman problem and the optimization of time tables. Biological applications are e.g. flux balance analysis with stoichiometric equations for thousands of metabolites and reactions (Orth *et al.*, 2010) or the optimization of cell-networks to identify distinctively expressed pathways (Schramm *et al.*, 2010). Furthermore, MILP can be used to infer gene regulation (Poos *et al.*, 2016; Schacht *et al.*, 2014), or gene regulatory modules in cell-networks (Beisser *et al.*, 2012). Also classification problems are applicable in form of e.g. support-vector machines (Saraiva *et al.*, 2017) or decision trees (Deeg *et al.*, 2017).

1.3 Telomere regulation

The ends of the eukaryotic linear chromosomes are protected by specific nucleoprotein structures called telomeres which consist of repetitive DNA sequences ending in a single-stranded 3’ G-rich overhang (Martinez and Blasco, 2011) and are well conserved between eukaryotes (Blackburn, 1990). They have two main functions: (i) to circumvent the end-replication problem and (ii) to protect the chromosomal ends for DNA-damage response (de Lange, 2009; O'Sullivan and Karlseder, 2010). These two functions are regulated by several proteins called shelterin complex binding to the telomeres (Okamoto and Seimiya, 2019). To protect the chromosomal ends, the single-stranded G-rich 3’ overhang invades into the double-stranded telomeric repeats forming a telomeric (t)-loop structure. At the invasion site a displacement (D-) loop is constructed (Martinez and Blasco, 2011). Telomeres share a common structure in *Saccharomyces cerevisiae* (Luke-Glaser

et al., 2012) and in mammals (de Lange, 2004). But the telomeric complex differs between *S. cerevisiae* and humans (de Lange, 2005), which is described in the following.

1.3.1 Telomere maintenance in yeast

S. cerevisiae is a well suited model organism to study telomere maintenance (Kupiec, 2014) because of its high homology to humans and its constitutively expressed telomerase (Ungar *et al.*, 2009). In *S. cerevisiae*, the telomeres are 300 ± 75 bp long (Wellinger and Zakian, 2012) and consist of repetitive 5'- $(TG_{1-3})_n$-3' DNA sequences (Taggart and Zakian, 2003). The telomeric complex of *S. cerevisiae* is formed by the telomerase, the CST complex, Ku complex, Rap-Rif1-Rif2 and the Sir complex (Figure 3). The telomerase includes the template RNA, TLC1, and the "Ever shorter telomere" proteins Est1, Est2 and Est3. The catalytic subunit is Est2, while Est1 and Est3 are associated proteins of TLC1 (Taggart and Zakian, 2003). Est1 can bind to the 3' single-stranded overhang, but also to different regions of TLC1. While the binding of Est1 is independent of Est2, Est3 can bind to TLC1 only by interacting to Est2 (Kupiec, 2014).

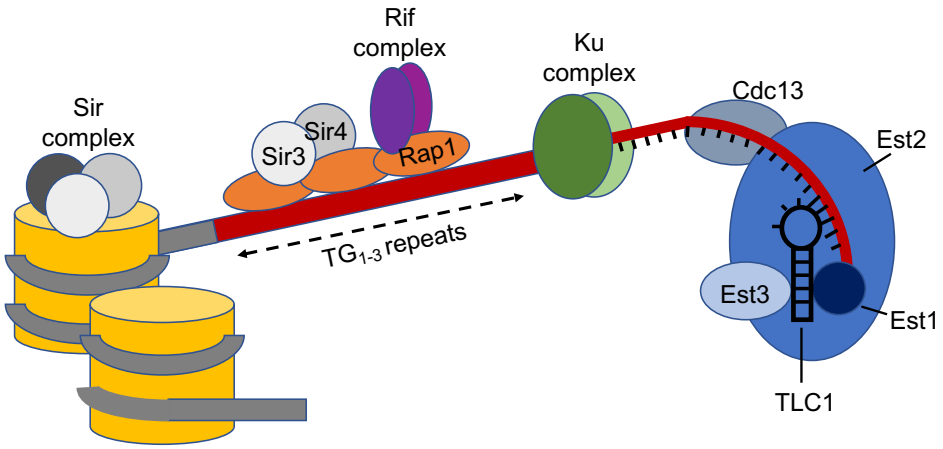

Figure 3. Telomeric complex and telomerase in *S. cerevisiae*.
Telomeric DNA (marked in red) consists of repetitive TG_{1-3} repeats. The telomerase contains the three *EST* genes and the template RNA TLC1. The capping protein Cdc13 binds to the single-stranded telomeric DNA to protect the telomeres from degradation, while the Ku complex protects for unwanted non-homologous end-joining. Rap1 binds to the double-stranded telomeric DNA and interacts with the Rif complex and sirtuin proteins. The number of bound Rap1 molecules is dependent on the number of TG_{1-3} repeats. In budding yeast, the nucleosomes (yellow) are only in the subtelomeric region (marked in grey).

Cdc13 (also called Est4) binds to telomeres in a sequence-specific manner and is involved in the capping process preventing the chromosomal ends from degradation. In addition, together with Est1 it recruits telomerase. Furthermore, Cdc13 forms a complex with Stn1 and Ten1 (CST complex) which is substantial for the telomeric capping process and protects from degradation by exonucleases and recombination processes (Teixeira, 2013). The Ku complex consists of the two proteins Yku70 and Yku80 and prevents the telomeres from unwanted non-homologous end-joining. The Ku proteins are further involved in TLC1 import in the nucleus. Together with the Sir complex they prevent damaged chromosomes and telomeres from Exo1 degradation (Kupiec, 2014). Rap1 is another essential protein of the telomeric complex binding to double-stranded telomeric DNA. Rap1 can interact with the gene silencing proteins Sir3 and Sir4 as well as the Rif proteins 1 and 2 (O'Reilly *et al.*, 1999). Both Rif proteins are also involved in the capping process (Teixeira, 2013). How many Rap1 molecules can bind to telomeres is dependent on the number of TG repeats (Bianchi and Shore, 2007) indicating that telomere length can be determined by counting the Rap1 molecules (Kupiec, 2014). It was shown that genes in close distance to telomeres undergo silencing which is called the telomere position effect (Pryde and Louis, 1999). Because of the heterochromatic structure in the subtelomeric regions, the promoter activity of the genes close to the telomeres is repressed, which can be up to 10-15 kilobases distant from the chromosome end. As Sir proteins play a role in gene silencing and interact with histones, they are recruited to telomeres via Rap1 (Kupiec, 2014). Nucleosomes can be found only in the subtelomeric region (Wright *et al.*, 1992). Telomere replication is cell cycle dependent and is coordinated with the replication of the whole genome. Because not all telomeres can be replicated in each cell cycle, short telomeres are preferred or in other words the telomeres with a low number of bound Rap1 molecules (Bianchi and Shore, 2007).

A deletion of the telomerase induces cellular senescence (Lundblad, 2002). In addition, it was shown that some yeast cells can survive without telomerase by using homologous recombination to maintain their telomeres. These cells are called survivors. They are either dependent on the RAD51 recombination pathway (type I survivor) or on RAD50 (type II survivor) (Chen *et al.*, 2001; Teng *et al.*, 2000). This recombination-dependent maintenance of the telomeres shows some similarities to the human ALT pathway (Lundblad, 2002).

TLM gene screening: Yeast is a unicellular organism with 6,275 open reading frames (ORFs). Some years ago, Winzeler et al. generated a collection of 4,700 yeast deletion strains by systematically deleting each non-essential gene individually (Winzeler *et al.*, 1999). This collection was extended by two additional studies where all 1,300 essential genes were deleted either hypomorphic (Breslow *et al.*, 2008) or both alleles of a gene were made temperature-sensitive (Ben-Aroya *et al.*, 2008). These mutant collections were then used for a genome-wide screening of genes involved in telomere maintenance. Askree *et al.*, Gatbonton *et al.* , Meng *et al*. and Ungar *et al.* measured the telomere length of each yeast mutant by Southern blot to identify genes whose corresponding deletion strain showed shorter or longer telomeres than the wild-type length of 350 bp (Askree *et al.*, 2004; Gatbonton *et al.*, 2006; Meng *et al.*, 2009; Ungar *et al.*, 2009). These genes were then called telomere maintenance genes (*TLM*) and their mutants *tlms*. Ben-Shirit *et al.* analyzed if the phenotype in a deletion strain of gene X is changes due to the effect of the neighbouring gene Y using the 'NGE inference via a network-based approach' (NIRVANA) (Ben-Shitrit *et al.*, 2012). Additionally, Shachar *et al.* used the *TLM* genes from (Askree *et al.*, 2004; Gatbonton *et al.*, 2006) for which PPI data was available to build a telomere length regulating subnetwork (TRS). As end-point of their TLM-related signaling pathways they defined the genes of the telomerase subunits as well as telomerase-interacting proteins. They identified TLM pathways that link *TLM* genes to telomere-binding proteins leading to a TRS with 327 proteins. A subset of the analyzed 180 *TLM* genes was found between the telomere-binding proteins and other TLM proteins, while most of the non-TLM proteins were necessary to connect the TLM proteins with the end-points (Shachar *et al.*, 2008).
In summary, systematic screens together with additional computational studies identified around 500 genes that effect telomere length when mutated (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Meng *et al.*, 2009; Shachar *et al.*, 2008; Ungar *et al.*, 2009), which is around 8% of the yeast genome. From the 500 *TLM* genes around 60% result in shorter telomeres and 40% in longer telomeres when mutated. Most of these proteins were not known before to play a role in homeostasis of telomere length, they are localized to several cell compartments and have many different biochemical functions. Still, because many of them are evolutionally conserved and have human orthologs, *TLM* genes could be interesting anticancer targets.

1.3.2 Telomere maintenance in humans

In humans, telomeres of somatic cells have a length of 10 to 15 kb (Palm and de Lange, 2008) and consist of repetitive 5'- $(TTAGGG)_n$ -3' sequences (Moyzis *et al.*, 1988). Telomerase contains the protein subunit *TERT* and the template RNA *TERC* (Figure 4) (Sandin and Rhodes, 2014) as well as accessory proteins like dyskerin (DKC1), TCAB1, NHP2, NOP10 and GAR1 (Shay and Wright, 2019).

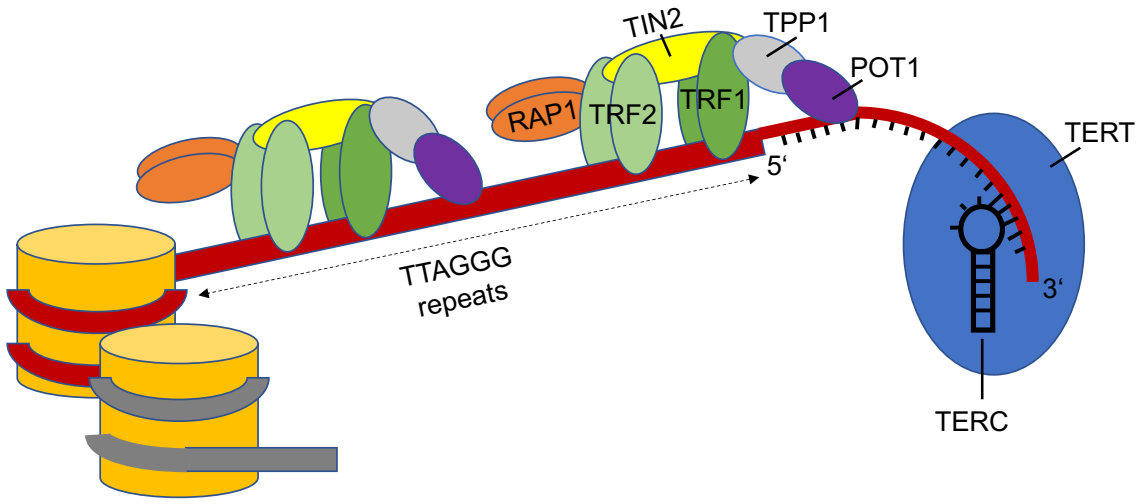

Figure 4. Telomerase and shelterin complex in human cells.
Telomeres contains a double-stranded region of TTAGGG repeats and a 150-200 nucleotide long single-stranded overhang of the G-rich strand. Telomerase consists of the protein subunit *TERT* and the template RNA *TERC*. The shelterin complex is built by the telomeric repeat binding factors (TRF) 1 and 2, repressor-activator protein (RAP) 1, protection of telomeres protein (POT) 1, TRF1-interactiong protein (TIN) 2 as well as the TIN2- and POT1-interacting protein (TPP1). TRF1 and TRF2 directly bind to the double-stranded telomeric repeat DNA, while POT1 binds to the single-stranded overhang. TIN2 binds to TRF1 and TRF2 recruiting the TPP1-POT1 complex.

TERC is constitutively expressed, but the *TERT* gene is silenced in somatic cells (Feng *et al.*, 1995; Kim *et al.*, 1994). *TERT* is only expressed during development, in germ and stem cells as well as in most cancer cells (Gaspar *et al.*, 2018). To elongate the telomeres, the telomerase complex recognizes the hydroxyl group (OH) at the 3'-end of the single-stranded overhang (Martinez and Blasco, 2011). Telomerase has been observed the shortest telomere (Hemann *et al.*, 2001). The shelterin complex consists of the following six components (Figure 4): the telomeric repeat binding factor I and II (TRF1, TRF2), TRF1-interacting protein II (TIN2),

protection of telomeres protein I (POT1), TIN2- and POT1-interacting protein (TPP1) and repressor/activator protein I (RAP1) (Martinez and Blasco, 2011; Palm and de Lange, 2008). Both TRF1 and TRF2 bind to the double-stranded telomeric DNA as homodimers and form a scaffold for the complex structure. Furthermore, TRF2 is involved in t-loop formation (Okamoto and Seimiya, 2019). TIN2 interacts with TRF1 and TRF2, while RAP1 can only bind to TRF2. TIN2 binds TPP1 recruiting POT1 to the telomeres. Compared to the other factors POT1 binds to the single-stranded telomeric DNA, which is 150-200 nucleotides long, at the 3' overhang in the loops (Palm and de Lange, 2008). TRF2 and POT1 prevent the telomeres from recombination processes (de Lange, 2009) and DNA damage response (Martinez and Blasco, 2015). TPP1 is also involved in the recruitment of telomerase to single-stranded telomeric DNA (Wang *et al.*, 2007; Xin *et al.*, 2007).

1.4 Telomere maintenance in cancer

In principle, telomeres function as a 'molecular clock'. Due to the end-replication problem and degradation they shorten with each cell division by 50-200 bp therefore limiting the life span of each cell (Levy *et al.*, 1992). The end-replication problem describes the incomplete replication of chromosome ends because DNA polymerase can only work from 5' to 3' direction and needs a primer antisense to the 3' end for starting the DNA synthesis (Martinez and Blasco, 2015). It was shown that after 50 to 60 cell cycles telomeres get critically short and replicative senescence or apoptosis is induced by p53 activation and chromosomal instability (Hayflick, 1965; Vaziri *et al.*, 1997; Wright *et al.*, 1996). This process prevents the cells from indefinite proliferation and outgrowth of abnormal cells (Blasco, 2005; Shay & Wright, 2000). Critically short telomeres become unprotected and Ataxia Telangiectasia Mutated (ATM) and Ataxia Telangiectasia and Rad3-Related Protein (ATR)- dependent DNA damage cascades are induced (d'Adda di Fagagna *et al.*, 2003; Kaul *et al.*, 2011; Takai *et al.*, 2003). Checkpoint kinases (CHK) 1 and 2 mark the uncapped ends by telomere induced foci (TIF) (Takai *et al.*, 2003). In primary fibroblasts, it was reported that only 5 dysfunctional telomeres are sufficient to induce replicative senescence (d'Adda di Fagagna *et al.*, 2003; Hayflick and Moorhead, 1961; Kaul *et al.*, 2011). However, some cells accumulating genetic

mutations in p53 or other checkpoint proteins can overcome senescence and proliferate further (Chin *et al.*, 1999; Preto *et al.*, 2004). Only a minor fraction of these cells acquire immortality and later carcinogenesis (Chin *et al.*, 1999). These cancer cells circumvent this safeguarding mechanism and develop ways to activate a telomere maintenance mechanism (TMM) (Gunes and Rudolph, 2013). Since the activation of TMM enable replicative immortality, it is a hallmark of cancer (Hanahan and Weinberg, 2011).

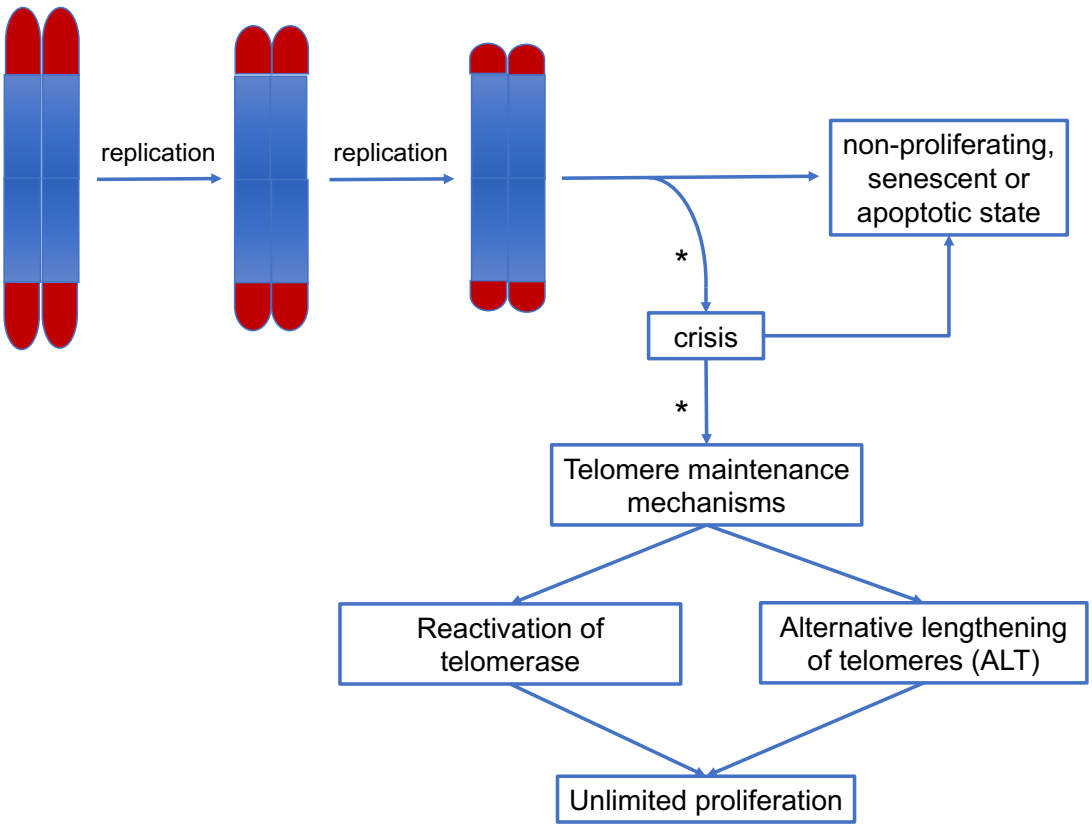

Figure 5. Active TMM leads to unlimited proliferation.
In somatic cells the telomeres shorten with each cell division and if they get critically short, the cell goes into senescence or apoptosis. Due to (epi)genetic aberrations (*) cells can escape senescence and further telomere shortening induces crisis. An accumulation of these aberrations can lead to malignant transformation into cancer cells and an activation of telomere maintenance mechanisms. Mostly, the telomerase is reactivated to elongate the telomeres. A minor fraction of cells maintains the telomeres by using the ALT pathway.

Telomere maintenance can be activated either by re-expression of telomerase, a reverse transcriptase (Greider and Blackburn, 1985), or by a telomerase-independent mechanism based on recombination processes termed alternative lengthening of telomeres (ALT) (Dunham *et al.*, 2000; Greider and Blackburn, 1985) (Figure 5). In 85-90 % of all cancers and in more than 70% of immortalized

human cell lines telomerase is expressed, while a subset of the remaining cancers utilize an ALT mechanism (Sobinoff and Pickett, 2017). Cancer entities with a high prevalence of ALT are e.g. sarcomas (64%), astrocytoma (63%) and pediatric glioblastoma multiforme (44%) (Chudasama *et al.*, 2018; De Vitis *et al.*, 2018; Heaphy *et al.*, 2011). In turn, there are also other cancer entities for which no ALT case has been reported so far. These are adenocarcinomas of the prostate, colon, stomach, pancreas and small intestine (Heaphy *et al.*, 2011). This shows that ALT occurs more often in tumors with neuroepithelial or mesenchymal origin (Heaphy *et al.*, 2011). It has been reported that there are also cell lines and primary tumors without telomere maintenance mechanisms (Dagg *et al.*, 2017). These are assigned to the term “ever shorter telomeres”. Their telomeres are very long and shorten with each cell division (Dagg *et al.*, 2017). It is assumed that telomere extension has been involved at some early time point and was deactivated at a later stage.

1.4.1 Re-activation of telomerase

As *TERT* is the limiting factor of telomerase activity (Feng *et al.*, 1995; Kim *et al.*, 1994), its activation is a highly interesting research topic to understand cancer development. The *TERT* gene is located on chromosome 5 and contains 16 exons (Cong *et al.*, 1999; Wick *et al.*, 1999). There exist several different splice transcripts of *TERT*, but only the full-length transcript shows reverse transcriptase activity (Saeboe-Larssen *et al.*, 2006). The core region of the *TERT* promoter is located between 330 bp upstream and 228 bp downstream of the transcription start site. The promoter of *TERT* is guanine-cytosine (GC) rich and lacks both a TATA box and a CAAT box (Cong *et al.*, 1999; Horikawa *et al.*, 1999; Takakura *et al.*, 1999). Several TF binding sites have been recognized in this region. TFs that have been reported as activators of *TERT* are e.g. c-Myc, the NF-κB complex, the STAT factors 3 and 5 as well as PAX5 and PAX8 or the estrogen receptor (Ramlee *et al.*, 2016). Transcriptional repressors of *TERT* include MAD1, SP3, CTCF, E2F1 and AR (Ramlee *et al.*, 2016). TFs like AP-1, EGR1, HIF-2A or SP1 have been identified as activators as well as inhibitors of *TERT*. For example, SP1 can have an activating role of *TERT* expression in telomerase-positive cells and a repressive role in telomerase-negative cells (Ramlee *et al.*, 2016). Furthermore, it has been reported that the methylation level of the *TERT* promoter region is correlated to *TERT* expression (Devereux *et al.*, 1999).

The mode of *TERT* activation differs between cancer entities. It has been reported that amplifications, rearrangements or promoter mutations can lead to an upregulation of *TERT*. *TERT* amplifications have been detected in ovarian cancer, lung adeno- and squamous cell carcinoma, oesophageal as well as adrenocortical carcinoma (Barthel *et al.*, 2017). *TERT* promoter rearrangements have been identified in high-risk neuroblastomas (Peifer *et al.*, 2015). An upregulation of *TERT* through non-coding mutations in the core promoter region was first reported in cutaneous melanoma. These C>T transitions occur at position 124 bp (C228T) or 146 bp (C250T) upstream of the translation start codon (ATG). These mutations are early events in tumor development (Shain *et al.*, 2015). Both mutations create a de-novo ETS binding site (CCGGAA/T binding motif) and a two- to four-fold higher *TERT* promoter activity in melanoma cells (Horn *et al.*, 2013; Huang *et al.*, 2013). It has been reported that both mutations are functionally distinct. C228T lead to a recruitment of the ETS family member GABPA, which cannot bind to the native *TERT* promoter (Mancini *et al.*, 2018). C250T generates an ETS binding site and a functional p52 site which requires ETS1 and ETS2 (Li *et al.*, 2015). In melanoma and thyroid cancer, *TERT* promoter mutations often co-occur with mutations in the BRAF gene (V600E) (Horn *et al.*, 2013; Huang *et al.*, 2013; Okamoto and Seimiya, 2019; Rusinek *et al.*, 2018).

Besides cutaneous melanoma, *TERT* promoter mutations frequently occur in cancers of the central nervous system (primary glioblastoma and oligodendroglioma), urothelial carcinoma, follicular cell-derived thyroid cancer and hepatocellular carcinoma as well as cutaneous basal and squamous cell carcinomas (Vinagre *et al.*, 2013). Glioma, melanoma and thyroid cancer patients with a *TERT* promoter mutation have a poorer prognosis (Batista *et al.*, 2016; Griewank *et al.*, 2014; Kim *et al.*, 2016) indicating that *TERT* promoter mutations can be used as clinical biomarker.

1.4.2 Alternative lengthening of telomeres (ALT) pathway

The exact mechanism of ALT is not yet fully understood. However, several molecular and cytological features have been identified that differ between telomerase-positive and ALT-positive, which are telomerase-negative, samples (Cesare and Reddel, 2010; Henson and Reddel, 2010) (Figure 6).

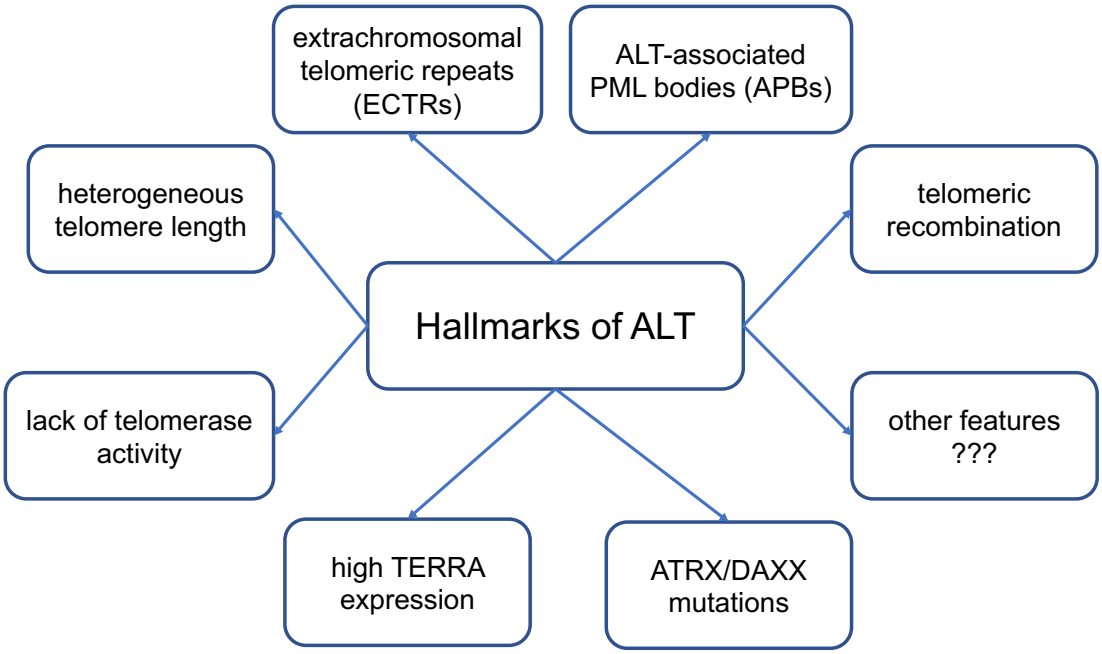

Figure 6. Hallmarks of ALT

Several different hallmarks of ALT have been identified so far. These hallmarks can be studied by cytological and molecular assays as well as sequencing readouts. But the exact mechanism of ALT is still unknown.

ALT cells typically show

(i) heterogeneous telomere lengths (Henson and Reddel, 2010). Telomerase-positive cells have telomeres with small range around the average length of usually 5 to 10 kb, while the telomere length of ALT cells spreads from undetectable by fluorescence in situ hybridization (FISH) to more than 50 kb. Another method for absolute telomere length detection is the telomere restriction fragment (TRF) analysis (Mender and Shay, 2015).

(ii) lack of telomerase activity, which can be measured by the telomere repeat amplification protocol (TRAP) assay (Mender and Shay, 2015). Furthermore, the expression of *TERT* can be detected by RNA-seq. Since *TERT* transcript levels are very low, *TERT* RNA-Seq has to interpreted with caution (Ducrest *et al.*, 2001; Fredriksson *et al.*, 2014).

(iii) high levels of extra-chromosomal telomeric repeats (ECTRs) (Cesare and Griffith, 2004; Tokutake *et al.*, 1998), which can be either linear or circular, double-stranded or single-stranded, G- or C-rich DNA fragments (Cesare and Reddel, 2010; Henson and Reddel, 2010). C-circles are a well-established

marker of ALT cells and can be detected by the so-called C-circle assay (Henson *et al.*, 2009; Henson *et al.*, 2017).

(iv) telomeric recombination visualized by chromosome orientation FISH (Nabetani and Ishikawa, 2011). These recombination processes can take place between sister chromatids or telomeres of different chromosomes. In addition, ALT cells allow homologous recombination processes at telomeres which are typically repressed by the shelterin complex (Lovejoy *et al.*, 2012).

(v) colocalizations of telomeres with promyelocytic leukemia (PML) bodies, the so-called ALT-associated PML bodies (APBs) (Chung *et al.*, 2012; Yeager *et al.*, 1999). PML bodies are present in normal cells, however the colocalization in telomeres is unique to ALT cells (cite some ALT paper). The average number of APBs differs between ALT cells, but most ALT cells have at least one APB (Henson and Reddel, 2010; Osterwald *et al.*, 2015).

(vi) high levels of the telomeric repeat-containing RNA (TERRA). TERRA is transcribed from the subtelomeres into the telomeres and was reported to inhibit telomerase (Ng *et al.*, 2009).

(vii) recurrent mutations in the alpha-thalassemia mental retardation X-linked protein (ATRX), the death-associated protein 6 (DAXX) (Heaphy *et al.*, 2011; Lovejoy *et al.*, 2012; Schwartzentruber *et al.*, 2012). Mutations in ATRX and DAXX are typically mutually exclusive (Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2012), also to *TERT* promoter mutations (Okamoto and Seimiya, 2019). However, it is not possible to induce ALT by knocking down any of these genes (Flynn *et al.*, 2015; Lovejoy *et al.*, 2012). However, ATRX loss can enable the activation of ALT (Napier *et al.*, 2015). Furthermore, it was reported that in an ATRX-negative cell line ALT can be suppressed by reexpression of ATRX (Clynes *et al.*, 2015; Deeg *et al.*, 2016), while inducing hTERT in ALT cells led to a repression of at least some ALT features (Perrem *et al.*, 1999).

(viii) mutations in the H3 histone family member 3A (H3F3A) often co-occur with ATRX mutations in pediatric glioblastoma samples (Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2012). Furthermore, it was reported that in HeLa cells with long telomeres ALT can be induced by a knockdown of the anti-silencing factor 1 (ASF1) (O'Sullivan *et al.*, 2014). Furthermore, it was reported that

the activation of the ALT pathway is independent of the expression level of *TERC* (Zhang *et al.*, 2011).

Currently, C-circles are one of the most reliable molecular markers for ALT. Since C-circles are prone to degradation, a lack of C-circle signal can be interpreted as false-negative. Therefore, a combination of the C-circle assay with the analysis of different other features is necessary to get a reliable TMM prediction.

1.5 Cancer entities analyzed in this study

1.5.1 Prostate cancer as a cancer entity with neither *TERT* promoter mutations nor ALT occurrence

After lung cancer (14.6 %), prostate cancer is with 13.5 % the second most common cancer type in men worldwide and it is the fifth most frequent leading cause of cancer death in men (6.7 %) (Bray *et al.*, 2018). For 2018 the International Agency for Research on Cancer listed around 1.3 million new cases (Bray *et al.*, 2018). Risk factors for a high incidence of prostate cancer are age, genetic susceptibility, family history as well as race (Al Olama *et al.*, 2014; Cancer Genome Atlas Research, 2015). The established screening of the prostate specific antigen (PSA) level considerably increases the potential of early diagnosis (Penney *et al.*, 2013). PSA is produced by the prostate and can be measured from the blood. A high level of PSA in the blood (> 4 ng/ml for 60-69 old men) is an indicator of prostate cancer and consequently a biopsy is highly recommended (Pezaro *et al.*, 2014). The PSA level in serum correlates with the pathological stage of the tumor and the Gleason Score. It was furthermore shown that a PSA-level of > 50 ng/ml is correlated with a poorer response to treatment and survival compared to other high-risk patients (Koo *et al.*, 2015). The PSA level is also used subsequent to a radical prostatectomy to determine the tumor recurrence (PSA-recurrence free survival) and is often used instead of the overall survival. After a radical prostatectomy the PSA level is low, but upon tumor recurrence the PSA-level highly increases again. Most of the patients have an indolent form of prostate cancer and are curable, while others have more aggressive cancer leading to metastasis and death (Al Olama *et al.*, 2014; Cancer Genome Atlas Research, 2015). Potential therapy options are radical

prostatectomy, radiation or brachy therapy or in some cases also active surveillance can be possible. For patients with metastasis also drug treatment is necessary, like chemotherapy (Docetaxel, Cabazitaxel) or androgen receptor inhibitors (Enzalutamid) (Board., 2002). The Gleason score describes the histological differences between healthy prostate cells and prostate cancer cells. There are five different grades (1-5) and the higher the grade the more aggressive is the tumor. The pathologist adds up the most frequent and the second most frequent pattern when regarding slides of tumor tissues. The poorest pattern is mentioned in addition, if it is not identical to the most frequent or the second most frequent pattern. Tumors with a Gleason score of 4+3, 4+4 or even higher are highly aggressive (Helpap and Egevad, 2006). The risk stratification system today combines Gleason score, pre-operative PSA-levels, and pathological as well as clinical staging, but it cannot adequately predict the patient's outcome (Cooperberg *et al.*, 2009). Previous studies of primary prostate cancer patients identified several recurrent genomic alterations like mutations, gene fusions, DNA copy-number changes, and rearrangements. The most common alteration is the *TMPRSS2-ERG* fusion (Tomlins *et al.*, 2005). This gene fusion was found mainly in early-onset prostate cancer patients maybe because of the increased androgen signaling in younger men (Weischenfeldt *et al.*, 2013). But a follow up study revealed that the age dependency of the ERG-fusions in limited to patients with a low Gleason score (≤ 3+4) (Steurer *et al.*, 2014). ERG-fusion negative patients showed an age-dependent accumulation of chromosomal deletions (Weischenfeldt *et al.*, 2013). *SPOP*, *TP53*, *FOXA1* as well as *PTEN* have been identified to be most frequently mutated in prostate cancer (Barbieri *et al.*, 2012). The identification of new biomarkers is needed to improve the risk stratification, the progression of the disease as well as therapy decisions.

1.5.2 Glioblastoma multiforme as a cancer entity with a high rate of ALT occurrence

Glioblastoma multiforme (GBM) is the most frequent malignant primary brain tumor in children and adults. GBMs are highly aggressive and are classified as stage IV in the World Health Organization scheme (Sturm *et al.*, 2014). With a median survival of less than 9 months, GBM patients have a very poor prognosis (Hakin-Smith *et al.*, 2003). Patients with ALT-positive GBM had an increased median survival rate (Hakin-Smith *et al.*, 2003; McDonald *et al.*, 2010), which indicates that ALT might

be an interesting prognostic marker for GBM. Pediatric (ped) and adult GBM both have a very poor prognosis, but recent studies showed differences regarding gene expression signatures, genetic mutations and DNA copy number (Gilheeney and Kieran, 2012; Sturm *et al.*, 2014). Mutations in ATRX and H3F3A occur more often in pedGBM than in adult GBMs (Heaphy *et al.*, 2011; Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2014). Mutations in H3F3A are highly specific to pedGBMs and often co-occur with mutations in TP53 and ATRX (Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2012). H3F3A encodes the histone variant H3.3, which is enriched in transcriptionally active regions, but also at telomeres and pericentromeres. A recurrent mutation in H3F3A affected mutually exclusive either an amino acid change in H3.3 of lysine 27 to methionine (K27M) or glycine 34 to arginine respectively valine (G34R/V (Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2012). Furthermore, pedGBM show a much higher frequency of ALT (44 %) compared to adult GBM (14 %) (Heaphy *et al.*, 2011). This together with the high frequency of ATRX mutations and H3F3A mutations makes pedGBM an interesting entity for studying the ALT mechanism.

1.6 Clinical aspects

1.6.1 Patient stratification

A deregulation of TFs or a mutation in their binding sites drive several human diseases, e.g. cancer (Lambert *et al.*, 2018) indicating that TFs are well suited as prognostic markers. Gene regulatory models are important to identify the TF to target gene interactions. Furthermore, the stratification of patients into molecular subtypes are essential to identify the exact regulatory mechanisms and can improve targeted therapies. One example for an established patient stratification system is a classifier based on methylation data for tumors of the central nervous system (Capper *et al.*, 2018).

In addition, a patient stratification based on the active telomere maintenance mechanism can also improve the prognostic information in several cancer types (Elkak *et al.*, 2006; Hakin-Smith *et al.*, 2003; Lundberg *et al.*, 2011; Poremba *et al.*, 2002). For instance, Lee *et al.* used whole genome sequencing (WGS) data to perform a quantitative telomere repeat variant analysis (Lee *et al.*, 2018). Besides

the canonical TTAGGG telomeric repeat there also exist repeat variants which differ by one nucleotide, e.g. TCAGGG, TGAGGG or TTGGGG. This telomere repeat variant content was then used to build a classifier to distinguish between ALT-positive and ALT-negative tumor samples. Sieverling *et al.* developed a classifier for the PanCancer Analysis of Whole Genomes (PCAWG) project to predict the active TMM of tumor samples with an ATRX/DAXX mutation (potentially ALT-positive) and with *TERT* expression by including measures for telomere content, the number of telomere insertions, the number of telomeric repeats of TGAGGG, TCAGGG, TTGGGG, TTCGGG and TTTGGG (identified with Telomere Hunter (Feuerbach *et al.*, 2016)) (Sieverling *et al.*, 2019).

Furthermore, *TERT* promoter mutations are the most common non-coding somatic mutations in cancer (Okamoto and Seimiya, 2019). As patients with a *TERT* promoter mutation show a poorer prognosis (Batista *et al.*, 2016; Griewank *et al.*, 2014; Kim *et al.*, 2016), the mutation status of the *TERT* promoter can be used as biomarker for patient stratification. In addition, a hypermethylation of the CpG (cg11625005) upstream of the transcription start site (TSS) of *TERT* resulted an expression of *TERT*. Since the re-activation of *TERT* is associated with tumor progression as well as a poor prognosis in pediatric brain tumors (Castelo-Branco *et al.*, 2013), the *TERT* CpG methylation pattern is also a potential biomarker.

1.6.2 Targeting telomere maintenance

As telomere maintenance is a hallmark of cancer and inactive in somatic cells, it is an interesting target for cancer therapy. Because most of the cancer cells maintain their telomeres by re-activating the telomerase, inhibition of telomerase seems to be an attractive strategy. So far, immunotherapy, gene therapy, small molecular inhibitors and G-quadruplex ligands have been developed. Some of them have also entered clinical trials (Chiappori *et al.*, 2015; Roth *et al.*, 2010; Ruden and Puri, 2013; Williams, 2013). For instance, the antisense oligo GRN163L (Imetelstat) targets *TERC* (Roth *et al.*, 2010), while for *TERT* e.g. the small-molecule inhibitor BIBR1532 has been developed (Pascolo *et al.*, 2002). However, there is no approved inhibitor for telomerase so far. One reason for this could be the long lag time between inhibition and apoptosis. Since it can take up to ~21-24 cell divisions, equal to 20 months (Uziel *et al.*, 2015), until telomeres reach a critically short length that can induce replicative senescence or apoptosis (Min *et al.*, 2017), the tumor

might have multiplied its size in the meantime. Furthermore, repression of one TMM, e.g. by inhibition of telomerase could result in activation of another TMM, e.g. the ALT pathway (Hu *et al.*, 2012; Kelland, 2005; Shay and Wright, 2019; Villa *et al.*, 2000). Therefore, the most effective treatment would inhibit both telomerase and the ALT pathway (De Vitis *et al.*, 2018). However, a targeted ALT therapy is lacking, because the complete mechanism is still unknown.

Surprisingly, 70 % of cancer cells have shorter telomeres as compared to normal cells and the remaining cells may have activated the ALT pathway (Barthel *et al.*, 2017). In principle, telomerase inhibitors should be more efficient in cancer samples with shorter telomeres as they induce senescence or apoptosis earlier. This shows that telomere length can act as a predictive marker (Frink *et al.*, 2016; Fujiwara *et al.*, 2018). A clinical study reported that Imetelstat treatment results in elongation of both the median progression-free and the overall survival of non-small lung cancer patients with short telomeres (Chiappori *et al.*, 2015). Shorter telomeres correlate also with malignancy. But there are also studies showing that genes involved in cancer progression, e.g. Interferon Stimulated Genes (ISGs), are activated due to maintenance of short telomeres. Telomere length is negatively correlated with a high activity of telomerase or a high expression of *TERT* or the shelterin genes (Butler *et al.*, 2012; Hu *et al.*, 2010). Limited telomere maintenance can also be explained by the ‘protein-counting’ mechanism. This means that there is a negative feedback loop of the telomere-bound shelterin complexes and the accessibility of telomerase (Marcand *et al.*, 1997; Shore and Bianchi, 2009) balancing the effects of the repressors with the length of the telomeres.

Scope of the thesis

All biological processes and signaling pathways in the cell are controlled by a complex network of transcriptional regulators. Typically, several regulators collaboratively regulate the expression of a gene by binding to its promoter or enhancer regions (Bauer *et al.*, 2011; Cheng *et al.*, 2012; Consortium *et al.*, 2009; Dong *et al.*, 2012; Oliveira *et al.*, 2008; Schacht *et al.*, 2014; Setty *et al.*, 2012). To address this additive co-operativity and to identify the regulatory interactions between TFs and their target genes, several methods have been developed. Some are based on linear regression (e.g. ISMARA (Balwierz *et al.*, 2014)), while others use a probabilistic framework or mutual information (e.g. ARACNE (Lachmann *et al.*, 2016)). None of these methods uses the potential of Mixed Integer Linear Programming (MILP) allowing to study the regression in L1 norm avoiding overestimating outliers and to implement constraints to get sparse models. Deregulation of transcriptional regulators can often cause disease, including cancer, making TFs putative drug targets, even if their exact mechanism of action is not known.

To improve prediction of regulatory processes, I developed the 'Mixed Integer linear Programming based Regulatory Interaction Predictor' (MIPRIP) (Poos *et al.*, 2019; Poos *et al.*, 2016) which can identify the most important regulators of a particular gene in a dataset. MIPRIP was further extended with

(i) a statistical downstream analysis to study regulatory processes between two or multiple datasets/conditions and

(ii) a modularity-based approach to identify the regulators that may not bind directly to the promoter of the gene of interest, but are collaterally involved in the regulation of the gene by interacting with other regulators.

One important hallmark of cancer cells is the ability of unlimited proliferation. The activation of a telomere maintenance mechanism (TMM) is crucial for cancer cells to enable replicative immortality (Hanahan and Weinberg, 2011). Telomeres are repetitive DNA sequences at the ends of eukaryotic chromosomes and they shorten gradually with each cell division eventually triggering replicative senescence or apoptosis. Compared to somatic cells, cancer cells overcome this restriction mainly by extension of their telomeres by either re-expressing the reverse transcriptase

telomerase or by using a mechanism based on homologous recombination called alternative lengthening of telomeres (ALT). Furthermore, mutations in the promoter of the telomerase reverse transcriptase (*TERT*) gene can also lead to an upregulation of *TERT* (Horn *et al.*, 2013; Huang *et al.*, 2013). The underlying regulatory processes of telomere maintenance are only partly understood. Besides this, it was shown that patient stratification according to TMM can improve the prognostic value for survival, e.g. ALT-positive glioblastoma patients showed a better prognosis compared to ALT-negative patients (McDonald *et al.*, 2010). Thus, understanding the mechanisms that maintain telomere length can have substantial medical implications, in particular for ageing and carcinogenesis.

In this thesis, I used gene regulatory network models to study telomere maintenance in the model organism *Saccharomyces cerevisiae* and in human cancer. To get a better understanding of the mechanism underlying telomerase re-activation, MIPRIP was applied to study the regulation of the telomerase, first in *S. cerevisiae* and second in 19 different cancer types. For the cancer types, I especially focused on melanoma skin cancer, an entity with a high fraction of *TERT* promoter mutations, and prostate cancer, a cancer type without *TERT* promoter mutations or ALT occurrence. Compared to many other cancer entities, pediatric glioblastoma (pedGBM) exhibits a high occurrence of ALT (Heaphy *et al.*, 2011) and is therefore well suited for the study of molecular differences between ALT and telomerase-positive cancers. Here, I developed a classification scheme to predict the active TMM (ALT-positive or ALT-negative) in pedGBM tumor samples based on typical ALT features extracted from sequencing data as well as cytological and molecular assays. After the classification of the pedGBM samples, the aim was to identify a gene expression and transcription factor activity signature for ALT-positive cancers that explain the differences in *TERT* activity between the different TMMs.

In summary, this thesis aimed to develop and apply new approaches based on Mixed Integer Linear Programming to identify the regulatory mechanisms of telomere maintenance in yeast as well as in different cancer entities, especially melanoma skin cancer, prostate cancer and pedGBM.

2 Materials and Methods

2.1 Software

During my PhD, I used the following software for data analysis and modeling (Table 1):

Table 1. Software used for the analysis

Software	Reference	Version
R	https://www.r-project.org/	3.5.1
RStudio	https://www.rstudio.com/	1.2.907
Gurobi	http://www.gurobi.com/	7.0.1
ARACNe-AP	(Lachmann *et al.*, 2016), https://github.com/califano-lab/ARACNe-AP	
Cytoscape	(Shannon *et al.*, 2003), https://cytoscape.org/	3.5.1
VIPER	(Alvarez *et al.*, 2016), doi.org/10.18129/B9.bioc.viper	3.8
gProfileR	(Reimand *et al.*, 2016), https://cran.r-project.org/web/packages/gProfileR/index.html	0.6.6
VennDiagram	https://CRAN.R-project.org/package=VennDiagram	1.6.20
DESeq2	(Love *et al.*, 2014), https://bioconductor.org/packages/release/bioc/html/DESeq2.html	1.16.1
MIPRIP	https://www.leibniz-hki.de/en/miprip.html; https://github.com/network-modeling/MIPRIP; (Poos *et al.*, 2019; Poos *et al.*, 2016)	1.0-2.0

2.2 Datasets

2.2.1 Yeast data

To model the regulation of the telomerase in *S. cerevisiae*, gene expression data was used together with annotation data of *TLM* genes.

Expression dataset: Pre-processed microarray gene expression data (Reimand *et al.*, 2010) of 269 yeast regulator deletion strains (strains BY4741, S288C and BYTET) was downloaded from Array Express (E-MTAB-109, www.ebi.ac.uk/arrayexpress/). Reimand *et al*. re-analyzed the dataset containing 588 two-color cDNA microarray hybridizations of 269 regulator mutants against a

reference sample, generated by Hu and coworkers (Hu *et al.*, 2007). All probes that could not be annotated as open reading frames were filtered out and duplicated or triplicated probes were averaged. The pre-processed dataset contains expression values of 6,253 protein-coding genes and was normalized using variance stabilization (VSN) (Huber *et al.*, 2002; Reimand *et al.*, 2010). For the modeling approach, a z-score transformation for each gene across all samples was performed. For this the mean of the gene expression values $\bar{g}_i$ of gene *i* over all samples was subtracted from each gene expression value g_{ik} of gene *i* and sample *k* divided by the standard deviation σ (equation 12),

$$z_{ik} = \frac{g_{ik} - \bar{g}_i}{\sigma_{g_i}}. \tag{12}$$

Annotation of *TLM* genes: Based on genome-wide screening and computational analysis around 500 genes were identified that lead to telomere shortening or telomere extension when mutated (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Shachar *et al.*, 2008; Ungar *et al.*, 2009). These genes were labelled as *TLM* genes and the mutants *tlms*. From the 269 deletion strains of the Reimand *et al.* dataset 18 deletion strains showed shorter telomeres than the wild-type (short *tlm* mutants), 11 elongated telomeres, and 240 wild-type telomere length (non-*TLMs* or controls) (Table S1).

2.2.2 RNA-seq data from The Cancer Genome Atlas (TCGA)

For the pan-cancer analysis of *TERT* regulation only the cancer entities from 'The Cancer Genome Atlas' (TCGA) with freely available transcriptome expression data of more than 100 primary tumor samples were selected. This resulted in a dataset comprising 19 different cancer entities (Table 2). For the pre-processed RNA-seq data of these cancer entities the usage restriction has been relaxed according to the TCGA publication guidelines from December 21, 2015 (http://cancergenome.nih.gov/publications/publicationguidelines) making the data publicly available. The normalized by 'RNA-Seq by Expectation Maximization' (RSEM) (Li and Dewey, 2011) and log2 transformed data was downloaded from the TCGA Genome Data Analysis Center (GDAC, http://gdac.broadinstitute.org/, release 2016-01-28) of the Broad Institute. Each dataset was reduced to the primary

tumor samples (ending -01 for solid tumors and -03 for blood-derived cancers based on the TCGA guidelines (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes)). For the prostate dataset, also 52 samples from normal tissue (ending -11) were available. In each dataset, all genes with more than 25 % missing entries and with low variances (standard deviation ≤ 0.5) were removed. Furthermore, a *z*-score transformation was performed for each gene across each dataset (see equation 1). For the modeling, all samples without a gene expression value for the gene of interest were removed.

Table 2. Cancer types in TCGA with more than 100 primary tumor samples, used for the pan-cancer analysis of *TERT*.

Cancer type	Number of tumor samples
Breast cancer (BRCA)	983
Cervical cancer (CESC)	300
Colorectal adenocarcinoma (COADREAD)	619
Cutaneous melanoma (SKCM)	103
Glioblastoma multiforme (GBM)	145
Head and neck squamous cell carcinoma (HNSC)	501
Liver hepatocellular carcinoma (LIHC)	342
Lung adenocarcinoma (LUAD)	491
Lung squamous cell carcinoma (LUSC)	489
Ovarian serous cystadenocarcinoma (OV)	294
Prostate adenocarcinoma (PRAD)	497
Stomach adenocarcinoma (STAD)	405
Urothelial bladder cancer (BLCA)	399
Uterine corpus endometrial carcinoma (UCEC)	532
Acute Myeloid Leukemia (LAML)	168
Testicular germ cell cancer (TGCT)	148
Esophageal cancer (ESCA)	178
Pancreatic ductal adenocarcinoma (PAAD)	145
Thymoma (THYM)	120

2.2.3 Next-generation sequencing (NGS) data of pediatric glioblastoma (pedGBM)

For the TMM classification into ALT and non-ALT samples, data from 57 pedGBM patients and 7 cell lines were used. Most samples were from the International Cancer Genome Consortium (ICGC) PedBrain Tumor Project (http://www.pedbraintumor.org/) (International Cancer Genome Consortium PedBrain Tumor Project, 2016). The sequencing data of these patients can be downloaded from European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under the accession number EGAS00001001139. Additional samples

(ICGC_GBM84, ICGC_GBM95, ICGC_GBM96, ICGC_GBM98, ICGC_GBM100, and cell lines NEM168, NEM165) were processed similar to the ICGC ones. For nearly all samples RNA-, DNA-seq and data from methylation profiles using Illumina HumanMethylation450 BeadChip was available. From the methylation data only the methylation level at CpG cg11625005 upstream of the *TERT* transcription start site was used.

2.2.4 RNA-seq data of chronic lymphocytic leukemia patients (CLL)

The RNA-seq dataset consisted of data from 20 CLL patients and 7 non-malignant B-cell control samples generated in the CancerEpiSys consortium (http://www.cancerepisys.org/cancerepisys/index.html). More information about the samples and the pre-processing of the data can be found in (Mallm *et al.*, 2019). For the activity calculation with VIPER (Alvarez *et al.*, 2016) raw read counts were normalized with variance-stabilization transformation using DESeq2 (Love *et al.*, 2014), while for MIPRIP gene expression values in 'Reads Per Kilo base per Million' (RPKM) mapped reads were used.

2.3 Assembling the initial regulatory networks

For the calculation of the regulator's activity value and for the MILP based models a network with all available regulator to target gene interactions was needed. For yeast and human a generic regulatory network was built mainly based on available ChIP-experiments.

2.3.1 Generic yeast regulatory network

A generic regulatory network for yeast was constructed based on regulator binding information from the YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) database (www.yeastract.com) and a study of Yu and Gerstein (Yu and Gerstein, 2006). In August 2015, YEASTRACT comprised binding data based on more than 1,300 publications and for the network only entries annotated as "DNA binding plus expression evidence" from ChIP-experiments or *in silico* refinements of this data (Harbison *et al.*, 2004; Lee *et al.*, 2002; Reimand *et al.*, 2010) were used. All regulators without any target gene in both sources were

filtered out. In total this network consisted of 203,234 interactions between 382 regulators and 6,346 target genes.

2.3.2 Generic human regulatory network

For the generic human regulatory network all available regulator to target gene interactions from the following seven data repositories were extracted:

- MetaCore™ (https://portal.genego.com/): the TF-target gene interactions are based on proven literature reports and are manually curated. The interactions are annotated as "direct" or "indirect" and are subdivided into “activating”, “inhibitory” or "unspecific".
- the ChIP Enrichment Analysis (ChEA) database (Lachmann *et al.*, 2010), which contains interactions from high-throughput ChIP-experiments.
- chromatin immunoprecipitation data from the ENCODE project (https://www.encodeproject.org/). Only entries which were found in at least 2 cell types were used.
- human ChIP-ChIP and ChIP-seq data from hmChIP (Chen *et al.*, 2011),
- the experimentally verified interactions from Human Transcriptional Regulation Interaction database (HTRIdb) (Bovolenta *et al.*, 2012),
- ChIP-seq data of long non-coding RNA and microRNA genes from ChIPbase (Yang *et al.*, 2013) and
- data from binding site predictions using the method of Total Binding Affinity (TBA) (Grassi *et al.*, 2015; Molineris *et al.*, 2011). Here, the TF’s binding probability over the whole promoter region of the gene is estimated. For TBA, a stringency cutoff of ≥ 1.5 was set.

The binding information of the different repositories were combined into a generic network of TFs and their target genes. Only TF *t* to target gene *i* interactions were selected if it was listed

(i) in MetaCore™ annotated as direct, or in Encode,

(ii) in at least two out of MetaCore™ (annotated as indirect), TBA (score≥1.5), ChEA or HTRIdb, or

(iii) in hmChIP and ChIPbase.

The interactions from the seven different repositories were incorporated based on the reliability of the sources. Interactions from MetaCore™ were manually curated

from literature reports, MetaCore's direct interactions ($x_{MCdir_{ti}}$) were weighted by a factor of 2 and MetaCore's indirect interactions ($x_{MCindir_{ti}}$) by a factor of 1. A factor of 1 was also used for interactions from ChEA ($x_{chea_{ti}}$), HTRIdb ($x_{htri_{ti}}$) and TBA ($x_{tba_{ti}}$). Interactions from Encode ($x_{enc_{ti}}$) were weighted by a factor of 0.5, and for interactions found in hmChIP ($x_{hm_{ti}}$) and ChIPbase ($x_{chip_{ti}}$) a factor of 0.25 was used (Poos *et al.*, 2019). This resulted in the overall edge strength score es_{ti}:

$$es_{ti} := 2 \cdot x_{MCdir_{ti}} + 0.5 \cdot x_{enc_{ti}} + a_{ti} \cdot \left(x_{MCindir_{ti}} + x_{chea_{ti}} + x_{htri_{ti}} + x_{tba_{ti}}\right) + 0.25 \cdot \left(x_{hm_{ti}} \cdot x_{chip_{ti}}\right) \quad (13)$$

with

$$a_{ti} := \begin{cases} 1 \; if \; \left(x_{MCindir_{ti}} + x_{chea_{ti}} + x_{htri_{ti}} + x_{tba_{ti}}\right) \geq 2 \\ 0 \; otherwise \end{cases}, \quad (14)$$

$$x_{tba_{ti}} := \begin{cases} 1 \; if \; z - score \geq 1.5 \\ 0 \; otherwise \end{cases}, \quad (15)$$

and $x_{MCdir_{ti}}, x_{MCindir_{ti}}, x_{enc_{ti}}, x_{chea_{ti}}, x_{htri_{ti}}, x_{hm_{ti}}, x_{chip_{ti}} \in \{0,1\}$. (16)

Regulators for which no target genes could be identified were filtered out.

2.4 **M**ixed **I**nteger linear **P**rogramming based **R**egulatory **I**nteraction **P**redictor (MIPRIP)

MIPRIP integrates a MILP based regulatory model with machine learning methods. The MILP model, in which the covariates are the putative regulators binding to the promoter of the gene of interest, is as follows:

$$\tilde{g}_{ik} = \beta_0 + \sum_{t=1}^{T} \beta_t \cdot es_{ti} \cdot act_{tk}, \quad (16)$$

where $\tilde{g}_{ik}$ was the predicted gene expression value of gene *i* in sample *k*, β_0 was an additive offset, *T* the number of all putative regulators binding to the gene's promoter, β_t was the optimization parameter of regulator *t*, es_{ti} the edge strength between regulator *t* and gene *i* based on the underlying network (yeast or human)

and act_{tk} the activity of regulator *t* in sample *k* calculated based on the gene expression values of the target genes (equation 31 and 32). The edge weight in the network was unequal 0 if there has been an interaction between the regulator *t* and the target gene *i* reported.

The objective was to minimize the difference of the measured gene expression value g_{ik} in the dataset and the predicted gene expression value $\tilde{g}_{ik}$, which is equal to the minimization of the error terms e_{ik} (L1 regression):

$$\min \sum_{k=1}^{l} |g_{ik} - \tilde{g}_{ik}| = \sum_{k=1}^{l} e_{ik}, \tag{17}$$

and all absolute values were transformed into two inequalities:

$$g_{ik} - \tilde{g}_{ik} - e_{ik} \leq 0 \tag{18}$$

$$-g_{ik} + \tilde{g}_{ik} - e_{ik} \leq 0 \tag{19}$$

A large variety of models with different sizes was constructed by constraining the number of regulators. For this a binary variable was introduced for each regulator t called x_t plus an additional constraint (eq. 20). If x_t is equal to 1, then regulator *t* can be selected by the model. The sum of all binary x_t variables is at most a specified number of regulators (*limit*),

$$x_1 + x_2 + \cdots + x_t \leq limit \; ; x_t \in \{0,1\}. \tag{20}$$

Typically, models starting with only one regulator up to a maximum of *n-2* putative regulators, where n was the number of samples, were constructed.

Furthermore, a variable called 'Big M' was utilized to define the bounds of the β_t of regulator t. The bounds of each β_t were set to $-1000 \leq \beta_t \leq 1000$.

$$\beta_t - 1000\, x_t \leq 0 \tag{21}$$

$$\beta_t + 1000\, x_t \geq 0 \tag{22}$$

To solve the optimization problem the Gurobi optimizer (www.gurobi.com, version 6.0-7.01) was used. To avoid overfitting a cross-validation was performed and the prediction performance was estimated by the Pearson correlation of the predicted

gene expression value from the training set and the measured gene expression value from the validation dataset. After all the resampling and cross-validation runs it was counted how often each regulator was selected in the models of each group over all the cross-validation runs. This results in a table where the regulators were ranked based on their frequency over all models.
The combination of this MILP model and machine learning methods in form of cross-validation and resampling techniques was implemented in R (https://www.r-project.org/). Because the here described analysis is limited to a single dataset, it is also called a single-mode MIPRIP analysis. Furthermore, MIPRIP was extended with (i) statistical analysis to compare the regulatory processes between two or multiple datasets/conditions (MIPRIP-Comparison (MIPRIP-Comp)) and (ii) a modularity-based approach to identify the regulatory subnetwork that can best explain the regulation of the gene of interest (MIPRIP-Network (MIPRIP-Net)) (see next sections).

2.4.1 MIPRIP-Comparison (MIPRIP-Comp)

MIPRIP-Comp can be used to compare the regulatory processes between different two (dual-mode) or multiple (multi-mode) datasets/conditions based on a statistical downstream analysis. The basic MIPRIP analysis of only one dataset (single-mode) as well as the dual- and the multi-mode of MIPRIP-Comp (Figure 7) are implemented in the R-package “MIPRIP2”. MIPRIP2 requires the R-package “slam” and the solver Gurobi (free for academic use). In all three MIPRIP-modes, models were built for 1 up to a pre-defined number of regulators by using a prior defined number of cross-validation runs separately for both datasets/conditions. For each dataset/condition it was counted how often each regulator was selected over all models. In the dual-mode MIPRIP analysis, a two-sided Fisher's exact test was performed with the regulator frequencies to identify significant regulators of dataset/condition 1 being found more often compared to dataset/condition 2 and *vice versa*. The *p*-values were corrected for multiple testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). In the multi-mode the count table with the regulator frequencies of all datasets/conditions was used to identify the most common regulators of the gene of interest across all datasets/conditions. For this a rank product test was performed based on the ranks of the counts of each regulator in each dataset/condition. To estimate if the calculated rank product value

over all datasets/conditions was higher than an observed value of a random distribution, 10,000 permutations of a rank product value calculated from random TF ranks of each dataset/condition were performed. An averaged expected value (E-value) was calculated by counting how often the rank product values in the permutations were below or equal to the observed value (Breitling *et al.*, 2004). To identify the specific regulators of each dataset/condition a one-sided Wilcoxon Test for each regulator was performed based on the count distributions and the resulting *p*-values were corrected for multiple testing (Benjamini and Hochberg, 1995). This leads to a list of regulators which were significantly selected more often in the models of the dataset/condition compared to the models of all other datasets/conditions.

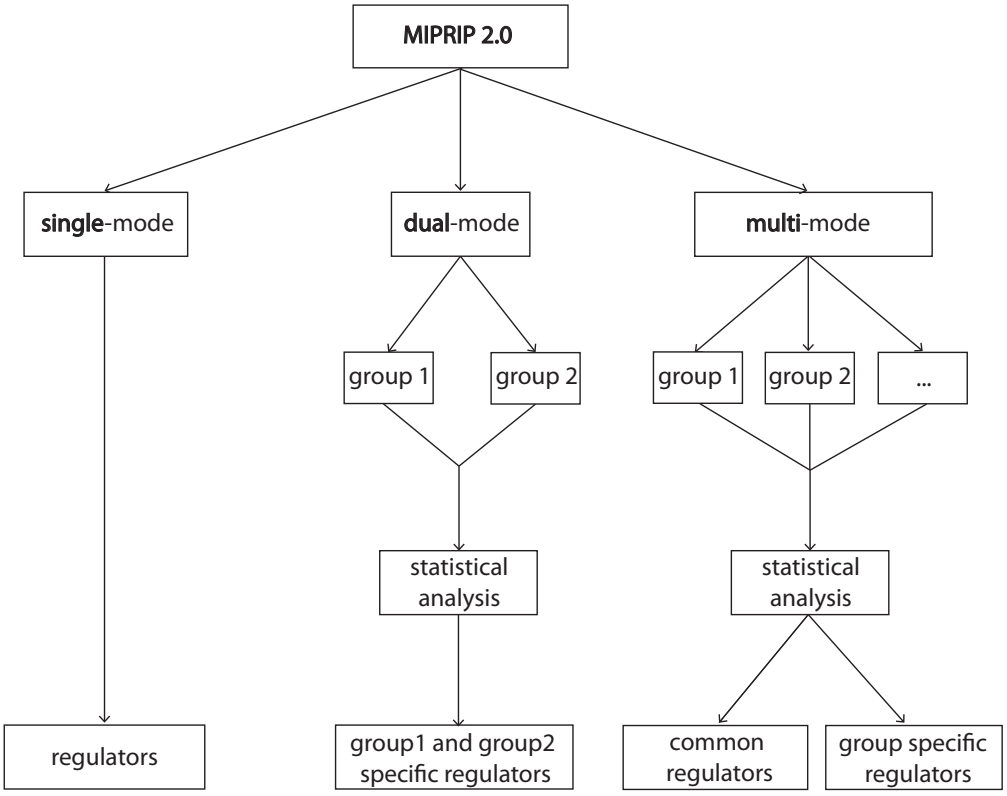

Figure 7. Three different application modes in the MIPRIP2 R-package.
The single-mode identifies the regulators which are used most often in the models to predict the gene expression of the gene of interest. This mode can be used only for a single dataset or a specific group of samples. The dual-mode was designed to study the regulatory processes between two datasets or two groups of samples (e.g. treatment vs. control). After the modeling the user receives a list of regulators which were significantly more often used for the prediction of group1 compared to group2 and *vice versa*. With the multi-mode the most common but also the group-specific regulators are predicted for multiple datasets/groups (e.g. for pan-cancer analysis). Image taken from (Poos *et al.*, 2019).

The R-package MIPRIP2 can be used like this:

```
> library("MIPRIP2")
> library("gurobi")
> miprip.result <- miprip.run(mode=c("single", "dual", "multi"), group_names=c(), target_gene, num_repeats=10, num_cv=3, num_parameter=10, gurobi_parameter=list(timeLimit = 5, OutputFlag=0), X=expression, ES=network)
```

,where the user specifies the mode (single, dual or multi), the groups (group_names), the target gene, the number of repeats and cross-validation runs (num_repeats, num_cv), the maximum number of regulators (num_parameter) and the loaded expression dataset (X) as well as the generic network (ES). Per default the activity values of the regulators are calculated within the MIPRIP 2.0 run, but the user can also provide an own activity matrix as in the first MIPRIP version. 'gurobi_parameter' defines the time limit of each optimization step.

So far there are two MIPRIP versions available at https://www.leibniz-hki.de/en/miprip.html or https://github.com/network-modeling/MIPRIP (Poos *et al.*, 2019; Poos *et al.*, 2016). MIPRIP version 1.0 was limited to a binary regulatory network and included a further inner cross-validation to improve the prediction performance. This means that the training set of the cross-validation was again divided into a training and a validation set to determine the regulator combination with the best performance.

2.4.2 MIPRIP-Network (MIPRIP-Net)

To construct a regulatory TF-TF network, MIPRIP was combined with the concept of modularity from Newman (Newman, 2006). This leads then to the best subnetwork of a particular gene consisting of direct and indirect regulators R_t. All regulators binding to the promoter of the particular gene are called direct, while the regulators of the regulators are called indirect regulators of the particular gene.

The MILP was as follows:

$$x_{t_1} + x_{t_2} - y_{t_1 t_2} \leq 1 \tag{23}$$

$$y_{t_1 t_2} \leq x_{t_1} \tag{24}$$

$$y_{t_1 t_2} \leq x_{t_2} \tag{25}$$

$$\sum_{t=1}^{T} x_t \leq limit \tag{26}$$

$$w_{t_1t_2} = cor\left(act_{t_1k}, act_{t_2k}\right) \cdot es_{t_1t_2} \quad (27)$$

$$\widetilde{w}_{t_1t_2} = w_{t_1t_2} - \frac{d_{t_1}\, d_{t_2}}{2m} \quad (28)$$

with $d_t = \sum w_{t_1,t_2}$

and $m = \frac{1}{2}\sum_{t=1}^{T} d_t$

$$x_t \in \{0,1\},\ y_{t_1t_2} \in \{0,1\}, \quad (29)$$

where *t* indicates the nodes (regulators), *w* are the edge weights, *d* the degree of the node and T the number of all regulators. *x* and *y* are binary parameters and indicate if the nodes and edges were selected.

Constraint (23) enforced that if node t_1 and node t_2 were in the module than also the edge between t_1 and t_2 had to be in the module. By constraints (24-25) it is ensured that only edges were selected for which both end nodes were inside the module. The size *T* of the module is constrained by equation (26). The goal of the modularity was to identify a highly connected module which can best explain the regulation of the particular gene of interest. Therefore, the sum of the edge weights between the connected nodes inside the modules was maximized and penalized if their end nodes had high degrees. The corresponding edge weights *w* were computed as described in (27-28) by multiplying the correlation of each regulator pair's activity over all investigated samples *k* with the corresponding edge weights in the generic network. Because this weight was not always the same between node (regulator) t_1 and t_2, the mean value of both directions was taken. All these weights were computed in a preprocessing step and were constants in the MILP. For the combined model of MIPRIP and modularity, all equations of MIPRIP (see 2.4) and the equations above were used. As objective function of the combined model the sum of objective functions of the single models were used:

$$Min\ \underbrace{\sum_{k=1}^{l} e_{i,k}}_{\textbf{MIPRIP}} - \lambda\ \underbrace{\sum_{t_1,t_2 \in V;\, t_1 \neq t_2} w'_{t_1t_2} \cdot\ y_{t_1t_2}}_{\textbf{modularity}}. \quad (30)$$

For this, variables of the direct regulators were the same for both optimization parts. The parameter lambda controlled the tradeoff between MIPRIP and the modularity by maximizing the sum of edge weights between the connected nodes in the

modularity network (equation 30). The performance of the model was determined as with MIPRIP alone. The best subnetwork consists of the combination of MIPRIP regulators (defined by the x_t) which was used most often in all the models and the corresponding modularity regulators.

2.5 Applications of the MIPRIP approach

MIPRIP was applied to study the regulation of the telomerase in (i) *S. cerevisiae* and (ii) different human cancer types, especially melanoma skin cancer and prostate cancer.

2.5.1 Applying MIPRIP-Comp to gene expression data of *S. cerevisiae*

MIPRIP-Comp was applied to regulator deletion strain data of *S. cerevisiae* to study the regulation of the ever shorter telomere genes *EST1*, *EST2* and *EST3*. For the yeast gene expression data, the activity act_{tk} of regulator *t* and deletion strain *k* was calculated using absolute gene expression values:

$$act_{tk} = \frac{\sum_{i=1}^{n} es_{ti} \cdot |g_{ik}|}{\sum_{i=1}^{n} es_{ti}} \text{ with } es_{ti} \in \{0,1\}, \quad (31)$$

where es_{ti} is the edge strength between regulator *t* and gene *i*, g_{ik} the gene expression value of gene *i* in strain *k*. The edge weight between regulator *t* and gene *i* was equal to 1 if there was an interaction in the generic yeast regulatory network, and otherwise zero. For the modeling, the dataset was split up into knockout strains showing short telomeres (short *tlm* mutants), long telomeres (long *tlm* mutants) and knockout strains with normal telomere length (control samples) based on the annotation of (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Shachar *et al.*, 2008; Ungar *et al.*, 2009). Deletion strains with a long telomere phenotype were excluded because telomere elongation is one of the hallmarks of cancer. This led to a dataset of 18 short *tlm* mutants and 240 control samples. For each *EST* gene, dual-mode MIPRIP-Comp analysis was performed.

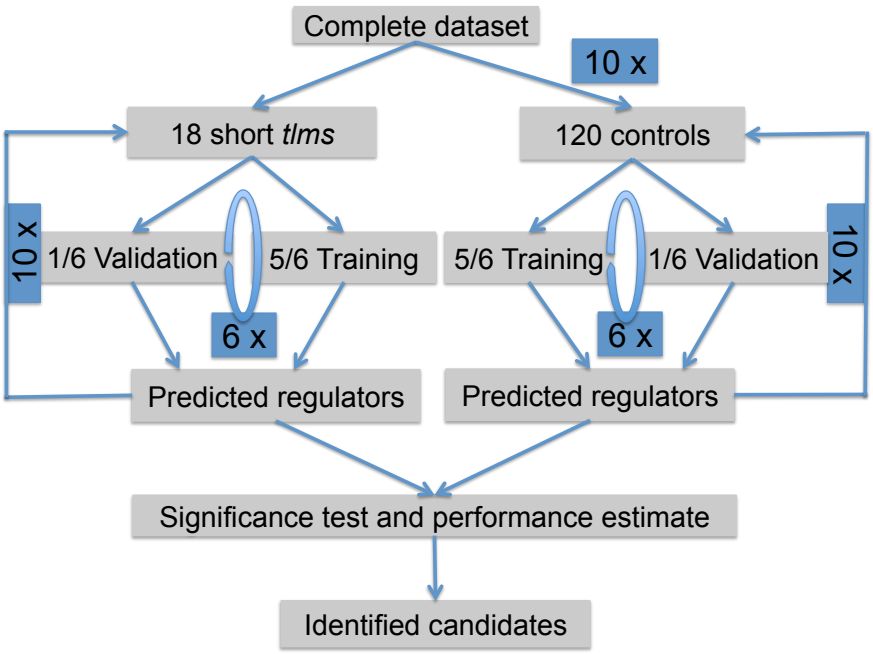

Figure 8. Schematic workflow of the yeast study.

The gene expression dataset was divided into knockout strains showing short telomeres (short *tlms*) and knockout strains with wild-type telomere length (controls). From the control dataset 120 samples were randomly selected. For each dataset a ten-times six-fold cross-validation was performed (plus a five-fold inner-cross-validation, not shown here). For the control dataset the process was further repeated ten-times. It was then counted how often each regulator was selected in the short *tlm* and the control models over all cross-validation runs. With these regulator frequencies a one-sided Wilcoxon Test was performed to identify regulators that were significantly more often selected in the short *tlm* models compared to the control models. Image taken from (Poos *et al.*, 2016)

First, ten-times 120 samples were randomly selected (drawing with replacement) from the control dataset (240 patietns). For the 18 short *tlm* mutant dataset and for each of the selected control dataset, models were constructed by using a ten-times sixth-fold cross-validation (Figure 8). Furthermore, the number of regulators was constrained from 1 up to 10. For each number of regulators, a further five-fold inner cross-validation was performed to determine the regulator combination with the best performance. The performance of the model was estimated by calculating the correlation of the predicted and the real gene expression values.

In summary, for each *EST* gene 60 different models were constructed for the short *tlm* mutant dataset and 600 for the control dataset. With the regulator frequencies a one-sided Wilcoxon Test was performed to identify the regulators which were significantly used more often in the models of the short *tlm* dataset compared to the control dataset.

2.5.2 Applying MIPRIP-Comp to gene expression data of different human cancer types

To study the regulation of the *TERT* gene in different cancer types, MIPRIP-Comp was applied to gene expression data from TCGA. For all the MIPRIP analysis of *TERT*, models were constructed by constraining the number of regulators from one regulator up to 10 regulators and a ten-times threefold cross-validation was performed. This yields to 300 models for each dataset. The activity of the *TERT* regulators was calculated by

$$act_{tk} = \frac{\sum_{i=1}^{n} es_{ti} \cdot g_{ik}}{\sum_{i=1}^{n} es_{ti}} \text{ with } es_{ti} \in \mathbb{R}^{\geq 0}, \quad (32)$$

where the gene expression value of *TERT* was excluded.

Pan-cancer *TERT* analysis: To study the regulation of *TERT* in 19 different cancer types a multi-mode MIPRIP-Comp analysis was performed with the pre-processed gene expression datasets of 19 different cancer types from TCGA (Table 2). This led to a list with the common *TERT* regulators across all datasets and a list with the specific *TERT* regulators of each cancer type compared to all other cancer types.

Skin cutaneous melanoma (SKCM) case study: Because SKCM is a cancer entity with a high frequency of *TERT* promoter mutations, the SKCM dataset (primary and metastatic samples) was split up into samples with and without *TERT* promoter mutation based on the data of (Cancer Genome Atlas, 2015). With these two subgroups a dual-mode MIPRIP-Comp analysis was performed leading to a table with regulators selected significantly more often in the samples with *TERT* promoter mutation compared to the wild-type *TERT* promoter samples and vice versa.

To validate the importance of ETS1 for the *TERT* promoter samples, microarray gene expression data of TF perturbation experiments with the melanoma cell line A375 was investigated. The microarray data (Affymetrix GeneChip Human Genome U133 Plus 2.0) was generated by Wang *et al.* for siRNA mediated knockdowns of 45 TFs and signaling molecules and contains expression values of the knockdown experiments (1 sample per knockdown, 48 h after transfection) and untreated (3

replicates) as well as siRNA control treated (3 replicates) samples (Wang *et al.*, 2012). The RMA-normalized (robust multi-array average) expression data was freely available at Gene Expression Omnibus (GSE31534). For this data the Affy probe-ids were mapped to gene symbols using BioMart (Smedley *et al.*, 2015). For multiple affy probe-ids of the same gene an average was computed. A fold change was calculated for *TERT* upon ETS1 knockdown compared to control.
As a comparison to MIPRIP, an ISMARA analysis (Balwierz *et al.*, 2014) was performed with the same TCGA SKCM samples with and without *TERT* promoter mutation. For ISMARA, gene expression values in 'Fragments Per Kilobase per Million' (FPKM) mapped reads from the TCGA SKCM dataset were downloaded from the GDC portal (https://portal.gdc.cancer.gov/, June 2018). Because only preprocessed data was available, it was not possible to use the web portal of ISMARA. Therefore, the developers started the ISMARA analysis using default settings.

Prostate cancer: A dual-mode MIPRIP-Comp analysis was performed for the prostate cancer (n=445) and the normal prostate (control, n=18) samples to identify the regulators which were used significantly more often in the prostate cancer models.

2.5.3 Applying MIPRIP-Net on gene expression data of prostate cancer

MIPRIP-Net was applied to the TCGA prostate cancer data to identify the best submodule explaining *TERT* regulation. For this the overlap of the prostate cancer specific *TERT* regulators from the pan-cancer analysis and the significant regulators of prostate cancer versus healthy prostate tissue was selected for the MIPRIP-Net model. To reduce computational complexity, the number of regulators of these 12 selected regulators was limited. For this a basic MIPRIP analysis was performed for each of the selected regulators with the same parameter setting as for the MIPRIP-Comp analysis and the regulators used in at least 20 % of the models were selected for the MIPRIP-Net approach. In the pre-processed expression matrix, CTCF and NR2F2 were filtered out because of low variances and TFAP2D because of too many missing expression values. For TFAP2D no MIPRIP model was possible, while for CTCF and NR2F2 the unfiltered gene expression data was used. This means that 12 regulators were used for the MIPRIP model and additional 72 for the

modularity part of the model. In total, models of 2 up to 20 regulators were constructed and a ten-times three-fold cross-validation was performed. To find the optimal tradeoff between the objective functions of MIPRIP and the modularity approach, models with 9 different λ-values (0.001, 0.01, 0.1, 0.3, 1, 3, 10, 100, 1000) were calculated. The optimal λ was determined as the intercept of the number of selected MIPRIP regulators over all models and the number of selected modularity regulators. For the optimized λ-value a combined model was constructed with the parameters of the other λ-values. The performance was determined similar to MIPRIP. The best subnetwork consisted of the most often selected MIPRIP regulator combination and the corresponding modularity regulators. The subnetwork which can best predict the expression of *TERT* was visualized with Cytoscape (Shannon *et al.*, 2003).

2.6 Distinguishing between different telomere maintenance mechanisms

To stratify pedGBM patients according to their TMM a decision tree-based classifier was constructed and after the classification an ALT signature was determined based on differentially activity (regulators) or expression levels (genes).

2.6.1 TMM classifier

A decision tree-based classifier was constructed to predict the TMM of pedGBM patients. For this TMM specific features from imaging data (presence of ultra-bright telomere foci from FISH, ATRX loss of expression from IHC staining), from DNA-sequencing data (chromothripsis, loss of function mutations in *ATRX* and *TP53*, K27M and G34R/V mutations in *H3F3A*; extracted from (International Cancer Genome Consortium PedBrain Tumor Project, 2016)), RNA-sequencing (*TERT* expression) and telomere qPCR (telomere content) as well as from the Illumina 450K array methylation data (*TERT* promoter methylation) were used as features to predict if a sample is ALT-positive (“ALT”) or ALT-negative (“non-ALT”) based on intelligent enumeration (Table S24). An incomplete list of TMM features was available for 57 pedGBM patients and 7 cell lines. All sequencing readouts were available for 44 samples, while for another 20 samples at least one sequencing readout was missing. The biological assays (ATRX IHC, telomere content, Ultra-

bright telomere foci and C-circle assay) could be performed only for a subset because of limited patients' material (Table 3). The C-circle assay and the *TERT* promoter mutation status were used to define the two classes (ALT *vs*. non-ALT). Samples with an activating *TERT* promoter mutation were in any case ALT-negative, while a positive C-circle signal indicated an ALT-positive sample. Samples with a negative C-circle result and no ultra-bright telomere foci were used as training set for the ALT-negative group. The results from the C-circle assay and the *TERT* promoter mutation status were not included as features for training and validation. Altogether, the training set contained 27 pedGBM patients (13 ALT and 14 non-ALT, including 2 samples with *TERT* promoter mutation) and 7 cell lines (5 ALT and 2 non-ALT). Most of the features could be easily translated into binary values. For the features "*TERT* promoter methylation", "*TERT* expression (RPKM)" and "telomere content", continuous values were available and for these features optimal thresholds had to be defined first. Based on the ALT and non-ALT samples of the training set the threshold with the fewest misclassified samples was determined by testing different thresholds. The thresholds were defined separately for the patients and the cell lines (Figure 23). For each feature combination all possible decision trees were calculated and the tree with the minimal number of misclassified samples and questions based on the training set was selected. The accuracy *Acc* of the tree was indicated by the number of correctly predicted samples in the training set and was determined by a leave-one-out cross-validation. The optimal tree was then derived by using all training samples. Since the feature information was not complete for each sample, decision trees with all different possibilities from 1 up to 9 features were constructed, which required different sample sizes. If the accuracy *Acc* did not improve with more features, the decision tree with the best feature subset was used. Furthermore, *p*-values for each tree were calculated based on the confusion matrix containing the numbers of correct and incorrect classified samples in the ALT and the non-ALT group over all cross-validation runs using Fisher's exact test. The *p*-values were corrected for multiple testing according to Benjamini and Hochberg (Benjamini and Hochberg, 1995). It is noted that for some feature combinations the *p*-value was not significant due to too low sample sizes (Table 3).

Table 3. Overview of the sample numbers with feature information in the whole dataset and in the training set.

Feature	# samples with feature present	# training samples with feature present
Chromothripsis	64	34
***TP53* mutation**	64	34
***TERT* promoter mutation**	58	32
***TERT* promoter methylation**	60	32
***TERT* expression (RPKM)**	44	32
***ATRX* mutation**	64	34
***H3F3A* mutation**	64	34
ATRX IHC	35	28
Telomere content	28	22
Ultra-bright telomere foci	25	20
C-circle assay	46	34

2.6.2 ALT gene signature

To identify a gene signature for ALT versus non-ALT pedGBM samples based on gene expression data, differentially expressed genes and regulators were calculated. The differential gene expression analysis was performed with the raw gene counts of the pedGBM RNA-seq data (training set plus predicted samples) using DESeq2 (Love *et al.*, 2014) and a *p*-value cutoff of 0.05. The regulator activities were calculated as described for the MIPRIP approach (equation 32) by using the generic human regulatory network to determine the target genes. Only differentially expressed target genes were used to calculate the activity of a regulator and activities were only calculated for regulators with at least 5 differentially expressed target genes. A Student's t-test was used to identify regulators with significant activity changes between ALT and non-ALT samples. For the regulators, a maximal significance level (*p*-value) of 0.05 after multiple testing correction with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) was used.

2.7 Comparison of MIPRIP with the well-established tools ISMARA and ARACNE/VIPER

To evaluate the advantages and disadvantages of MIPRIP a comparison with the well-established tools ARACNE (Lachmann *et al.*, 2016) and VIPER (Alvarez *et al.*, 2016) was performed. This was done based on the CLL RNA-seq dataset.

To calculate regulator activities with VIPER, first a B-cell specific GRN of around 6,000 regulators and their target genes was constructed based on MI by using the ARACNe-AP algorithm (Lachmann *et al.*, 2016). For the network computation publicly available gene expression data of 264 B-cell samples including samples of B-cell lymphomas, non-malignant B-cells and also cell lines (Basso *et al.*, 2010) were used together with the precompiled list of 5,927 TFs, transcriptional co-factors and signaling pathway related genes defined by Alvarez and coworkers based on Gene Ontology (GO) annotations (Alvarez *et al.*, 2016). The calculation of the GRN was performed within 100 bootstraps, a permutation seed of one and a MI cutoff of $p=10^{-8}$. The B-cell specific GRN contains 214,405 interactions between 3,862 regulators and 12,119 target genes. Based on this B-cell specific ARACNE network and the normalized in-house RNA-seq data of 20 CLL and 7 non-malignant B-cell patients the activity of each regulator in each sample was calculated by using the VIPER algorithm (Alvarez *et al.*, 2016). Activity values could be computed for 2,804 regulators and a two-sided student's t-test was used to identify regulators with a significantly different activity between the CLL and the non-malignant B-cell samples. The *p*-values were corrected for multiple testing using Benjamini and Hochberg method (Benjamini and Hochberg, 1995)

In comparison to the regulator activities calculated with VIPER, regulator activities were calculated for the same dataset as described for MIPRIP based on the generic human regulatory network (see above). Only regulators for which an activity value could be calculated with VIPER and MIPRIP were used for the comparison.

Furthermore, a differential expression analysis was performed using DESeq2 (Love *et al.*, 2014) between the CLL and the non-malignant B-cell samples. For this the raw counts of the in-house RNA-seq data were used and significant difference was determined based on a *p*- value cutoff of 0.05 for the regulators. To focus on target genes that showed a high difference between CLL and non-malignant B-cell

samples, a *p*-value cutoff of 0.01 plus a log fold change (LFC) of -1.7 < LFC > 1.7 was defined.

For the regulators with significant activity changes between the CLL and the non-malignant B-cell samples identified by both VIPER and the MIPRIP framework a gene set enrichment analysis (GSEA) was performed. For the GSEA the R-package “gProfileR” (Reimand *et al.*, 2016) was used with the pathway annotation of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000).

3 Results

For my PhD thesis, I have developed the tool "**M**ixed **I**nteger linear **P**rogramming based **R**egulatory **I**nteraction **P**redictor" (MIPRIP) to study the regulatory mechanisms of gene expression. As a case study, MIPRIP was applied to study the regulation of telomere maintenance.

3.1 Mixed Integer linear Programming based Regulatory Interaction Predictor (MIPRIP)

Typically, several regulators are involved in the expression of a gene by binding to its promoter (Bauer *et al.*, 2011; Cheng *et al.*, 2012; Consortium *et al.*, 2009; Dong *et al.*, 2012; Oliveira *et al.*, 2008; Schacht *et al.*, 2014; Setty *et al.*, 2012). To identify the most relevant regulators of a gene of interest taking into consideration the additive co-operativity of the regulators, a Mixed Integer Linear Programming based approach (MILP) was developed. A MILP based regression model avoids over-emphasizing outliers because the error penalties are linear (L1 norm) and not quadratic (L2 norm) as in a lasso model.

The basic idea behind the MILP model is that the gene expression of the gene of interest can be predicted by a linear model with all the regulators R_1 to R_n binding to the gene's promoter as covariates (Figure 9). For this purpose, all of the putative regulators for the gene of interest were extracted from a generic regulatory network which was constructed beforehand (described in Material and Methods 2.3).

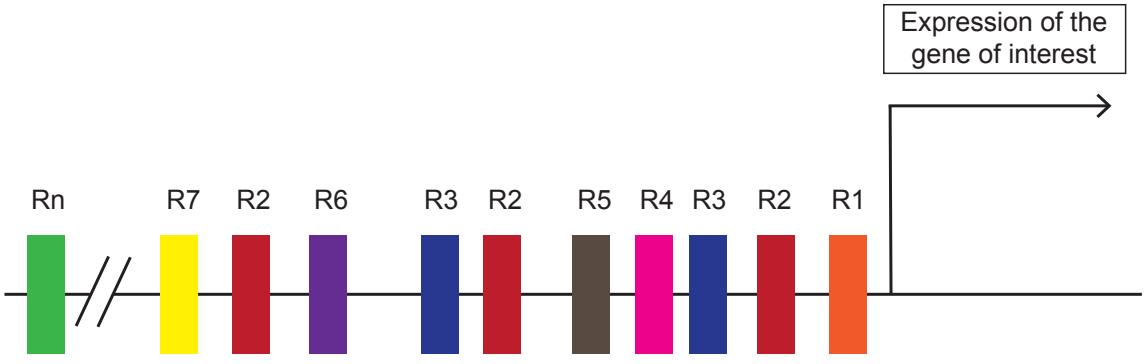

Figure 9. Basic principle of MIPRIP.
The gene expression of any gene of interest is predicted by a linear model with the regulators R_1 to R_n binding to the gene's promoter as covariates.

The described MILP model above was combined with machine learning methods in form of resampling and cross-validation to gain a large variety of different models. Furthermore, the number of regulators was restricted such that models of different sizes could be generated and to obtain sparse models. For this purpose, an additional binary parameter was used. For example, a limit of 5 regulators means that the optimizer tests all possible combinations of those 5 regulators out of a much larger number of putative regulators and identifies the combination which can best predict the gene expression for the gene of interest. A basic MIPRIP analysis predicts the most important regulators of a particular gene in one dataset or one group of samples.

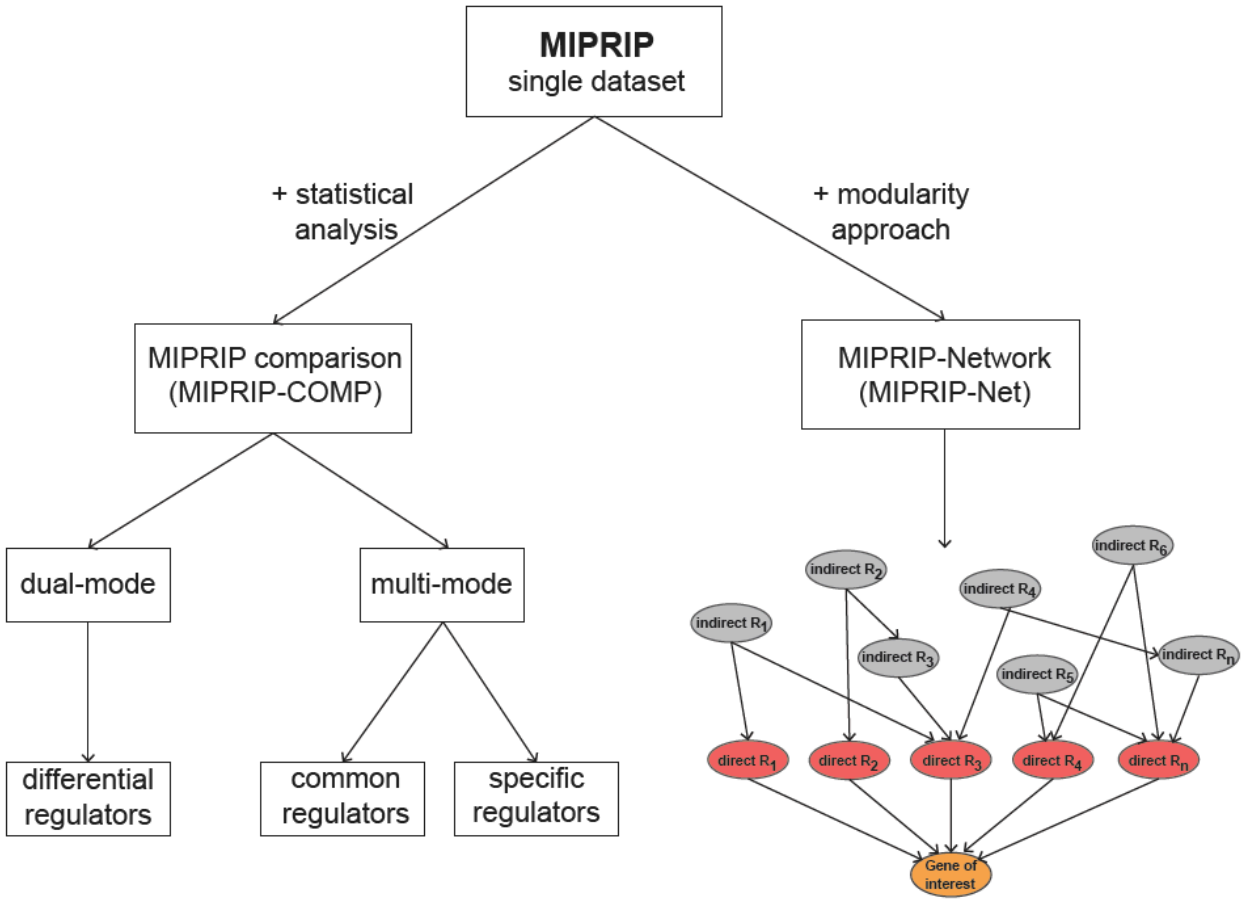

Figure 10. Overview of the MIPRIP framework.
Besides the basic MIPRIP analysis, there are two extensions available. MIPRIP-Comp compares the regulatory processes of a particular gene in two or multiple datasets/conditions. Furthermore, MIPRIP was extended with a modularity-based approach to identify the highly connected subnetwork that can best predict the gene expression of the particular gene (MIPRIP-Net).

To compare the regulatory processes between two or multiple datasets/conditions MIPRIP was extended with a statistical downstream analysis, called MIPRIP-

Comparison (MIPRIP-Comp). In MIPRIP-Comp, the regulatory processes between two datasets/conditions (dual-mode) or between multiple datasets/conditions (multi-mode) can be studied. In addition, MIPRIP was extended with a modularity-based approach to include direct and indirect regulators of the particular gene into a regulatory network (Figure 10).

3.1.1 Regulatory networks used for MIPRIP

To get all the putative regulators of a particular gene, it is essential to have a global regulatory network with all the regulator-target gene interactions as background. We constructed generic regulatory networks for *S. cerevisiae* (Poos *et al.*, 2016), human and mouse (Poos *et al.*, 2019) mainly based on ChIP-binding data. For the yeast network the regulatory interactions of around 400 regulators were extracted from the YEASTRACT database and from (Yu and Gerstein, 2006). The generic human regulatory network integrates regulatory interactions from ChIP-binding data (ChEA, Encode, hmChIP, HTRIdb and ChIPbase), publications (MetaCore™) and computational predictions (TBA). Most interactions were extracted from Encode, followed by ChIPbase and hmChIP (Figure 11A). Here, the highly reliable MetaCore™ interactions represent only 4 % of all extracted TF-target gene interactions. We defined three criteria to integrate only reliable interactions into the generic regulatory network. This means that only TF-target gene interactions found in

(i) MetaCore™ direct or Encode (criteria 1),
(ii) at least 2 out of ChEA, MetaCore™ indirect, HTRIdb and TBA ($z \geq 1.5$) (criteria 2),
(iii) ChIPbase and hmChIP (criteria 3)

were used to construct the generic regulatory network (Poos *et al.*, 2019). Because of these criteria, several interactions were filtered out (Figure 11B).

The compiled generic human regulatory network consisted of 1,160 regulators, 31,915 target genes and 618,537 interactions and was constructed by Theresa Kordaß from the group of Prof. König (University Hospital Jena). The regulators with the highest number of target genes are CTCF (# targets: 16,483), POLR2A (# targets: 16,076) and TAF1 (# targets: 13,956) indicating that these are among other master regulators, while more than half of the regulators had less than 25 target genes (Figure 12). This shows that regulatory interactions between TFs and their

target genes structure as a scale-free network with hubs as master regulators (Babu *et al.*, 2004). It is to be noted that the number of target genes for a regulator seems to be highly dependent on the number of experiments performed for the regulator.

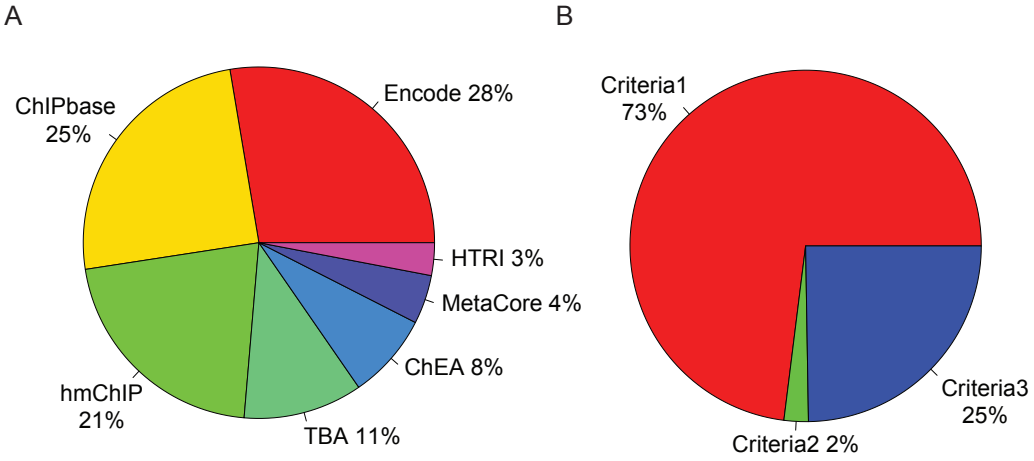

Figure 11. TF-target gene interactions.
(A) extracted from the 7 different sources and (B) fulfilling the defined criteria.

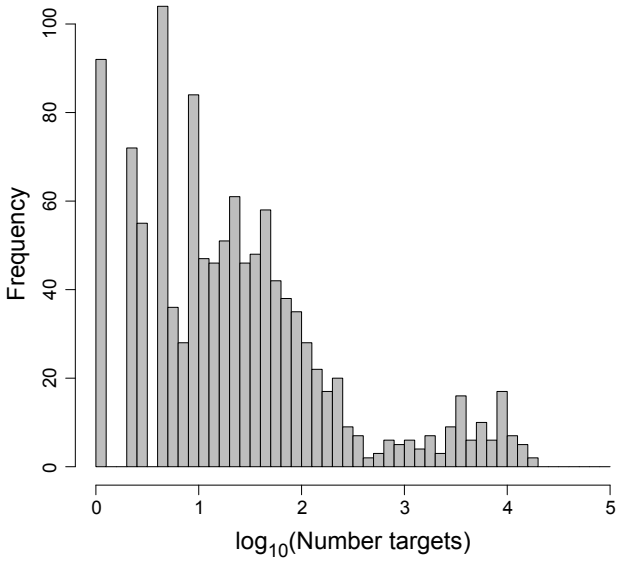

Figure 12. Number of target genes identified for the 1,160 TFs.
The histogram shows the scale-free structure of the generic human regulatory network. Some regulators had up to 17,000 target genes, while more than half of the regulators had less than 25 target genes.

The mouse regulatory network was generated by Amol Kolte analogous to the human generic regulatory network but was not used here for the presented applications.

3.1.2 Regulator activities

Because regulators are post-transcriptional modified and sometimes lowly expressed, the activity of a regulator was calculated based on the expression of the regulator's target genes, similar to previous studies (Balwierz *et al.*, 2014). The activity of a regulator describes the cumulative effect of a regulator on all its target genes. The activity values are normalized by the sum of all target genes to balance regulators with extreme (very high or low) number of target genes. If many target genes of the regulator are differentially expressed, then the regulator itself is more active. The target genes of each regulator were extracted from the generic regulatory network. The activity values of each regulator t in each sample *k* were then used for the modeling instead of the gene expression value of the regulator. In a previous study, it was shown that by implementing a binary switch, the regulator's activity value was used more often by the solver than the gene expression value of the regulator. This led to a better prediction of the gene expression for the gene of interest (Schacht *et al.*, 2014).

3.1.3 MIPRIP-Comparison (MIPRIP-Comp)

MIPRIP-Comp extends the applications of MIPRIP. The basic MIPRIP analysis was only applicable to one dataset or one group of samples (single-mode). Here, the user receives a table how often the regulator was selected in the models over the cross-validation runs and the performance of different sized models. The extension of MIPRIP with statistical downstream analysis allowed the comparison of the regulatory processes between two or more datasets. Using a two-sided Fisher's Exact Test, in the dual-mode, the regulators that were significantly more often used in the models of dataset/condition 1 compared to the models of dataset/condition 2 and *vice versa* were identified. Furthermore, the multi-mode option allows a MIPRIP analysis for more than two datasets/conditions (e.g. pan-cancer analysis). It is combined with a statistical analysis pipeline where the user receives (i) the most common regulators of a gene of interest over all datasets and (ii) the specific regulators of the gene of interest in one dataset/condition compared to all other datasets/conditions. The basic MIPRIP analysis as well as MIPRIP-Comp analysis are implemented in the R-package 'MIPRIP2'. MIPRIP2 is freely available at http://www.leibniz-hki.de/en/miprip.html and https://github.com/network-

modeling/MIPRIP, together with the generic regulatory networks, a user's manual and some example data.

In summary, the MIPRIP2 R-package predicts the most important regulators of a particular gene in one dataset, between two datasets and/or multiple datasets. It can be easily applied to identify crucial regulators of gene expression in yeast, human or mouse.

3.1.4 MIPRIP-Network (MIPRIP-Net)

MIPRIP-Net is a combination of the basic MIPRIP model and a modularity-based approach. Originally, modularity was introduced by Newman and was used for clustering of different modules (Newman, 2006). Here, it is used to identify the best-connected submodule regulating the particular gene. Because biological processes are complex and regulators are highly interacting with each other or with other co-factors to regulate the expression of a particular gene, MIPRIP-Net integrates not only the regulators directly binding to the promoter of the particular gene but also additional regulators directly interacting with the 'direct' regulators but only indirectly with the target genes (Figure 13).

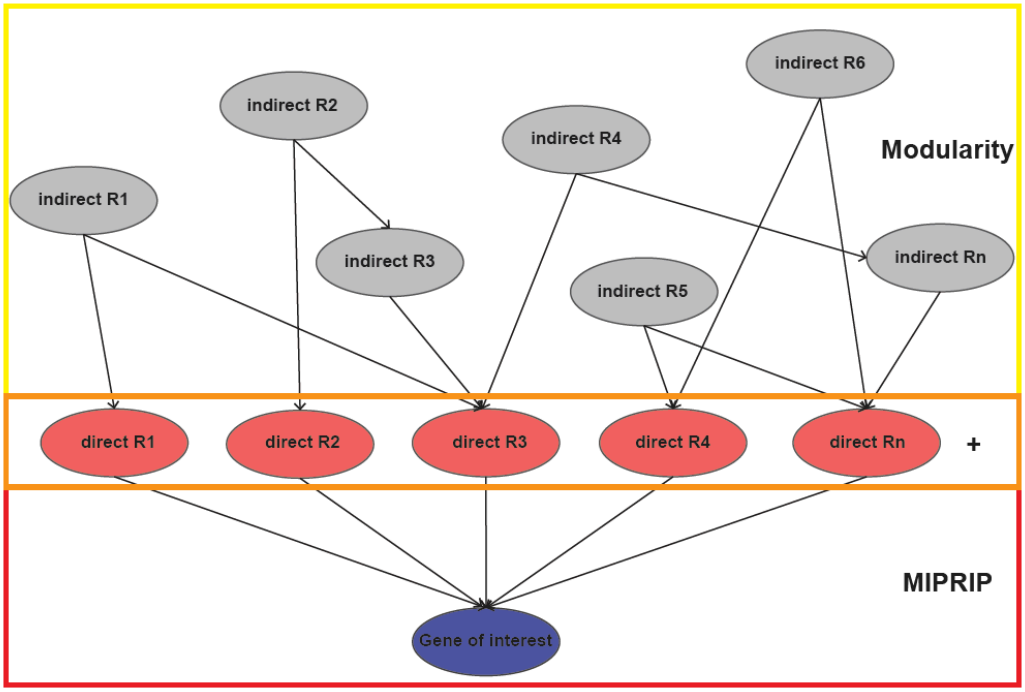

Figure 13. Basic principle of MIPRIP-Net.

MIPRIP-Net is a combination of the established MIPRIP tool and a modularity-based approach to include also the additional regulators like co-factors into a regulatory network. The direct regulators are constrained to be the same for both optimizations.

MIPRIP-Net selects the highly-connected subnetwork, which can best predict the gene expression of the particular gene by (i) minimizing the error term from MIPRIP and (ii) maximizing the sum of edge weights between the connected nodes (modularity task). These two objectives are combined by a tradeoff parameter, which has to be optimized first to find the right balance between high connectivity and accurate prediction of the direct regulators. Based on the modularity an edge is only selected if both end nodes are within the module.

3.2 Telomerase regulation in yeast

Telomere maintenance is a hallmark of cancer cells to enable their replicative immortality. *S. cerevisiae* is a well-suited model system to study telomere maintenance which shows a high homology to humans and has a constitutively expressed telomerase (Teixeira, 2013). MIPRIP was first applied to yeast gene expression data to study the regulation of the telomerase genes (*EST1*, *EST2* and *EST3*). Around 500 yeast genes have been identified which affect telomere length when deleted (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Shachar *et al.*, 2008; Ungar *et al.*, 2009). These genes are called telomere length maintenance (*TLM*) genes. *TLM* genes leading to telomere shortening after deletion ("short *tlm* mutants") are positive regulators of telomere maintenance and could also be potential anticancer targets. In this study, the regulation of *EST* genes was analyzed in yeast deletion strains with shorter telomeres (short *tlms*) compared to deletion strains with wild-type telomere length (controls). For this the microarray gene expression data (Reimand *et al.*, 2010) containing only regulator (TFs and chromatin modifier) deletion strains was divided into a short *tlm* mutant and a control dataset. The Venn diagram (Figure 14) shows the overlap between the regulator deletion strains in the microarray gene expression data, the list of putative regulators of the *EST* genes extracted from the generic yeast regulatory network and the list of *TLM* genes. In total, the dataset consisted of 240 control samples and only 29 *tlm* samples (18 short *tlms* and 11 long *tlms*).

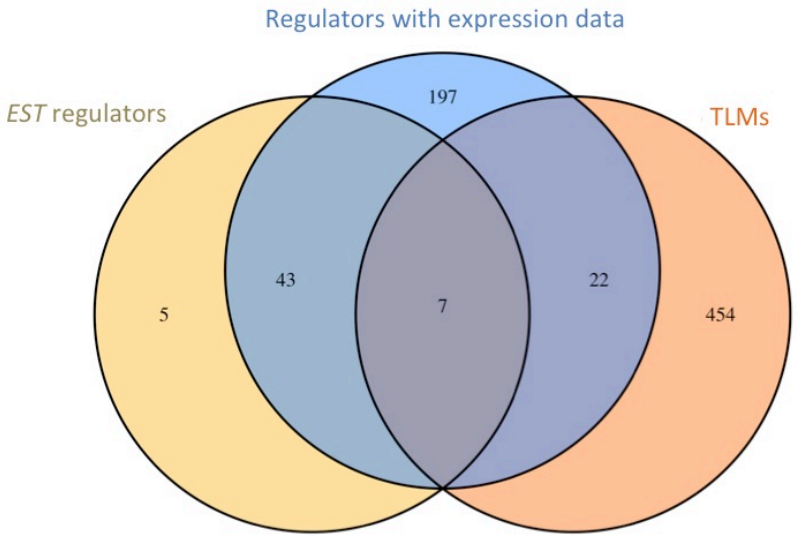

Figure 14. *TLM* genes in the mutant dataset.
The Venn diagram shows the overlap between the putative regulators of the *EST* genes (extracted from the yeast regulatory network, yellow), the regulator deletion strains in the expression dataset (blue) and *TLM* gene list (orange). Image taken from (Poos *et al.*, 2016).

3.2.1 MIPRIP analysis of the *EST* genes

For each *EST* gene a dual-mode MIPRIP-Comp analysis was performed with the short *tlm* mutant dataset and the control dataset. This means that the gene expression value of each *EST* gene in each sample was predicted by a linear model consisting of the sample-specific activity values of the putative regulators. The putative regulators of the *EST* genes were selected from the yeast regulatory network. Each *EST* gene had 25 putative regulators in the regulatory network. For each dataset models were calculated for 1-10 selected regulators and a ten-times six-fold cross-validation was performed. This resulted in 60 models for the short *tlm* mutant. Because of the large number of control samples, models were calculated with a random subset of 120 control samples and this process was repeated ten-times, leading to 600 models for the control samples. For all cross-validation runs, the number of regulators selected in the short *tlm* mutant models was counted, similar for the control models. With these regulator frequencies a significance test was performed to identify the regulators which were selected significantly more often in the short *tlm* mutant models compared to the control models. This resulted in 32 regulators (Table 4) for the three *EST* genes which were selected significantly

more often in the short *tlm* models compared to the control models during all the cross-validation runs. The performance of the models was estimated as the correlation of the predicted gene expression values from the models and the expression values from the dataset and is shown exemplarily for *EST1* (Figure 15). Looking at the correlations of the short *tlm* and the control models together over all cross-validation runs, the mean performance was similar for the models with 1-10 regulators (Figure 15A) and over all models r=0.51 (Figure 15B). Models predicting *EST1* expression showed the highest performance, while for *EST2* (r=0.32) and *EST3* (r=0.12) the performance was lower. But the correlations of all *EST* genes were highly significant (p<2.2 E-16).

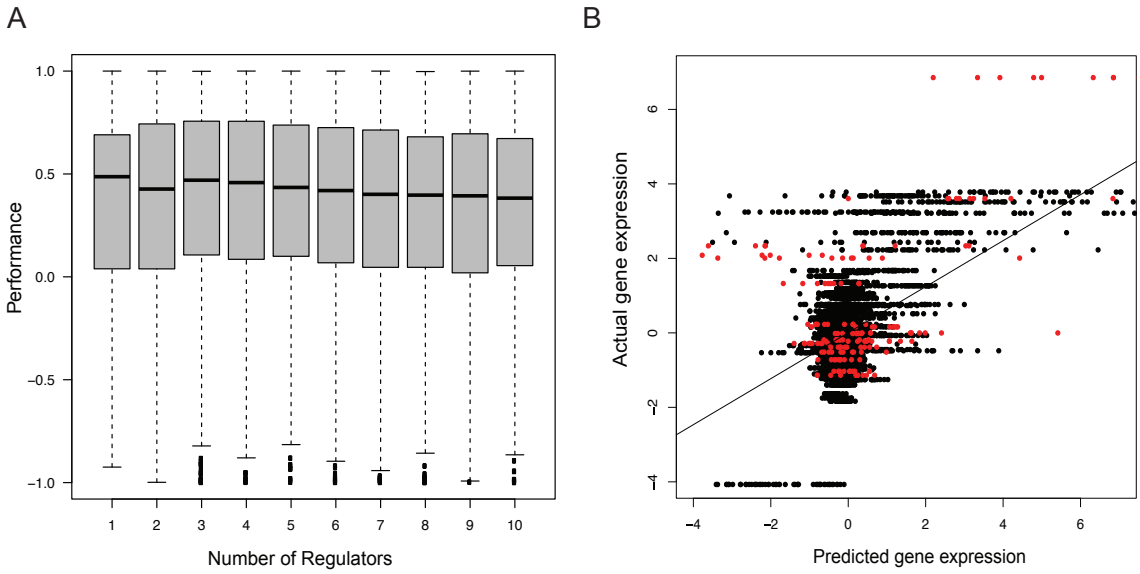

Figure 15. Performance of *EST1* models.
(A) Correlation of predicted gene expression value and *EST1* expression values in the microarray dataset for models of 1 up to 10 regulators over all cross-validation runs (short *tlm* plus control models). (B) Scatterplot of the predicted vs. actual gene expression of *EST1* of the short *tlm* (red) and the control (black) models. Images taken from (Poos *et al.*, 2016).

To focus on regulators which highly influence the expression level of the predicted *EST* gene, regulators with

(i) a strong expression level in the knockout sample (absolute z-score > 1),
(ii) less than 1,000 putative target genes or
(iii) a *TLM* annotation

were selected as most interesting. Regulators fulfilling criteria (i) and (ii) are marked in bold in Table 4, while regulators with a *TLM* annotation (criteria (iii)) are additionally marked in red (short *TLM* genes) or blue (long *TLM* genes). Several of the significant *EST* regulators are by themselves *TLM* genes and affect telomere length after mutation. The identified regulators Sum1, Hst1, Srb2 and Sin3 led to telomere shortening when mutated (Askree *et al.*, 2004; Gatbonton *et al.*, 2006), while a deletion of *DIG1* showed longer telomeres (Gatbonton *et al.*, 2006). A positive z-score indicates an upregulated expression of the *EST* gene after the regulator knockout, suggesting that this regulator is an inhibitor of the gene. A negative z-score shows the opposite effect that this regulator is an activator of the corresponding *EST* gene.

Table 4. Significant regulators of the three *EST* genes for the short *tlm* samples compared to samples with wild-type telomere length (controls).

	Regulator	Z-score*	Significance (P)**	Number of targets
EST1	**Sum1*****	6.85	1.96 E-29	579
	Hst1***	3.61	1.96 E-29	219
	Msn4	-0.63	7.61 E-13	2483
	Mig1	0.14	2.48 E-11	423
	Gcn4	-0.13	2.64 E-10	2712
	Ste12	-****	1.28 E-9	3673
	Rfx1	-0.45	1.51 E-8	660
	Srb2***	2.08	1.14 E-7	785
	Sfp1	3.24	4.86 E-4	4199
	Cup2	-0.30	3.18 E-3	548
	Swi3	2.69	9.43 E-3	1737
	Mbp1	0.76	3.71 E-2	665
EST2	Gcn4	-0.22	9.47 E-16	2712
	Gln3	-2.67	1.57 E-12	981
	Rme1	-0.31	6.33 E-11	399
	Yrm1	-	4.50 E-10	2509
	Pdr3	-0.44	3.04 E-9	929
	Msn4	-1.17	5.48 E-9	2483
	Msn2	0.05	5.48 E-9	3260
	Pdr1	0.31	1.67 E-8	1318
	Arg81	0.42	1.12 E-7	335
	Ste12	-	4.78 E-7	3673
	Rtg3	0.16	8.55 E-7	646

	Tec1	-0.22	1.53 E-6	3669
	Sfp1	-0.06	3.12 E-5	4199
	Abf1	-0.29	3.20 E-5	2715
	Swi5	-2.68	3.79 E-5	1871
	Ace2	-1.70	2.91 E-4	4683
	Nrg2	-	4.19 E-2	331
EST3	**Dig1*****	-1.87	1.77 E-24	334
	Sok2	0.28	8.61 E-24	2160
	Sin3***	-3.11	4.62 E-16	1759
	Msn2	-0.46	5.27 E-14	3260
	Ste12	-	2.38 E-12	3673
	Ixr1	-0.17	4.17 E-11	1633
	Msn4	0.62	9.20 E-10	2483
	Mga1	-0.38	1.09 E-8	674
	Hir1	-2.19	3.55 E-5	306
	Srb2***	-2.38	4.67 E-4	785
	Ume6	3.64	5.50 E-4	826
	Ace2	1.05	1.93 E-2	4683

* Effect of the knockout of the regulator on the expression of the *EST* genes (positive z-score = up-regulation of the corresponding *EST* gene; negative z-score = down-regulation of the corresponding *EST* gene); ** Multiple testing corrected (Benjamini-Hochberg); *** red: short *tlm* mutant, blue: long *tlm* mutant; **** For some genes, no expression data was available.

For *EST1*, Sum1 (P=1.96 E-29), Hst1 (P=1.96 E-29) and Srb2 (P=1.14 E-7) were identified as most important regulators of the short *tlm* mutants. Investigating the literature (Pubmed, www.ncbi.org) for the selected regulators with the keywords “telomere”, “telomerase” and each of the *EST* gene symbols, Sum1 was identified as a chromatin silencing factor and initiation factor of replication. Furthermore, Sum1 plays a role in the regulation of middle-sporulation genes and in a complex together with Hst1 and Rfm1 it represses genes through histone deacetylation (Bedalov *et al.*, 2003; Li *et al.*, 2013; McCord *et al.*, 2003; Zill and Rine, 2008). Sum1, Hst1 and Sir2 have specific co-enriched binding sites and can interact with Rap1 indicating that they play similar roles in telomere maintenance (Bedalov *et al.*, 2003; Li *et al.*, 2013; McCord *et al.*, 2003; Zill and Rine, 2008). Besides Sum1 and Hst1, Srb2, a subunit of the RNA polymerase II mediator complex, was identified as an important regulator of *EST1* and *EST3*. Regarding telomere maintenance, Srb2 has been reported to play a direct role in TLC1 transcription or an indirect role in TLC1

accumulation (Mozdy *et al.*, 2008). EST1 was highly upregulated in the *sum1*, *hst1* and *srb2* deletion strains. For *EST2*, among the significant regulators, only Gln3 had a strong knockout effect and less than 1,000 target genes. Gln3 controls the level of the Ku complex and is therefore involved in telomere shortening upon starvation by Tor Complex 1 (TORC1) (Ungar *et al.*, 2011). As significant regulators of *EST3*, Ume6, Sin3, Srb2, Hir1 and Dig1 were identified by the modeling approach. While *EST1* is upregulated in the srb2 knockout strain, *EST3* showed the opposite effect. From the above-mentioned significant regulators of *EST3* only Ume6 had an inhibitory effect on *EST3* expression, all others were activators of *EST3*. Sin3 has been reported to form histone deacetylase complexes together with Rpd2 and Rpd3. Sin3 can act as an activator or a repressor of transcription (Sun and Hampsey, 1999) and is involved in gene silencing, DNA repair processes as well as telomere maintenance. Sin3 and Rpd3 together affect silencing at telomeres (Sun and Hampsey, 1999). Ume6 is also an important regulator of early meiotic genes and can interact with Rpd3, similar to Sin3. Ume6 further plays a role in chromatin remodeling and recruits Rpd3 and Sin3 to form the histone deacetylase complex (Kadosh and Struhl, 1997).

In summary, the MIPRIP analysis led to promising regulators of the *EST* genes and several of them are involved in the regulation of chromatin and histone modifications.

3.2.2 Validation of the regulators identified for *EST1*

EST1 showed the highest expression level in the *sum1* knockout strain (z-score=6.85, $\log_2$-fold change = 0.926, Figure 16A) compared to all other regulator knockouts in the dataset. Furthermore, *EST1* was upregulated in the knockout strains of the two other regulators *hst1* (z-score=3.61) and *srb2* (z-score=2.08), which showed that all three regulator candidates are putative inhibitors of *EST1* expression. This inhibitory effect is surprising because the knockout strains of *SUM1*, *HST1* and *SRB2* showed a shorter telomere length compared to the control samples.

To validate the effect of Sum1 on *EST1* expression levels, our collaboration partner Andre Maicher from the group of Prof. Luke (IMB, Mainz) performed expression studies on two *sum1* knockout strains. Because it has been reported that Sum1 is involved in the regulation of the mating-type of the yeast strain (Chi and Shore,

1996), the expression of *EST1* was investigated in a wild-type and a *sum1* mutant of both mating-types, Mat a and Mat α, by RT-PCR. In both mating types, *EST1* was highly upregulated in the *sum1* mutant (4.37-fold ± 0.67 SEM for Mat a and 6.00-fold ± 0.48 SEM for Mat α) (Figure 16B) validating the inhibitory effect of Sum1 on *EST1* expression (Poos *et al.*, 2016).

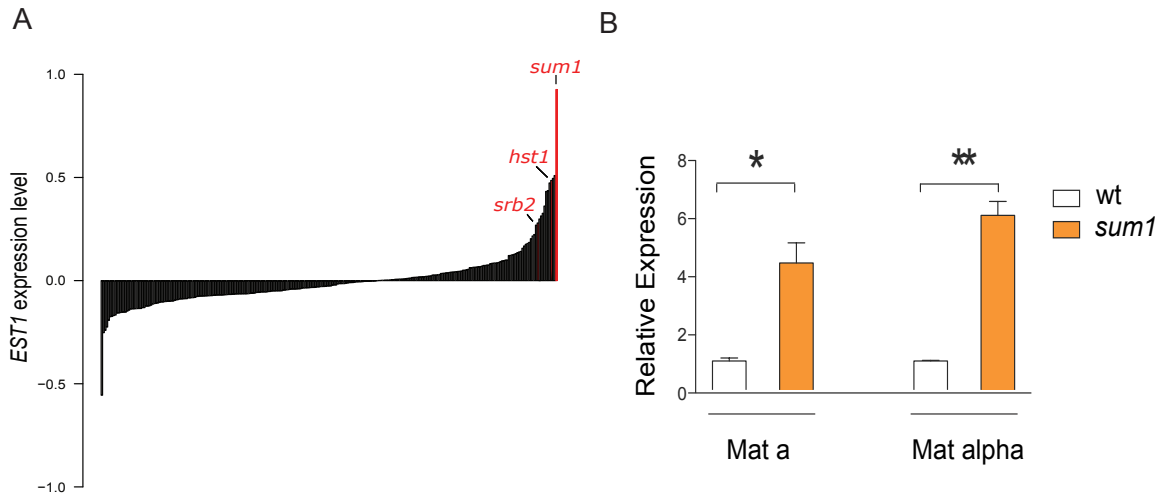

Figure 16. *EST1* expression

(A) in all regulator deletion strains from the microarray data (log_2-fold change) and (B) in wild-type and *sum1* mutants of both mating types (left: mating-type a; right: mating-type α) (measured by RT-qPCR and normalized to actin). The experiment was performed in three replicates and the error bars indicate the standard error (SEM). Significant differences were observed with a two-tailed unpaired t-test with Welch´s correction. Images taken from (Poos *et al.*, 2016).

3.3 Investigating telomerase regulation across 19 different human cancer types

Telomere maintenance is one of the hallmarks of cancer which enables replicative immortality. I used gene expression data of 19 different primary cancer types (Table 2) from TCGA to study the regulation of the telomerase reverse transcriptase (*TERT*) gene. Based on the generic human regulatory network, which we generated based on ChIP-binding data from different sources, publications and computational predictions, I identified 75 putative regulators binding to the *TERT* gene's promoter (Table S3). From the 75 putative regulators 60 were extracted from the manually curated database MetaCore™, and 30 were extracted from our generic regulatory network that were also described elsewhere (Ramlee *et al.*, 2016). The overlap between both studies was of significance (Fisher's Exact Test *p*=6.01E-23). Nearly

all of the overlapping regulators were found in MetaCore™, only CTCF was extracted from Encode data as a *TERT* regulator.

3.3.1 MIPRIP analysis of *TERT* in 19 different subtypes

To study the regulation of *TERT* in the selected 19 different cancer types, I performed a multi-mode MIPRIP-Comp analysis. For each cancer entity regulatory models for *TERT* were constructed by using a ten-times three-fold cross-validation. Furthermore, the number of regulators was restricted from 1-10 regulators. In total, 300 different models were created for each cancer type. The correlation between the predicted expression value of *TERT* and the expression value of *TERT* from the RNA-seq data was calculated, to get an estimate of the performance of the models. The obtained performance was r=0.4 or higher for most of the cancer types (Figure 17A).

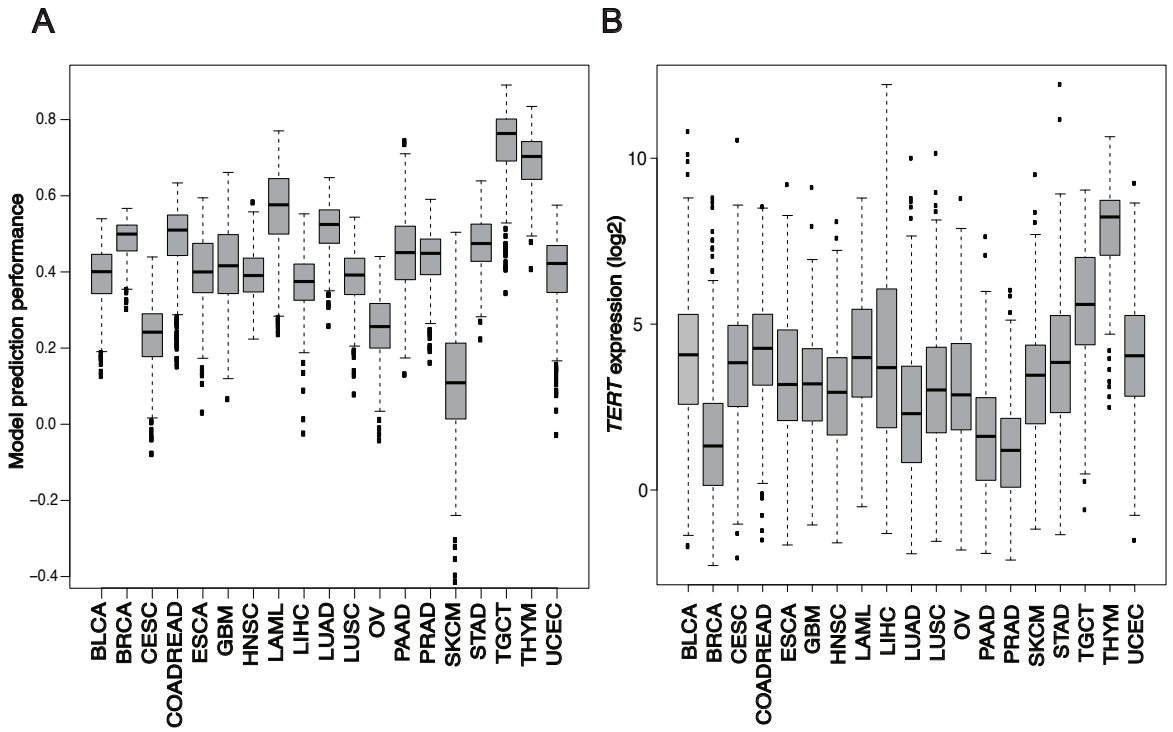

Figure 17. Prediction performance and *TERT* expression in the different cancer entities

(A) Performance of the regulatory models for *TERT* and (B) expression of *TERT* in all the 19 different cancer entities. Plots taken from (Poos *et al.*, 2019).

The best performance was observed for testicular germ cell cancer (TGCT) (r=0.75) and thymoma (THYM) (r=0.7), while cervical (CESC), ovary (OV) and melanoma skin (SKCM) cancer showed the lowest performance. Furthermore, I investigated the expression of *TERT* in all the 19 different cancer types. Here, I observed the highest *TERT* expression in TCGT and THYM and the lowest *TERT* expression in breast (BRCA), pancreas (PAAD) and prostate (PRAD) cancer (Figure 17B). Looking at the modeling performance in relation to *TERT* expression, best performance was observed for the two cancer entities (TCGT and THYM) with the highest *TERT* expression. However, a poor performance was not associated with an extremely low *TERT* expression. While the expression of *TERT* in SKCM was comparable to the other cancer entities, the performance of the regulatory models was poorest (r=0.1). One reason for this could be that SKCM is a cancer entity with a high frequency of *TERT* promoter mutations indicating that there exist cancer subtypes with different *TERT* regulation processes which was investigated in Results section 3.3.2. The significant *TERT* regulators of each cancer entity are listed in Table S4-Table S21.

3.3.2 Common *TERT* regulators over all 19 different cancer types

A Rank product test was performed to identify significant common regulators. For this purpose, the regulators were ordered for each cancer entity based on their frequency in the 300 models. These ranked lists were then compared, and significance was tested based on a permutation-test. This led to the following nine common *TERT* regulators (Table 5): The Paired Box Proteins PAX5 and PAX8, the E2F factors 2 and 4, AR, BATF, SMARCB1, TAF1 as well as MXI1.

Table 5. Predicted *TERT* regulators common to all 19 different cancer entities

TF	E-value
E2F4	0
AR	1.00 E-04
PAX5	4.00 E-04
E2F2	6.00 E-04
BATF	3.20 E-03
PAX8	6.30 E-03
SMARCB1	1.38 E-02
MXI1	1.87 E-02
TAF1	2.12 E-02

To validate the identified common *TERT* regulators with data from the literature, a Pubmed search (https://www.ncbi.nlm.nih.gov/pubmed/) was performed. For this purpose, all regulator gene symbols from Table 5 were queried together with "TERT" and the terms “telomerase”, “human” and “regulation”: (E2F4 OR AR OR PAX5 OR E2F2 OR BATF OR PAX8 OR SMARCB1 OR MXI1 OR TAF1) AND TERT AND telomerase AND human AND regulation. The received number of Pubmed hits was compared to a query without the common regulators (TERT AND telomerase AND human AND regulation). As background the same two queries were performed without the “TERT” gene symbol. With the results of the Pubmed hits a Fisher’s exact test was used to validate if the nine as common identified regulators were significantly more often found together with *TERT* than without *TERT*. For the identified common *TERT* regulators 21 out of 1,002 articles of *TERT* were found in Pubmed indicating a significant enrichment (*p*-value=0.013, Table 6).

Table 6. Confusion matrix for the Fisher’s Exact Test based on the results of the Pubmed query

	Found with the query containing the nine predicted regulators	**Found only with the query which did not contain the nine predicted regulators**
Query with “TERT”	21	981
Query w/o “TERT”	25	2,483

3.3.3 Melanoma skin cancer as a cancer entity with a high fraction of *TERT* promoter mutations

For melanoma skin cancer (SKCM) a high frequency of *TERT* promoter mutations has been discovered. The promoter mutation occurs mainly at position 124 bp and 146 bp upstream of the translational start codon and generates a further binding site for TFs of the ETS-family (Horn *et al.*, 2013; Huang *et al.*, 2013). As described in the multi-mode MIPRIP-Comp analysis above SKCM was the cancer entity with the lowest prediction performance. Because of the high rate of *TERT* promoter mutations in SKCM and the observation that samples with a *TERT* promoter mutation lead to a higher *TERT* expression (Figure 18A, *p*-value=5.33 E-03), it was expected that there exist different regulatory mechanisms in the samples with and

without a *TERT* promoter mutation. This would lead to different regulatory subtypes and would be in line with a very low modeling performance.

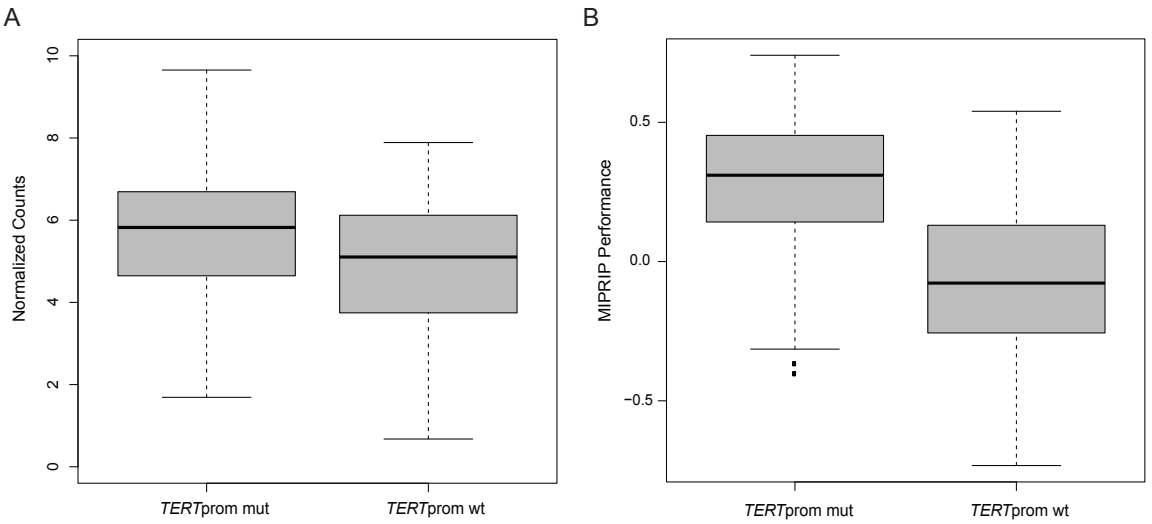

Figure 18. Melanoma samples with and without *TERT* promoter mutation
(A) *TERT* expression in the SKCM samples with a *TERT* promoter mutation (mut) and without *TERT* promoter mutation (wt). (B) Performance of the dual-mode MIPRIP analysis of the *TERT* promoter mutated and wild-type SKCM samples.

To study the regulatory processes of *TERT* in SKCM samples with a *TERT* promoter mutation and with wild-type *TERT* promoter, I performed a dual-mode MIPRIP-Comp analysis. Based on the study of (Cancer Genome Atlas, 2015) the *TERT* promoter status was available for 115 SKCM patients from the TCGA cohort. From these 115 samples 74 showed a *TERT* promoter mutation and 41 had a wild-type *TERT* promoter. The dual-mode MIPRIP analysis was performed with these two groups and a ten-times three-fold cross-validation together with a restriction of regulators from 1 up to 10. This leads to a much better performance for the samples with a *TERT* promoter mutation (r=0.3), while for the samples without the *TERT* promoter mutation the performance was still very low (r= -0.1) (Figure 18B). Looking at the regulators which were significantly used more often in the models of the samples with compared to the samples without *TERT* promoter mutation, I identified 12 regulators as significantly more often used in the mutated group and 17 for the wild-type group (Table 7). For the mutated group AR, E2F1, JUND and ETS1 were the most significant hits, while HMGA2, HIF1, RUNX2 and TAL1 were highly significant for the wild-type samples. Especially ETS1 was in line with the literature

because a *TERT* promoter mutation generates an ETS1 binding site at the mutated position (Horn *et al.*, 2013; Huang *et al.*, 2013). This study showed that splitting up the dataset into cancer subtypes leads to more reliable results and also better prediction performances. In summary, the dual-mode MIPRIP-Comp analysis is well suited to study regulatory processes between two conditions.

Table 7. *TERT* regulators of melanoma samples with (mut) and without (wt) *TERT* promoter mutation

Regulators in mut	P-value	Regulators in wt	p-value
AR	3.97 E-37	HMGA2	1.05 E-16
E2F1	3.00 E-29	HIF.1	1.03 E-15
JUND	2.86 E-25	RUNX2	2.88 E-12
SMARCB1	1.85 E-15	TAL1	1.54 E-09
ETS1	4.46 E-13	ESR2	3.92 E-09
SIN3AK20	1.42 E-06	AP-2	3.28 E-06
REST	3.64 E-06	MITF	2.59 E-05
MAZ	7.85 E-06	WT1	2.80 E-05
E2F2	9.20 E-05	SMAD3	5.76 E-05
TAF1	1.36 E-04	TFAP2D	2.78 E-04
BCL11A	4.31 E-04	PAX8	6.04 E-04
MYB	4.73 E-04	GRHL2	7.16 E-04
		TP53	1.21 E-03
		TCF7	1.68 E-03
		MZF1	3.44 E-03
		TFAP2C	3.68 E-03
		NR2F2	7.96 E-03

An ETS1 knockdown in a melanoma cell line with a *TERT* promoter mutation (Wang *et al.*, 2012) showed that the *TERT* expression is lower in the ETS1 knockdown cells compared to controls (fold change: 0.82). This indicates that ETS1 is involved in *TERT* expression in samples with a *TERT* promoter mutation as suggested by our modeling analysis.

As comparison to MIPRIP, the same melanoma samples with and without *TERT* promoter mutation were used for an analysis with the well-established tool ISMARA (Balwierz *et al.*, 2014). Similar to MIPRIP, ISMARA calculates the activity of regulators based on their target genes, but here the target genes are inferred from motif binding information. ISMARA calculated the activities of each regulator separate for each sample and performed then an average estimation over all samples of each group (*TERT* promoter with *vs.* without mutation). For the

melanoma samples with *TERT* promoter mutation ISMARA predicted 20 regulators for *TERT* (Table 8). Compared to the MIPRIP results, only the three regulators SIN3A, MAZ and WT1 were found with both tools. A *TERT* promoter mutation leads to a further binding site of TFs of the ETS family (Horn *et al.*, 2013; Huang *et al.*, 2013). From the ETS family of TFs, ISMARA predicted GABPA, ELF2 and ELF5 with very low significance, while MIPRIP identified ETS1 as a highly significant regulator of *TERT* in the melanoma samples with a *TERT* promoter mutation. In summary, the overlap between MIPRIP and ISMARA was low and especially the highly significant MIPRIP hit ETS1 was in high agreement with the literature.

Table 8. *TERT* regulators predicted with ISMARA for the SKCM gene expression data of samples with and without a *TERT* promoter mutation.

Regulator	Score
MXI1_MYC_MYCN	3.40
KLF16_SP2	1.48
ELK4_ETV5_ELK1_ELK3_ELF4	1.41
PLAGL1	1.37
SIX4	1.31
GMEB2	1.11
MNT_HEY1_HEY2	1.05
TCF12_ASCL2	1.05
CTCF_CTCFL	0.85
RCOR1_MTA3	0.85
SIN3A_CHD1	0.84
ARNT	0.73
MAZ_ZNF281_GTF2F1	0.41
AHR_ARNT2	0.38
HES1	0.33
TCF3_MYOG	0.32
IKZF1	0.23
MYF6	0.10
WT1_MTF1_ZBTB7B	0.04
ELF2_GABPA_ELF5	0.01

Regulators marked in bold were also identified with MIPRIP as significant between both groups. For example, MXI1_MYC_MYCN indicates the motif name. The associated genes of this motif are Mxi, Myc and Mycn. All 3 have the same motif. The Score of each motif indicates the significance level that ISMARA assigns to each motif based on the gene expression data and is quantified as Z-value.

3.3.4 Prostate cancer as a cancer entity with neither *TERT* promoter mutations nor ALT occurrence

So far, in prostate cancer neither *TERT* promoter mutations nor ALT occurrence has been reported (Heaphy *et al.*, 2011). Therefore, prostate cancer is well suited to study the regulation of the telomerase. To identify the most important regulators of *TERT* in prostate cancer, I performed a dual-mode MIPRIP-Comp analysis for prostate cancer compared to healthy prostate samples from TCGA. As in the pan-cancer analysis models were constructed for 1 up to 10 regulators and using a ten-times three-fold cross-validation. This resulted in 17 regulators which were used significantly more often in the prostate cancer models compared to the normal prostate models (Table 9) and 40 significant regulators for the normal prostate models *versus* the prostate cancer models (Table S22). The regulators PITX1, MITF, AR and TFAP2C were identified as most significant *TERT* regulators in prostate cancer, while TAF9, AP-2, ETS2 and HIF1A were significantly more often used in healthy prostate tissue models compared to the prostate cancer models.

Table 9. Significant *TERT* regulators of prostate cancer *versus* normal prostate tissue.

Regulators	Frequency	Frequency	p-value
PITX1	186	35	1.56 E-37
MITF	119	28	5.97 E-17
AR	92	21	1.26 E-12
TFAP2C	72	11	1.67 E-12
E2F2	92	24	1.31 E-11
NR2F2	97	27	1.31 E-11
SMARCB1	88	24	1.15 E-10
CEBPA	65	20	6.08 E-07
BHLHE40	53	16	8.26 E-06
CTCF	48	15	4.13 E-05
ETS1	63	26	7.43 E-05
MXI1	27	5	1.75 E-04
POLR2A	34	9	2.23 E-04
RAD21	32	11	2.37 E-03
IRF1	31	12	6.38 E-03
TFAP2D	34	18	3.91 E-02
MAX	36	20	4.62 E-02

The regulators which were specific for prostate cancer in the pan-cancer analysis are marked in bold.

In the pan-cancer analysis above, I predicted the significant *TERT* regulators for each cancer type compared to all other cancer types. Specific for prostate cancer I identified 17 significant *TERT* regulators (Table 10). 12 out of these 17 prostate cancer specific *TERT* regulators were also significant for prostate cancer when compared to healthy prostate tissue (from the dual-mode MIPRIP analysis). KLF2 (*p*=7.09E-04), TFAP2A (*p*=4.75E-02), ZBTB48 (*p*=8.13E-02), MEN1 (*p*=3.37E-02) as well as the NF-κB complex (NFKB.P50.P65) (*p*=1.62 E-02) were additionally identified in the pan-cancer analysis as specific for prostate cancer. From these additional regulators ZBTB48 and NFKB.P50.P65 were used more often in the normal prostate models than in the cancer models, while KLF2, MEN1 and TFAP2A were not significant at all for prostate cancer versus normal prostate tissue. PITX1 was the most significant *TERT* regulator of prostate cancer in the dual-mode and in the multi-mode MIPRIP analysis.

Table 10. Specific *TERT* regulators of prostate cancer *versus* all other cancer types based on the multi-mode MIPRIP analysis.

TF	Adjusted *p*-value
PITX1	2.79 E-21
ETS1	3.04 E-19
MITF	2.56 E-17
NR2F2	8.28 E-16
IRF1	3.38 E-13
TFAP2D	4.24 E-10
CEBPA	2.09 E-08
E2F2	1.02 E-07
BHLHE40	5.67 E-06
KLF2	7.09 E-04
TFAP2C	2.35 E-03
AR	5.22 E-03
ZBTB48	8.13 E-03
NFKB.P50.P65	1.62 E-02
CTCF	2.48 E-02
MEN1	3.37 E-02
TFAP2A	4.75 E-02

The overlap between the pan-cancer analysis and the prostate cancer *vs*. healthy prostate tissue MIPRIP analysis is marked in bold.

In the multi-mode MIPRIP analysis, besides prostate cancer, PITX1 was a significant *TERT* regulator only in head and neck carcinoma (HNSC, p=1.46 E-16, Table S10), ovary (OV, p=2.21 E-02, Table S15) and cervical cancer (CESC, p=1.19 E-03, Table S6).
In summary, the 12 *TERT* regulators overlapping in both MIPRIP studies seem to be highly specific for prostate cancer and were used for further modeling.

3.4 MIPRIP-Net identifies the gene regulatory network of *TERT* in prostate cancer

In the MIPRIP-Comp analysis above I identified 12 direct regulators of *TERT* that are highly specific for prostate cancer. These regulators were used for the MIPRIP-Net analysis to identify a highly connected subnetwork that predicts the regulation of *TERT* best.

3.4.1 MIPRIP-Net analysis

To get a broader view on the *TERT* regulation, I applied the newly developed MIPRIP-Net approach to the prostate cancer data. The goal was to integrate direct and indirect regulators (e.g. co-regulators) that are involved in the expression of *TERT* into one regulatory subnetwork. Direct regulators mean that these regulators can bind to the *TERT* promoter and were extracted from the generic human regulatory network. Indirect regulators bind to the direct regulators, but not directly to *TERT*. MIPRIP-Net is a combination of the MIPRIP approach and modularity. This means that on the one hand the regulators were identified that can best predict the expression of *TERT*, and on the other hand that the regulators are highly connected within each other and with their regulators.
To limit the number of indirect regulators, a basic MIPRIP analysis was performed separately for each of the 12 regulators with the same parameter setting as described above. To note, for TFAP2D no MIPRIP analysis was possible because for TFAP2D no gene expression data was available. From the MIPRIP runs, the regulators that were used in at least 20 % of the models (Table 11) were selected for the MIPRIP-Net approach. This resulted in 72 additional (indirect) regulators. Some of the selected regulators can also bind directly to the *TERT* promoter and

are marked in red in *Table 11*. The highest overlap between the selected regulators and putative *TERT* regulators was found for PITX1, MITF, NR2F2 and IRF1 (Table 11).

Table 11. MIPRIP-Comp dual-mode analysis of the 12 significant *TERT* regulators identified specifically for prostate cancer.

	Regulators used in at least 20% of the models	# regulators	# TERT regulators
PITX1	SMARCC1, TAF1, HEY1, POLR2A, FOXO1, HNF4A, ESR1, RBBP5, SMAD1, SMARCB1	10	5
AR	MAFF, MAFK, ZBTB17, CREB3, GATA2, TCF4, CTCF, EGR1	8	2
MITF	MXI1, ZNF263, SMC3, TAL1, MYC, EP300, MAX	7	4
CTCF	MAX, PRDM16, YY1, RBBP5, REST, POU2F2, FOXP2, EP300	8	3
BHLHE40	ARNTL, HIF1A, SIN3AK20, EGR1, NCOR1, AR, CEBPB, GABPA, ZNF143	9	4
ETS1	ETV2, PAX5, FOS, CEBPB, USF1, FOXA1, TCF7L2, IRF4, GATA2	9	1
CEBPA	SP1, CLOCK, IKZF1, MYC, NCOR1, FOXP2, JUN, SREBF1, MAZ	9	3
E2F2	E2F4, PML, E2F7, MAFK, ELF1, HEY1, EBF1, E2F6, MAFF, TCF12	10	4
NR2F2	MXI1, TP53, USF1, E2F4, SF1, FOXP2, SIN3AK20, ZNF263	8	4
IRF1	NFKB.P50.P65, IRF2, SPI1, EGR1, MYB	5	3
TFAP2C	TP63, MAX, RAD21, RBPJ, SP1, POU5F1, ZFP36L1, MTA1, E2F1, EZH2, SETDB1	11	3

The regulators that can bind to the *TERT* promoter are marked in red. For TFAP2D no gene expression data was available and hence no MIPRIP model could be constructed. Therefore, regulators of TFAP2D were only included in the MIPRIP-Net model if they were found in at least 20% of the models of the other regulators.

The MIPRIP-Net models were constructed with the 12 prostate cancer specific *TERT* regulators and the 72 additional regulators by restricting the number of regulators from 2 up to 20 and performing a ten-times three-fold cross-validation. The performance of the MIPRIP-Net models was estimated from MIPRIP, but the combination with the modularity influences the regulators which were selected by the MIPRIP model. These two objectives (good model for MIPRIP and good model for the regulatory network) were combined by a tradeoff parameter λ which had to be optimized first for each gene of interest and the used dataset. For this

optimization, I constructed MIPRIP-Net models for 9 different λ parameters (0.001, 0.01, 0.1, 0.3, 1, 3, 10, 100 and 1000). For each λ the performance over all models and the number of regulators was counted separately for the MIPRIP and the modularity part of the combined model.

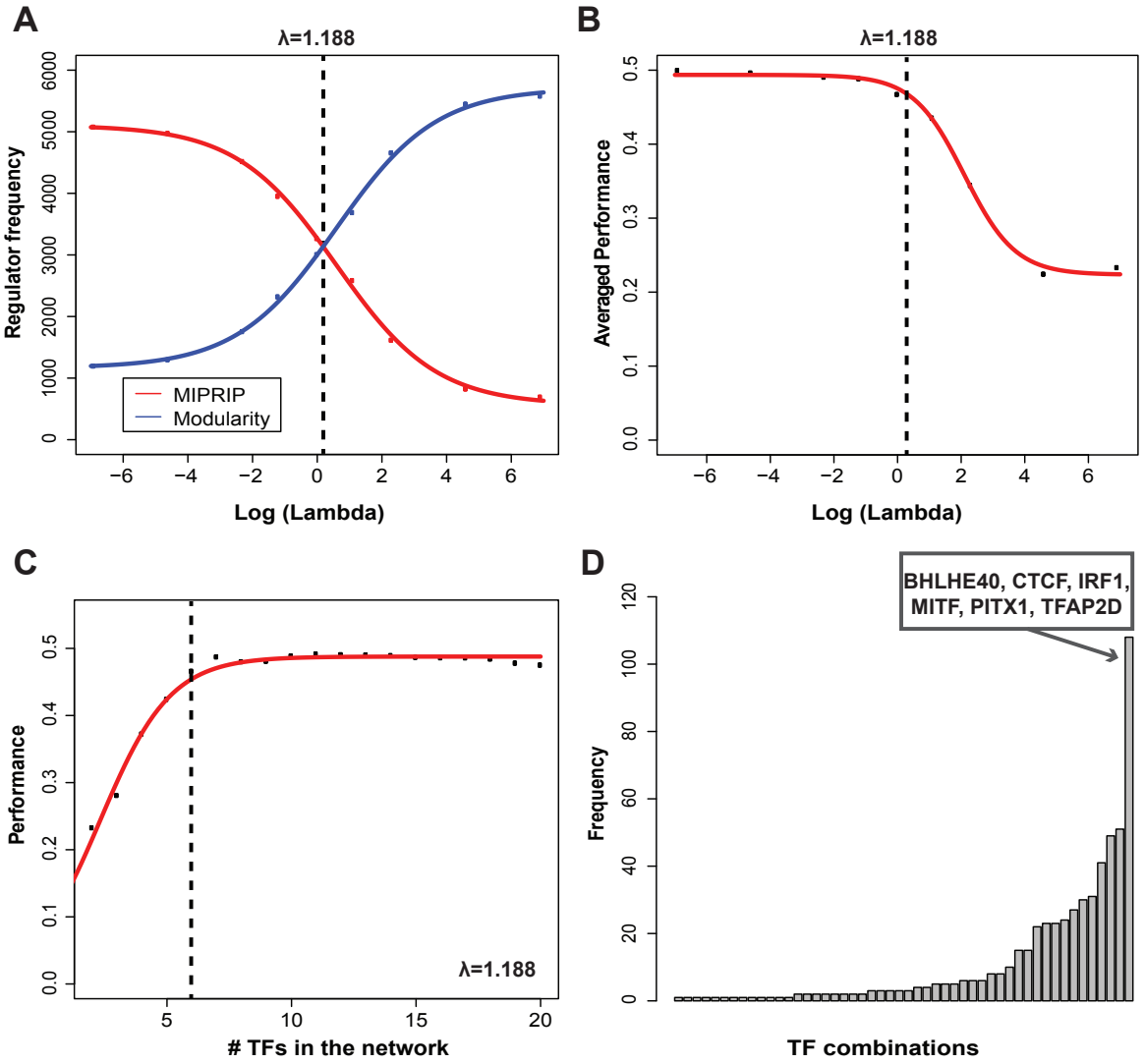

Figure 19. Optimization of the MIPRIP-Net models.

(A) The number of selected regulators from the MIPRIP model (red curve) and from the modularity model (blue model) are plotted for different λ parameters. The λ value was determined where both curves were intersecting. (B) The performance over all models was plotted for the different λ-values. The determined λ from plot (A) showed a good performance. (C) Models were generated with λ=1.188 and the performance was plotted. At least six MIPRIP regulators are necessary to get a good prediction of *TERT* expression. (D) Histogram which MIPRIP regulator combination was used most often over all models.

Plotting the counted number of the MIPRIP (red curve) and the modularity (blue curve) part of the model for the 9 different λ values, both curves were intersecting at λ = 1.188 (Figure 19A). Therefore, the optimized λ = 1.188 leads to a balance between both objectives. For low λ values the MIPRIP-Net models were dominated by MIPRIP, while high λ values led to a modularity driven regulator selection. This means that there would be many indirect and only very few direct *TERT* regulators, which led to a very low performance (Figure 19B). For λ values of 1, which would mean that the MIPRIP and the modularity optimization were weighted equally, the performance was constant at r=0.5. But for higher λ values the performance drops down to around r=0.25 (Figure 19B). The optimized λ=1.188 from Figure 19A still led to a performance of r=0.48 and was used to identify the best subnetwork of *TERT*. For this, further MIPRIP-Net models were constructed with λ=1.188. As shown in (Figure 19C) at least six selected MIPRIP regulators were necessary to get a good prediction of *TERT* expression (r=0.45). I then counted which combination of six MIPRIP regulators was selected most often over all models (Figure 19D). This led to a distinct result. The combination “BHLHE40, CTCF, IRF1, MITF, PITX1 and TFAP2D” was selected in 108 of the 570 models.

I then investigated which combination of modularity regulators were selected most often together with this combination of direct regulators. This led to the following 14 regulators: E2F4, MAZ, POLR2A, POU2F2, SMARCB1, TAF1 and REST, which can also bind to the *TERT* promoter, but were not selected by MIPRIP, and the indirect *TERT* regulators EP300, MAFK, PML, SMC3, USF1, YY1 and ZNF263. The six significant regulators of the MIPRIP-Comp analysis (marked in red) and the putative *TERT* regulators (orange) as well as the additional regulators (grey) from the modularity part are visualized in a regulatory network explaining *TERT* regulation (Figure 20). The edge weights of the interactions between the significant and putative *TERT* regulators and the *TERT* gene are based on the scores of the generic human regulatory network, while the connections between the regulators are weighted based on the correlations of their activity profiles multiplied with the edge scores in the generic network. The putative *TERT* regulators from the modularity approach (marked in orange) were not identified as prostate cancer specific *TERT* regulators above, but they had several interactions to the significant *TERT* regulators of the MIPRIP-Comp analysis. To investigate if the additional

regulators are also known to be linked to telomere maintenance, the TelNet database (Braun *et al.*, 2018) was queried. The TelNet database is a large collection of manually curated genes that play a role in telomere biology. From the indirect regulators PML, SMC3 and USF1 were found in TelNet. Based on the cBioPortal for Cancer Genomics (http://www.cbioportal.org/) I investigated if the regulators in the network are frequently mutated or deleted in prostate cancer. Here, I found that POLR2A was deleted in around 8 % of all prostate cancer patients.

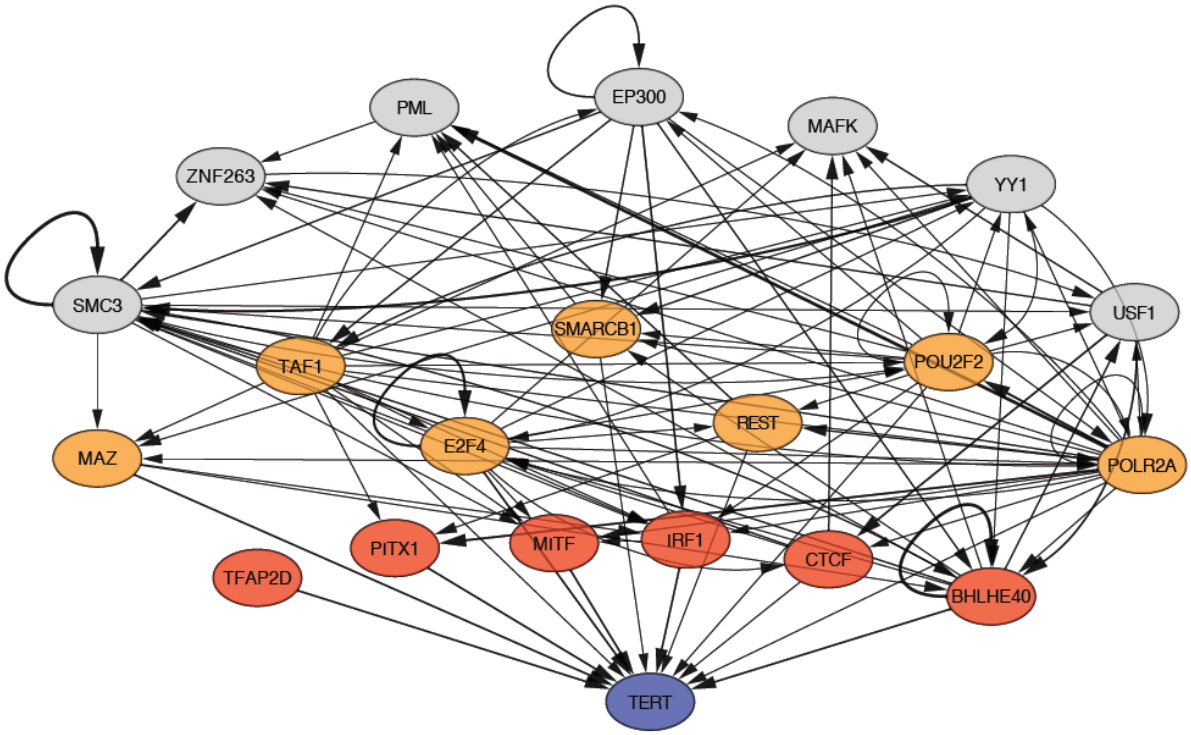

Figure 20. Regulatory network model best explaining *TERT* regulation in prostate cancer constructed with MIPRIP-Net.

Significant regulators from the MIPRIP-Comp analysis are marked in red, while the regulators added by the modularity approach are marked in orange if they can bind to the *TERT* promoter and in grey if they only bind to the other regulators. The edges between the significant and putative regulators and *TERT* are weighted based on the generic regulatory network, while the interactions between the regulators are weighted based on the correlation of their activity values multiplied with the scores in the generic regulatory network.

In summary, using the newly developed MIPRIP-Net approach I identified a subnetwork of 20 regulators best explaining the regulation of *TERT*. From these regulators BHLHE40, CTCF, IRF2, MITF, PITX1 and TFAP2D play a direct role and are studied in more detail in the next section.

3.4.2 Clinical validation of the identified *TERT* regulators

The 12 prostate cancer specific regulators of *TERT* from the MIPRIP-Comp analysis were validated in the department of Guido Sauter/Ronald Simon at the University Hospital in Hamburg-Eppendorf for their potential as prognostic markers. A special focus was on the 6 significant direct *TERT* regulators, which were part of the regulatory subnetwork in Figure 20. Using tissue-microarrays (TMAs) of 17,747 patients an immuno-histochemical (IHC) analysis was performed for the regulators and the staining intensities (mostly negative, low, high) were correlated to histopathological and molecular features (e.g. ERG-fusion gene, *PTEN* deletions).
As PITX1 was the most significant *TERT* regulator in prostate cancer, it was analyzed first. PITX1 was upregulated in around two-thirds of the tumor samples compared to normal prostate epithelial tissue. But over all evaluated tumor samples PITX1 was highly expressed in only 4 % of the tumor samples (Table S23). PITX1 upregulation was tested to be associated with several features indicating tumor aggressiveness (high Gleason grade, advanced tumor stage, presence of lymph node metastasis, higher pre-operative PSA-levels, positive surgical margin (histological presence of cancer cells at the inked margin of the radical prostatectomy specimen)). Comparing the PITX1 expression in ERG-fusion positive and negative tumor samples, we found out that PITX1 was upregulated in around 80 % of the ERG-fusion positive samples, while only 55 % of the ERG-fusion negative samples showed a PITX1 expression. Furthermore, we found out that ERG-negative tumor samples were genomic instable because they were associated with 10q23 (*PTEN*), 5q21 (*CHD1*), 6q15 (*MAP3K7*) and 3p13 (*FOXP1*) deletions, while in ERG-fusion positive tumors we only found a slightly upregulation of PITX1 in samples with these deletions (Figure S1). PITX1 upregulation is associated with a higher cell proliferation independent of the Gleason score. The Kaplan-Meier analysis revealed that patients with a high PITX1 expression have a poorer PSA-recurrence free survival compared to patients with a low or negative PITX1 expression (p-value < 0.0001) (Figure 21A). The patient's outcome was not dependent on the ERG-fusion status. To estimate if PITX1 can provide an added value to the established clinical-pathological prognostic parameters, four different multivariate models were calculated to resemble typical scenarios. Scenario 1 utilizes all available parameters after radical prostatectomy (pathological tumor stage, Gleason grade, lymph node and surgical margin status as well as

preoperative PSA level) and PITX1 expression. Scenario 2 excludes the nodal status because lymph node dissection is not standardized in the surgical prostate cancer therapy. To model the preoperative situation, Scenario 3 was set up. It included PITX1 expression, clinical tumor stage, preoperative PSA level and the Gleason grade obtained from the prostatectomy specimens, while in Scenario 4 the preoperative Gleason grade obtained from the biopsies were used. In general, the postoperative determination of the Gleason grade is more precise than the preoperative determination (Epstein *et al.*, 2012). It turned out that PITX1 expression gave a significant added value in the Cox proportional hazards regression analysis for all four scenarios in the ERG-negative tumor samples, while in the ERG-positive tumor samples it yielded only a significant added value in the preoperative stage (scenario 4). This shows that PITX1 is a good prognostic marker, especially at the biopsy level and before surgery.

For IRF1 only around 2 % of the stained tumor samples showed a high IRF1 expression among all tumor samples, while TFAP2D was upregulated in around 75 % of the tumor samples. For both regulators an upregulated expression is associated with a poorer prognosis (Figure 21B+C). For CTCF, a high expression is associated with poor outcome and tumor aggressiveness, especially in ERG-fusion negative prostate cancer samples (Figure 21D) (Höflmayer *et al.*, 2019). For MITF and BHLHE40 no suitable antibody for TMA IHC-staining could be found. This showed that 4 out of the 6 identified *TERT* regulators from the MIPRIP-Net analysis were novel prognostic markers for prostate cancer and could help especially at the biopsy level to improve the prognosis progression and hence the therapy decision, e.g. if it is necessary to perform a radical prostatectomy. Especially a combination of several markers can lead to more reliable results. A Kaplan-Meier analysis with PITX1, CTCF, IRF1 and TFAP2D together showed that an upregulation of at least 3 of the 4 regulators is associated with a highly significant decreased PSA-recurrence free survival compared to patients for which less than 2 of these markers were expressed (Figure 21E). Hence, these 4 markers may act cooperatively as suggested by the MIPRIP analysis.

In summary, the modeling analysis identified new prognostic markers for prostate cancer.

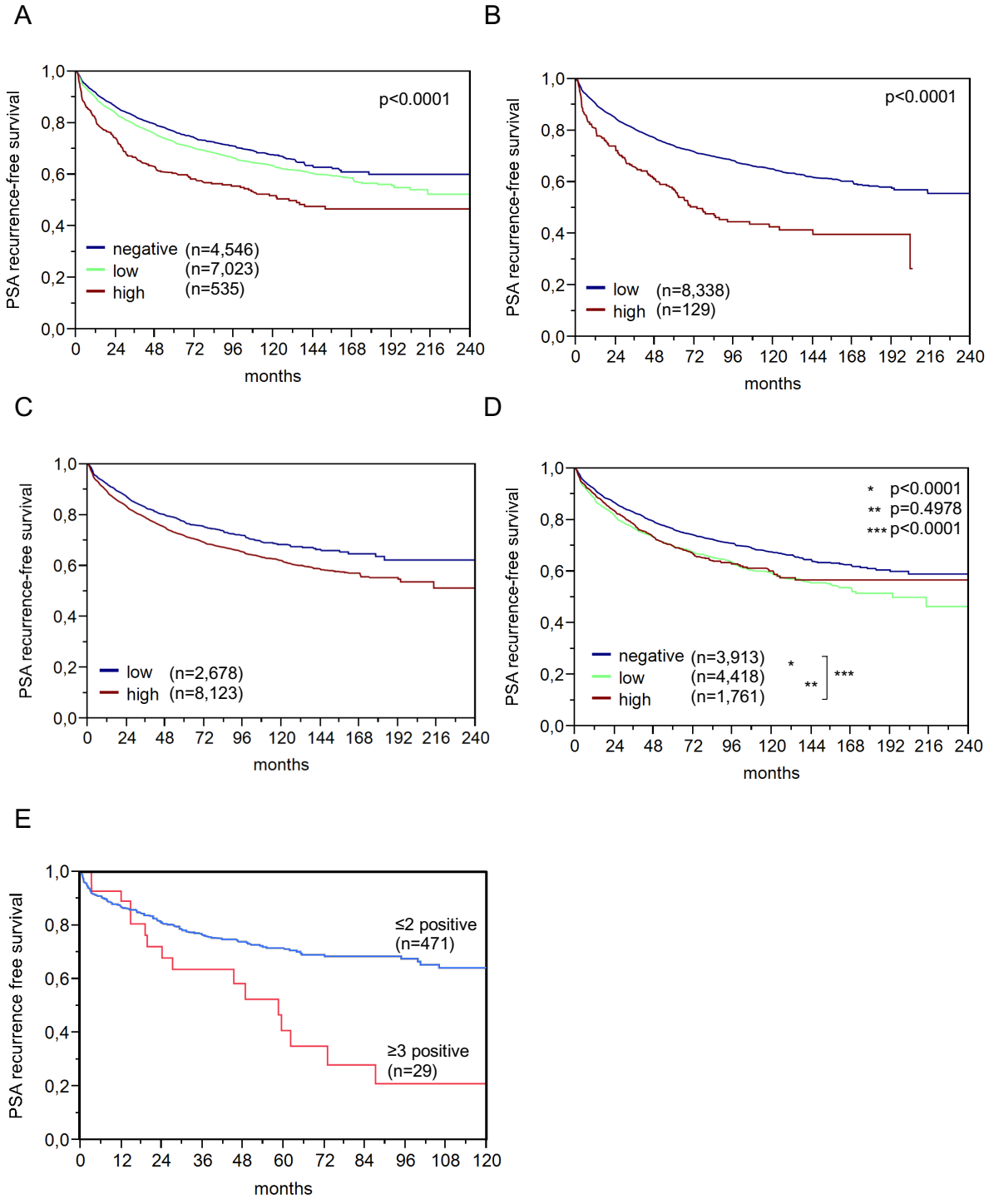

Figure 21. Kaplan-Meier analysis

of (A) PITX1, (B) IRF1, (C) TFAP2D and (D) CTCF over all patients. The PSA-recurrence free-survival is used instead of overall survival. (E) PSA-recurrence free survival of patients with a high expression of a combination of these 4 markers. (Plots were provided by the department of Guido Sauter (University Hospital Hamburg-Eppendorf).)

3.5 Telomere maintenance classification

PedGBM show a high frequency of ALT (44 %) (Heaphy *et al.*, 2011) and frequent mutations in ATRX as well as H3F3A (Schwartzentruber *et al.*, 2012). Therefore, pedGBM is an interesting entity to study the ALT mechanism.

3.5.1 Construction of a TMM classifier

The classification of patient's sample according to their active TMM is required to improve patient stratification. Therefore, I constructed a classifier based on typical TMM related features to distinguish between ALT positive (ALT) and ALT negative (non-ALT) samples. For this purpose, data from seven pedGBM derived cell lines with different genetic backgrounds were used together with 57 pedGBM patient samples from the PedBrain ICGC project (Bender *et al.*, 2013; Grasso *et al.*, 2015). All samples were characterized regarding their TMM features. In the cell lines the TMM status was determined by several functional assays (telomerase activity (TRAP-assay), telomere content (qPCR), C-circle levels, ATRX protein expression (IHC), heterogeneous telomere length (TRF blot), average number of ALT-associated PML nuclear bodies (APBs), TERRA levels and aberrant H3.3S31p as well as sequencing readouts (*TERT* expression from RNA-seq), mutational status of the genes *ATRX*, *H3F3A*, *TP53* and *TERT* promoter mutations as well as chromothripsis (derived by WGS) (Deeg *et al.*, 2017). A number of 5 cell lines showed a *H3F3A* mutation, 3 cell lines an ATRX mutation and one cell line had a *TERT* promoter mutation (C250T). From the cell lines two were characterized as non-ALT (SF188 and KNS42), while the others were ALT-positive. Not all the functional assays could be performed for the patient samples because of limited material and technical limitations. Hence, for these samples, data about TRAP-assays, TRF-blots, TERRA-levels and APBs was not available. The C-circle assay has been performed for 27 patients (by Inn Chung in the lab of Prof. Karsten Rippe, DKFZ, Heidelberg) and in addition the detection of ultra-bright telomere foci has been performed by telomere FISH for 20 patients (data provided by the lab of Prof. Pfister, DKFZ, Heidelberg). Sequencing readouts in form of RNA-seq (*TERT* expression), methylation data (*TERT* promoter methylation) and DNA-seq (mutations in the genes ATRX, H3F3A, TP53 and *TERT* promoter as well as chromothripsis) were available for almost all patients. The determination of the ALT

status was done with the results from the C-circle assay and ultra-bright telomere foci. A positive C-circle assay indicated an ALT-positive sample (Henson *et al.*, 2009), while a negative C-circle assay together with no ultra-bright telomere foci defined an ALT-negative sample. The combination of the C-circle assay and telomere FISH was performed because a lack of C-circle signal can also be observed from the rapid degradation of the single-stranded C-circles and might thus be interpreted as a false-negative result. Altogether, the results from the C-circle assay and the telomere FISH showed a high overlap in line with telomere FISH being a well-established marker for ALT-positive cells (Heaphy *et al.*, 2011). Samples with an activating *TERT* promoter mutation were classified as telomerase-positive. It has been reported that *TERT* promoter mutations and ALT occurrence were described as mutually exclusive (Killela *et al.*, 2013). In the present dataset this was also the case, since the samples with *TERT* promoter mutation for which a C-circle assay could be performed had a negative C-circle signal. Therefore, telomerase-positive samples were equated as ALT-negative samples. Altogether, based on these criteria 13 samples could be classified as ALT-positive and 12 as ALT-negative and two additional samples showed a *TERT* promoter mutation (also ALT-negative). A high fraction of the ALT-positive samples showed a high correlation for a loss of function mutation in the gene *ATRX* (11 out of 13 ALT-positive samples) and *TP53* (12 out of 13 ALT samples), a higher prevalence of *H3F3A* mutations (4 out of the 6 *H3F3A* mutated samples are ALT positive), a lower *TERT* expression (Figure 23A), lower *TERT* promoter methylation (Figure 23B) and a higher telomere content (Figure 23C). For chromothripsis there was no correlation with the occurrence of ALT.

These 27 pedGBM samples together with the 7 cell lines were used as training set for the decision tree-based classifier (Table S24). Here, I systematically tested all combinations of TMM features to identify feature combinations which were able to distinguish between ALT-positive and ALT-negative samples. Because of limited patient's material the datasets are frequently incomplete. However, the TMM prediction potential of single available features is required e.g. for treatment decisions. It is noted that the classifier can distinguish only between ALT-positive and ALT-negative samples, with the latter including samples with the ever shorter telomere (*EST*) phenotype (Dagg *et al.*, 2017).

The *TERT* methylation and the telomere content information was different for the 7 cell lines compared to the patient samples. The *TERT* methylation was much higher in the cell lines than in the patient samples. Regarding the telomere content, a normalization to healthy control samples was not possible in the cell lines. Nevertheless, all other features were in high agreement with the patient samples. The C-circle and the *TERT* promoter mutation were not used for the training of the classifier, in turn, these two features were used to define the classes “ALT” and “non-ALT” as shown in Figure 22.

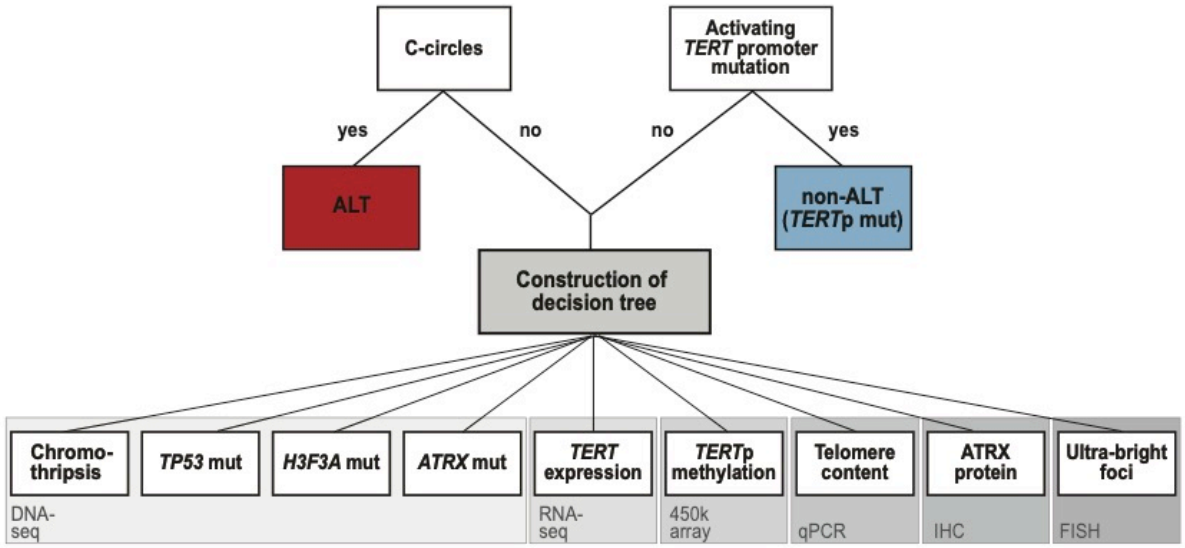

Figure 22. Scheme for classification of primary pedGBM samples according to their TMM status.

A positive result in the C-circle assay determines the class “ALT”, while an activating TERT promoter mutation shows that this sample is non-ALT. If none of these two criteria are true, the sample is classified based on the available features. Image taken from (Deeg et al., 2017).

Most of the described features were binary variables, e.g. mutated or wild-type. In general, the presence of a feature was set to 1 (e.g. mutation present or chromothripsis observed), the absence to 0. In contrast, the features *TERT* expression, *TERT* promoter methylation and telomere content were continuous variables. For these variables a threshold was defined for the training sample set by minimizing the number of misclassified samples when testing different thresholds. The thresholds for the three continuous variables are shown in Figure 23. The *TERT* expression in the RNA-seq data was very low, but still higher in the non-ALT samples indicating that telomerase is active in these samples. The threshold was RPKM > 0.01 for the non-ALT samples in the pedGBM patients and cell lines (Figure

23A+D). As described in (Castelo-Branco *et al.*, 2013) methylation at position cg11625005 of the *TERT* transcription start site is a marker for *TERT* expression. The ALT-negative samples showed a higher *TERT* promoter methylation compared to the ALT-positive samples. This shows that telomerase expression is correlated to its promoter DNA-methylation. The threshold here was determined to be 0.22 for the patient samples (Figure 23B) and 0.80 for the cell lines (Figure 23E). As in the pedGBM cell lines, the telomere content of the pedGBM patient samples was higher in ALT-positive compared to ALT-negative samples. This could only be measured for the samples for which tumor and matched control blood samples were available. The threshold for the ratio was 0 (Figure 23C). For the cell lines the threshold for the telomere content was 0.10 (Figure 23F).

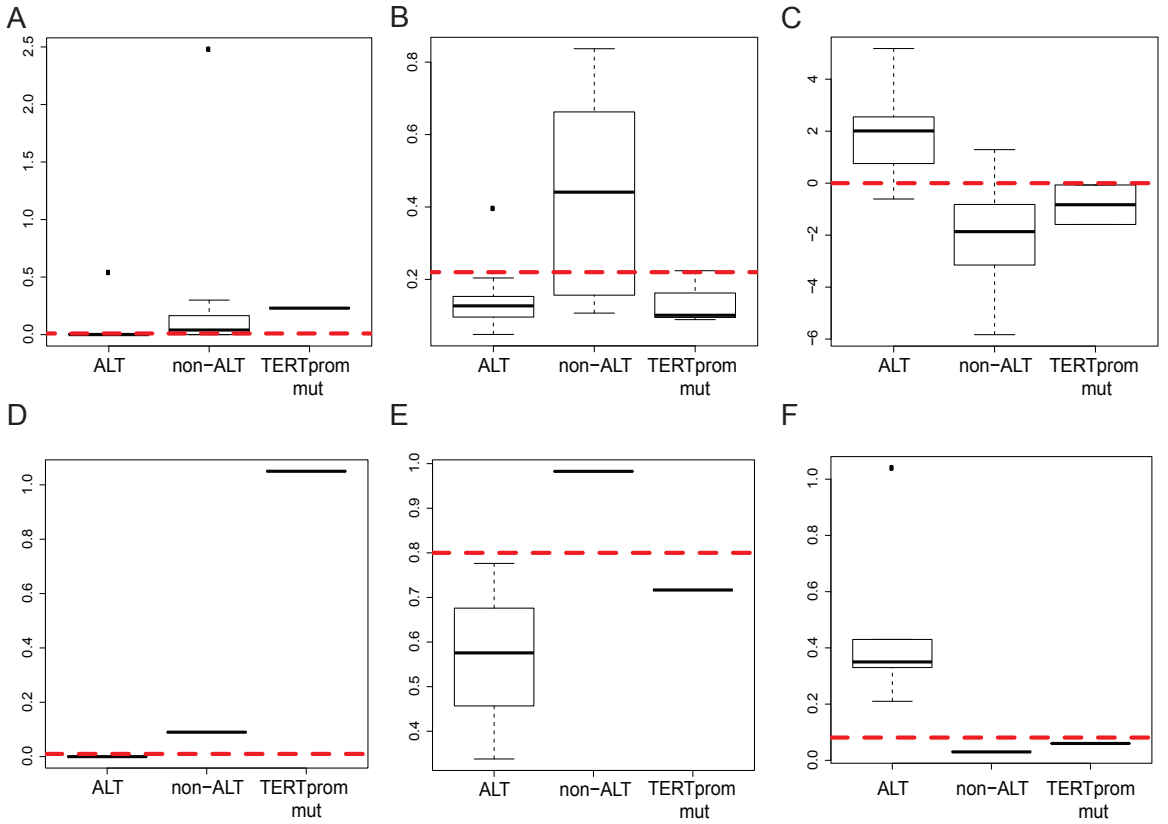

Figure 23. Thresholds of the continuous features.
(A) *TERT* expression in pedGBMs; (B) *TERT* promoter methylation in pedGBMs; (C) Telomere content in pedGBMs; (D) *TERT* expression in the pedGBM cell lines; (E) *TERT* promoter methylation in the pedGBM cell lines; (F) Telomere content in the pedGBM cell lines for the ALT, non-ALT and *TERT* promoter mutated patient samples/cell lines. The red dashed line indicates the threshold which was used to construct the decision trees.

For all possible feature combinations decision trees were constructed. Because of some missing feature information, only training samples with all available features

were used leading to trees calculated with different sample sizes. The accuracy *Acc* of the classifier was determined by the number of correct predicted samples based on a leave-one-out cross-validation with the training samples. Additionally, based on a confusion matrix a *p*-value (*p*) was calculated for each tree to test if this tree is better than a random tree. If a selected feature set led to a poor prediction performance the best performing subset was used. The reason for this can be e.g. a difference in the sample sizes for one of the features.

Constructing a tree with only one feature, either presence of ultra-bright foci (*Acc*=0.90), ATRX protein expression and or telomere content (each *Acc*=0.86) as well as ATRX mutation (*Acc*=0.85), could best predict the TMM status (ALT or non-ALT) in the pedGBM samples, while chromothripsis or H3F3A mutation alone led to a non-reliable prediction (each *Acc* = 0.59, $p > 0.05$). For ultra-bright telomere foci and telomere content the sample size was very low with 20 and 22 samples, respectively, but the predictions were still accurate (Table 12).

Table 12. Performance and significance level if the 9 TMM features were used alone for the classification into ALT and non-ALT.

Feature	# samples	Accuracy	p-value
Ultra-bright telomere foci	20	0.90	7.22 E-04
ATRX IHC	28	0.86	1.53 E-04
Telomere content	22	0.86	1.91 E-03
ATRX mutation	34	0.85	2.80 E-05
TERT expression	32	0.78	1.67 E-03
TERT promoter methylation (cg11625005)	32	0.78	2.05 E-03
TP53 mutation	34	0.71	1.45 E-02
H3F3A mutation	34	0.59	n.s.
Chromothripsis	34	0.59	n.s.

n.s. = not significant

Combination of features could improve the prediction. For example, a combination of the features *TP53* mutation (alone *Acc* =0.71) and *TERT* expression (alone *Acc* =0.78) and chromothripsis (alone *Acc*=0.59) showed a highly increased prediction performance (*Acc* = 0.89, *p* = 6.88 E-05, # samples = 28) (Figure 24A). ATRX protein (IHC) instead of chromothripsis led to the same performance (Figure 24B). This is surprising because ATRX protein expression was the feature with the second-best performance, while chromothripsis alone was the least accurate. This

showed that the combination of different features into one decision tree was powerful. It is noted that all 4 features from Figure 24 A+B together (*TERT* expression, *TP53* mutation, ATRX protein and chromothripsis) did not improve the performance, likewise an addition of H3F3A mutation did neither. An example for a tree with three layers was the combination of *TERT* promoter methylation with chromothripsis and *TERT* expression (Figure 24C). This combination led to a performance of *Acc*=0.8 (p = 0.0025, # samples = 30) which was a slightly higher than *TERT* promoter methylation alone (*Acc*=0.78).

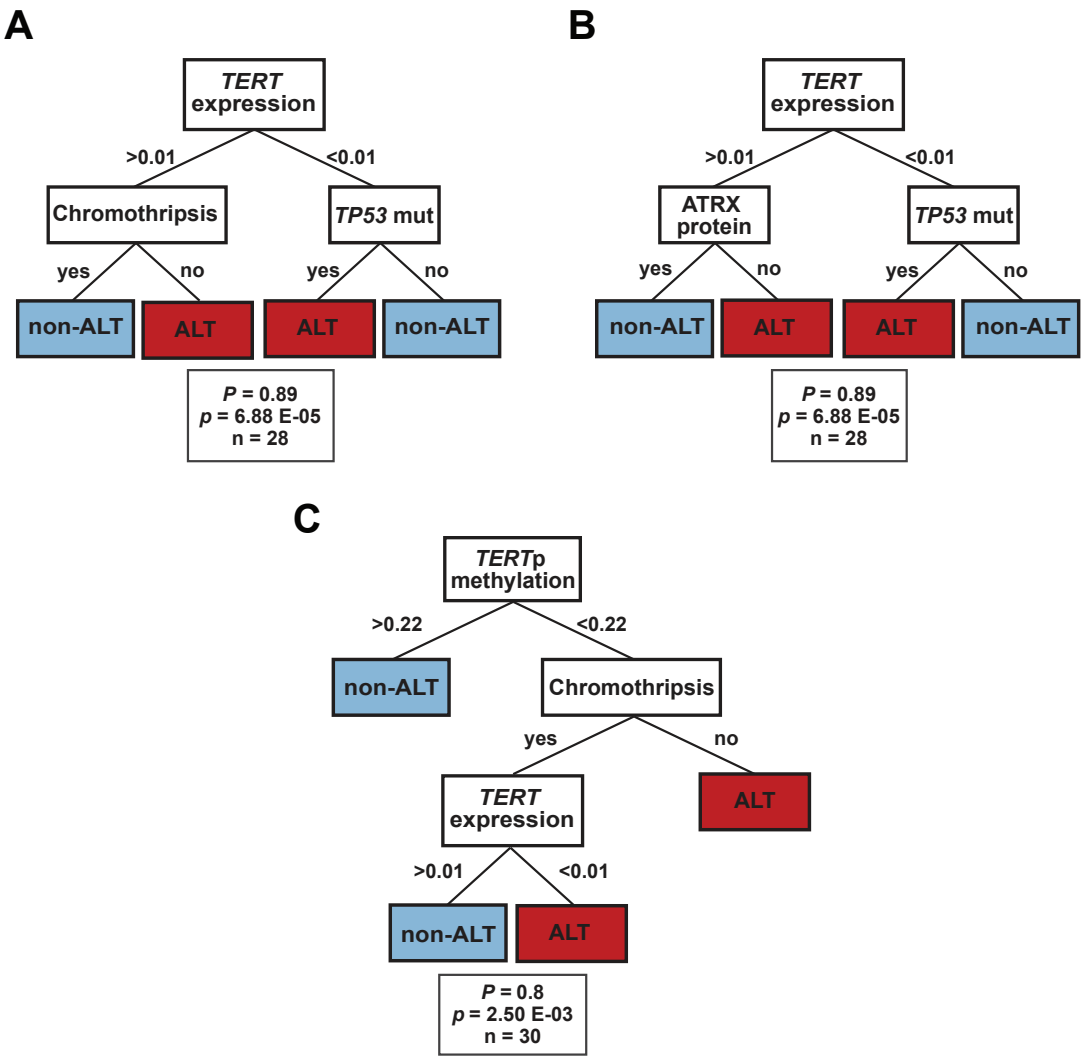

Figure 24. Three feature combinations leading to an improved performance.
(A) The feature combination of *TERT* expression, chromothripsis and *TP53* mutation led to a highly improved performance (*Acc*=0.89) compared to the features alone. (B) The tree with the feature combination *TERT* expression, ATRX protein detection and *TP53* mutation shows the same performance then the tree in (A). (C) The combination *TERT* promoter methylation, chromothripsis and *TERT* expression can also improve the performance to some extent.

The classifier was implemented as a web-based tool called 'Predicting Alt IN Tumors' (PAINT) (http://www.cancertelsys.org/paint/index.html) by Nick Kepper from the group of Prof. Rippe (DKFZ, Heidelberg). The user can select all available TMM features of his/her pedGBM sample and receives a probability (performance *P* and *p*-value) with the sample being ALT or non-ALT. If the *p*-value of the selected feature combination is not significant, PAINT displays the feature subset with the best performance and its *p*-value. Next, PAINT was used to predict the TMM status (ALT or non-ALT) of the remaining 30 pedGBM patient samples. This resulted in 12 ALT and 18 non-ALT samples (Table S25).

In summary, features like detection of ultra-bright foci or ATRX mutation status alone led to a high performance, while the combination of TP53 mutation status, chromothripsis and *TERT* promoter methylation could greatly improve the prediction. This in general showed that a combination of sequencing-based readouts led to a highly reliable prediction. Sequencing based readouts have been shown to be well suited for the clinical routine without the need of additional assays.

3.5.2 ALT gene signature in pedGBM patient samples

Next, a differential analysis was performed based on (i) gene expression level and (ii) regulator activities, to investigate differences in gene expression levels between ALT-positive and ALT-negative samples. For this purpose, all patient samples with available RNA-seq expression data were used, resulting in 34 patient samples. The TMM status (ALT, non-ALT) of the 34 samples was determined from the training set or was predicted with PAINT as described above. This resulted in 14 ALT-positive and 20 ALT-negative samples. Because the difference between cell lines and patient samples was too high, all cell lines were excluded from the analysis (Figure S2). First, a differential gene expression analysis was performed, which resulted in 115 differentially expressed genes (p-value<0.01; 366 genes for p-value<0.05). Clustering of differentially expressed genes (p-value<0.01) did not lead to a clear separation of ALT-positive and ALT-negative samples. There were few genes (e.g. DRG2, PTCHD4 and CPAMD8) which were higher expressed in the ALT-negative samples, while most genes had a low expression in both groups. From the 115 differentially expressed genes (p-value < 0.01) only 5 could be found in the TelNet database (Braun *et al.*, 2018) (13 out of the 366 differentially expressed genes with p-value < 0.05). These 5 genes were TERT, PCK1, HIST1H1A,

CCDC155 and ABCC12. Besides *TERT*, only for PCK1, repressor of telomerase (Cerone *et al.*, 2011), a direct relation to telomere maintenance has been reported. The differentially expressed genes with a *p*-value < 0.05 were used to calculate the activities of the regulators with the MIPRIP framework. A significant change (*p*-value<0.01) in their activity between the ALT-positive and the ALT-negative samples was observed for 15 regulators. The activity values of the 15 regulators led to two clearly separated clusters. One cluster with upregulated regulators and the other with downregulated regulators.

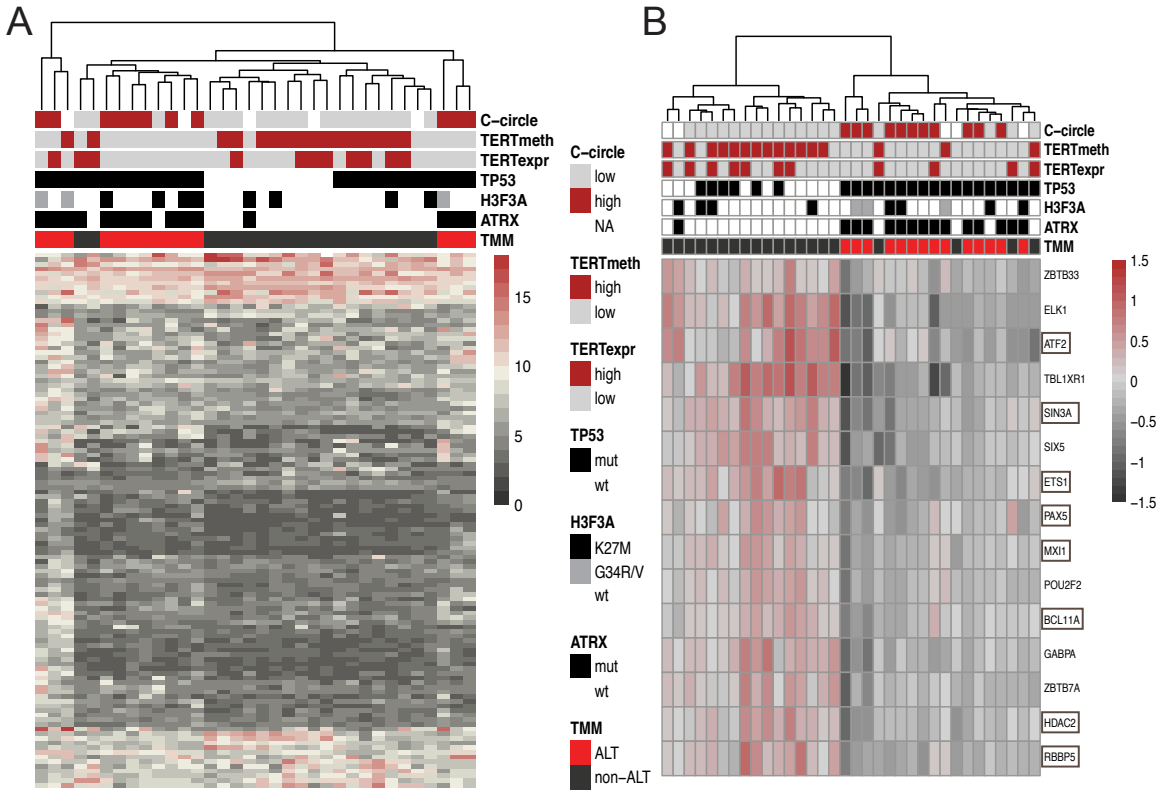

Figure 25. TMM gene signature of pedGBM.

(A) Clustering of pedGBM patient samples into ALT and non-ALT based on differential gene expression (*p*-value < 0.01) for which RNA-seq data was available. For this analysis the cell lines were excluded. (B) Clustering of pedGBM patient samples into ALT and non-ALT based on the activities of the regulators showed a significant change in the ALT compared to the non-ALT samples (*p*-value < 0.01). For this purpose, the regulator activities were calculated as described for MIPRIP by using only the differentially expressed target genes (*p*-value < 0.05) of the regulators. The analysis was restricted to regulators with at least 5 differentially expressed target genes. The boxes indicate that the regulator was found in the TelNet database (Braun *et al.*, 2018).

The cluster with upregulated regulators only contains ALT-negative samples, while the downregulated regulator cluster includes all ALT-positive samples plus three

outlier (ALT-negative samples). In the TelNet database, 8 out of the 15 regulators could be identified as telomere maintenance genes. These are *ATF2, SIN3A, ETS1, PAX5, MXI1, BCL11A, HDAC2* and *RBBP5*. The six regulators SIN3A, ETS1, PAX5, MXI1, POU2F2 and BCL11A are putative regulators of *TERT* based on the generic human regulatory network. These 6 out of 15 regulators were upregulated in almost all ALT-negative samples indicating a possible activating role of *TERT* expression in these samples. *PAX5* and *MXI1* were identified in the pan-cancer MIPRIP analysis as a common *TERT* regulator, while for melanoma samples with a *TERT* promoter mutations *ETS1* is a highly significant hit. *POU2F2* was part of the gene regulatory network of *TERT* in prostate cancer.

In summary, the calculation of regulator activities was better suited than differential gene expression analysis to identify a signature that can distinguish between ALT and non-ALT pedGBM samples.

3.6 Comparison of MIPRIP with ARACNE/VIPER based on chronic lymphocytic leukemia

ARACNE and VIPER are well-established tools from the Califano Lab to study gene regulation. ARACNE constructs a de-novo GRN based on gene expression data and VIPER uses these networks to calculate regulator activities.

The comparison of MIPRIP with ARACNE/VIPER could not be performed based on the regulation of *TERT.* Because of the low *TERT* expression it was not possible to identify reliable regulators of *TERT* in the ARACNE network. Instead, I used gene expression data of 20 CLL patients and 7 non-malignant B-cell samples for the comparison of the both approaches. For the comparison the following two questions were addressed:

(i) How different is the generic human regulatory network compared to the B-cell specific ARACNE network?

(ii) How is the overlap between the regulator activities calculated with MIPRIP and VIPER?

For the latter case the overlap of significantly differential active regulators between CLL and non-malignant B-cell samples from MIPRIP and VIPER was estimated. First, I computed a B-cell specific ARACNE network with published microarray data of 264 malignant and non-malignant B-cell samples (Basso *et al.*, 2010). In total,

214,405 interactions could be identified for 3,862 regulators and 12,119 target genes. The regulators include transcription factors, transcriptional co-factors and also signaling pathway related genes (Alvarez *et al.*, 2016), while the generic human regulatory network used for MIPRIP is limited to TFs.

To compare the generic human regulatory network with the ARACNE B-cell network the number of target genes was compared for both networks. Strikingly in the generic human regulatory network there were several regulators with a much higher number of target genes than in the ARACNE B-cell specific network. For the generic regulatory network the average was 533 target genes (maximum = 16,483 for CTCF), but three-quarter of the regulators (872 out of the 1,160) had less than 500 target genes (Figure 12). In the B-cell specific ARACNE network all regulators had less than 500 target genes (mean ≈ 56 target genes) (Figure 26). This showed that the generic human regulatory network contained only some master regulators like CTCF, YY1 or MYC with huge numbers of target genes, while most regulators in both networks showed a comparable number of target genes.

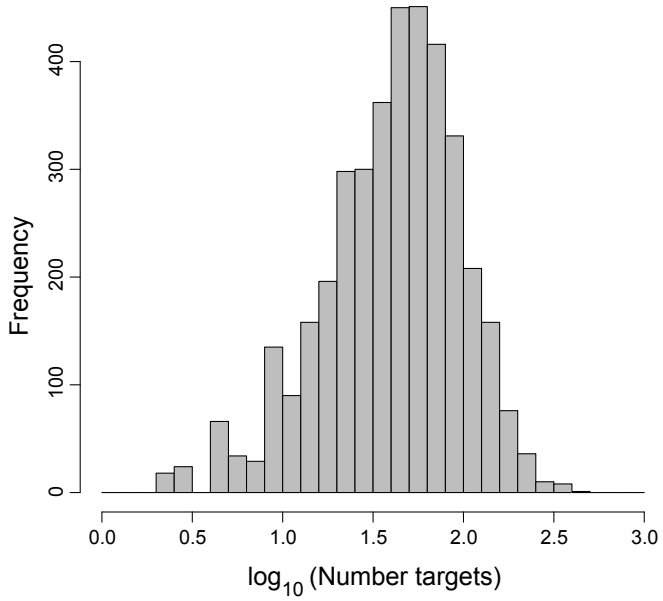

Figure 26. Histogram of the number of targets for all 3,885 regulators in the ARACNE B-cell network.

The ARACNE B-cell specific network was then used together with the in-house RNA-seq data to calculate the activity values of the regulators using the VIPER R-package (Alvarez *et al.*, 2016). Altogether, regulator activity values could be calculated for 2,804 regulators and 1,588 regulators (p-value ≤ 0.05) showed a significant change in their activity between the CLL and the non-malignant B-cell

samples. To evaluate the tissue-specificity of the constructed B-cell network, a pathway analysis was performed with the regulators showing the highest activity values in the non-malignant B-cell samples based on the VIPER analysis and the whole network as universe. This yielded the pathway “B cell receptor signaling pathway” as top hit, followed by “Pathways in Cancer” and “Thyroid hormone signaling pathway” (Table 13).

Table 13. Top 5 significant KEGG pathways of the regulators with the highest activity in the non-malignant B-cell samples.
This was calculated with the B-cell specific network, an ARACNE AML, GBM and PRAD network. The position of the B cell receptor signaling pathway is also shown for each network.

Network	*Position*	*KEGG pathway*	*p-value*
B-cell	1	B cell receptor signaling pathway	2.55E-09
	2	Pathways in cancer	3.37E-09
	3	Thyroid hormone signaling pathway	1.70E-08
	4	Hepatitis B	1.70E-08
	5	Cellular senescence	9.14E-08
AML	1	Pathways in cancer	1.49E-11
	2	Kaposi’s sarcoma-associated herpesvirus infection	1.13E-10
	3	Chemokine signaling pathway	2.32E-10
	4	Thyroid hormone signaling pathway	2.90E-10
	5	Autophagy – animal	9.77E-10
	6	B cell receptor signaling pathway	1.18E-09
GBM	1	Thyroid hormone signaling pathway	2.00 E-07
	2	Transcriptional misregulation in cancer	4.19 E-07
	3	Pathways in cancer	6.18 E-07
	4	Herpes simplex infection	6.18 E-07
	6	Viral carcinogenesis	7.74 E-07
	27	B cell receptor signaling pathway	2.52 E-04
PRAD	1	Basal transcription factors	8.55E-10
	2	Hepatocellular carcinoma	4.68E-08
	3	Herpes simplex infection	5.64E-08
	4	Colorectal cancer	1.00E-07
	5	Neurotrophin signaling pathway	1.30E-07
	65	B cell receptor signaling pathway	4.74 E-03

The pathway “B cell receptor signaling” as top hit indicates that the ARACNE network constructed from B-cell expression data is highly tissue-specific.

Furthermore, I calculated the regulator activities for the CLL and the non-malignant B-cell samples with a publicly available ARACNE network (R-package 'aracne.networks') for acute myeloid leukemia (AML), for glioblastoma (GBM) and prostate cancer (PRAD). For each of the three networks, the regulators with the highest activity in the B-cell samples were used for the pathway analysis. For the AML, the GBM and the PRAD network, the 5 most significant pathways are shown in Table 13. The "B-cell regulator signaling pathway" was ranked on position 6 for the AML network, on position 27 for the GBM network and on position 65 for the PRAD network. This shows that the network constructed with the microarray B-cell gene expression data is highly tissue-specific and suited best to study the deregulated regulatory processes between CLL and non-malignant B-cell samples. As comparison I calculated regulator activities in the manner of MIPRIP for the same CLL and non-malignant B-cell samples by determining the target genes from the generic human regulatory network. With the MIPRIP activity calculation, I identified 763 out of 1,154 regulators significantly changing their activity between the CLL and the non-malignant B-cell samples. To note, compared to ARACNE/VIPER the MIPRIP activity calculation is restricted to TFs. For the comparison I limited the list to the regulators for which activity values could be computed with both methods. This resulted in an overlap of 461 regulators (Figure 27A), from which 300 showed a significantly different activity between the CLL and non-malignant B-cell samples calculated with MIPRIP and for 254 calculated with VIPER. The overlap between the regulators with significant activity changes was 159 (Figure 27B). With the 159 regulators showing a significant change in their activity between the CLL and the non-malignant B-cell samples a pathway analysis was performed. Pathways like "Transcriptional misregulation in cancer", "Th17 cell differentiation" and "TCF-beta signaling pathway" were identified among others as highly enriched (Figure 27C). Furthermore, looking at the differential expressed target genes of the 159 regulators in both networks a high overlap could be identified. The 159 regulators had 2,660 differential expressed target genes in the generic network and 964 differential expressed target genes in the ARACNE B-cell specific network. But nearly all differentially expressed target genes (89 %) from the B-cell network were also target genes in the generic network (Figure 27D). This comparison again showed that the regulators in the generic regulatory network had more putative target genes than in the ARACNE network, but also that the target genes are highly overlapping.

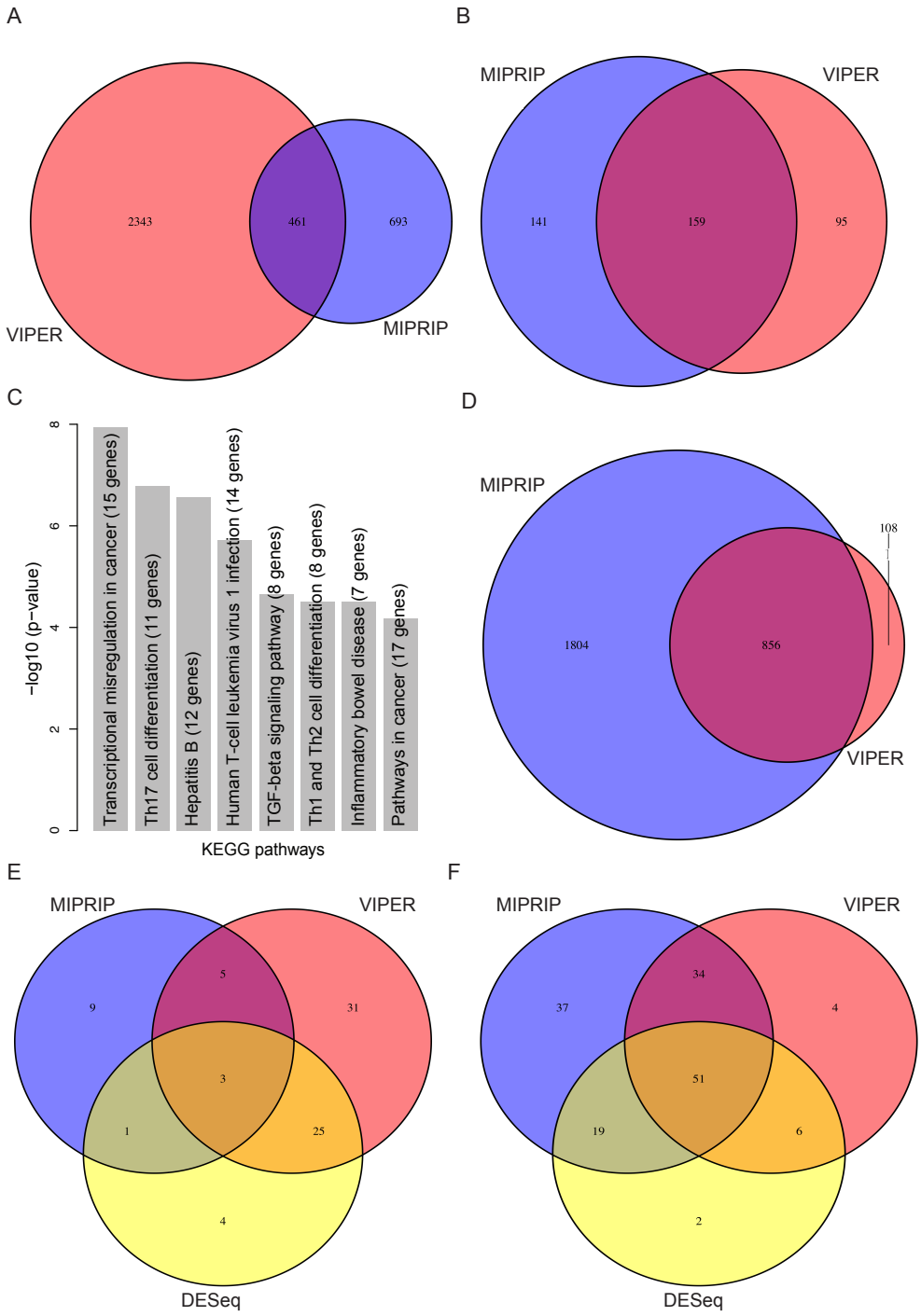

Figure 27. Comparison between the activity calculation with MIPRIP and VIPER.

(A) Venn diagram showing the overlap of regulators for which activity values could be calculated. (B) Venn Diagram showing the number of regulators with a significantly higher activity in the CLL samples compared to the non-malignant B-cell samples calculated with MIPRIP and VIPER. (C) Top ranked KEGG pathways of the 159 regulators identified as significant with MIPRIP and VIPER. (D) Overlap of the target genes of the 159 significantly differential active regulators identified with MIPRIP and VIPER. (E) Significantly up- and (F) downregulated regulators based on the activity calculation with MIPRIP and VIPER compared to differential gene expression analysis with DESeq2. This comparison was performed for the 159 regulators with a significant activity change between CLL and non-malignant B-cell samples with VIPER or MIPRIP.

Comparing the regulators with significantly higher (Figure 27E) and lower (Figure 27F) activity values in CLL *vs.* non-malignant B-cell samples with the results from the differential gene expression analysis (performed with DESeq2), most regulators were downregulated in CLL (MIPRIP=141 genes, VIPER=95 genes and DESeq2=78 genes). Especially in MIPRIP, the number of regulators with a higher activity in the CLL samples was low (n=18 genes). The overlap between the activities calculated within the MIPRIP framework and VIPER was higher than between the significantly differential expressed regulators from the DESeq analysis and the regulators with significant activity change in MIPRIP/VIPER, especially for the regulators with lower activity in the CLL samples.

In summary, there was a high overlap between the activities calculated with MIPRIP and VIPER, although the number of target genes were very different between the generic human regulatory network and the B-cell specific network computed with ARACNE. Around one third of the regulators for which activity values could be computed with both approaches showed a significantly change in their activities between the CLL and the non-malignant B-cell samples.

4 Discussion

In this thesis, I developed a new approach based on Mixed Integer Linear Programming combined with machine learning methods to predict significant regulators of any given gene. Using this approach to study telomere maintenance, novel regulators of the telomerase in *S. cerevisiae* and different human cancer types could be identified. Furthermore, based on a classification scheme pedGBM patients were grouped into ALT and non-ALT and a regulator signature to distinguish between both groups was identified. Here, I will discuss the advantages and disadvantages of the new approach as well as the clinical relevance of the novel insights on telomere maintenance.

4.1 Mixed Integer linear Programming based Regulatory Interaction Predictor

The 'Mixed Integer linear Programming based Regulatory Interaction Predictor' (MIPRIP) presents a novel approach of linear regulation models based on Mixed Integer Linear Programming (MILP) embedded into machine learning methods to identify the most important regulators of a gene. It uses the advantage of L1 norm for regression avoiding overestimating outliers and of the implementation of constraints to get sparse models. The basic idea of MIPRIP is to identify the most relevant regulators of a target gene by predicting the target gene's expression using a linear model in which the covariates are all potential regulators putatively binding to its promoter. These potential regulators were extracted from a generic regulatory network which has been constructed in the lab of Prof. Dr. König (University Hospital Jena) for *S. cerevisiae*, human and mouse. The TF-target gene interactions are mainly based on ChIP-binding data, but results from literature research and computational binding predictions were also considered. For yeast the generic regulatory network is binary. This means that if there is an interaction then the edge weight is 1 and 0 otherwise. For human and mouse, the TF-target gene interactions were extracted from several databases and integrated into one generic network. In both networks, the edge weights were selected and weighted based on the reliability of the database and co-occurrences. One possible improvement of the generic regulatory network could be to optimize the edge weights from the different sources.

For instance, the experimentally validated direct interactions from MetaCore™ could be used as a gold standard. The weighting of the interactions from the other sources could be based on the overlap with the gold standard. For more than half of the regulators less than 25 target genes were identified, while a small subset had more than 10,000 target genes (e.g. MYC, CTCF or YY1). These regulators are so-called master regulators. This is in agreement with the finding that regulatory interactions between TFs and their target genes organize as a scale-free network with hubs as master regulators (Babu *et al.*, 2004). It is to be noted that the number of target genes is highly dependent on how well different regulators have been studied. The computed generic networks were used to identify the putative regulators of a particular gene and to calculate the activity of each regulator. Similar to Balwierz *et al.* (Balwierz *et al.*, 2014) or Alvarez *et al.* (Alvarez *et al.*, 2016) the activity of a regulator is defined as the cumulative effect on its target genes. As previously shown using the activity value of a regulator instead of its gene expression value led to a better prediction of the gene expression of the gene of interest (Schacht *et al.*, 2014). A reason for this could be that regulators can act cooperatively with other TFs or signaling molecules and are modulated by post-transcriptional modifications or protein stability, which makes the activity values more informative than the gene expression values.

A recent study by Trescher *et al*. (Trescher *et al.*, 2017) compared a previous version of MIPRIP (Schacht *et al.*, 2014) with four other methods. All these methods determine TF activity from gene expression data using pre-defined TF-target gene interactions as the basis for linear regression or probabilistic models. The overlap between the identified regulators with the different tools was very low, although the approaches are methodological quite similar. One of these tools (RACER) (Li *et al.*, 2014) integrates data of mRNA expression, miRNA expression data, copy number variations and DNA methylation. Li *et al.* compared the results from RACER using different combinations of input variables in form of mRNA and miRNA expression data, copy number variations and DNA methylation data. If they exclude TF regulation from their model the performance was reduced considerably. This suggests that TFs seem to be most important for regulation. Still, prediction of the regulatory mechanism for the gene of interest may be improved by incorporating miRNA expression data, copy number variations and DNA methylation. Indeed, the MIPRIP implementation allows to incorporate also additional information extracted

from DNA methylation, miRNA expression and binding data, gene copy numbers, as well as other epigenetic regulation. This can be a promising future project.

As transcriptional processes are highly tissue-specific and within one tissue also condition specific, I extended MIPRIP with a downstream statistical hypothesis testing framework to compare the predicted regulators between two or more datasets/conditions (MIPRIP-Comp). MIPRIP-Comp allows to identify significant regulators between two datasets/conditions (e.g. treatment *vs.* control) (dual-mode) as well as between multiple datasets/conditions (e.g. pan-cancer analysis) (multi-mode). A multi-mode MIPRIP-Comp analysis predicts (i) the most common regulators of a particular gene over all datasets/conditions and (ii) the specific regulators of each dataset/condition compared to all other datasets/conditions (Poos *et al.*, 2019). The application of MIPRIP on telomerase regulation showed that MIPRIP performed well and led to novel regulators (see below).

Furthermore, MIPRIP was combined with a modularity-based approach (MIPRIP-Network (MIPRIP-Net)) because typically several co-regulators are involved in the expression of a gene. These co-regulators interact with the regulators binding to the gene's promoter and influence their activity. Modularity was introduced by Newman about ten years ago to cluster gene networks (Newman, 2006). In combination with MIPRIP, it was used to identify the subnetwork, which (i) can best predict the gene expression of the particular gene and (ii) integrates highly connected regulators that influence the expression of the particular gene directly by binding to its promoter or indirectly by interacting with other regulators. Both objectives can be easily combined by a tradeoff parameter, which shows the power of MILP compared to other approaches. Solving such an optimization problem with more than one objective function is called Pareto-optimization. Here, both objectives are understood as separate functions and are integrated by a weighting factor into one combined objective. This led to a reduction of the objectives and an optimal solution of the combined objective function is determined. To avoid an arbitrary weighting factor a separate optimization of several weighting factor combinations is necessary. Typically, there is more than one clear solution. Therefore, the set of solutions is called Pareto-optimum. The Pareto optimum restricts the search space to find the best compromise between all objectives, so that an improvement of one objective does not lead to a deterioration of the other. For the MIPRIP-Net approach, models of 9 different weighting factors were constructed. The optimal weighting

factor was evaluated based on the number of selected regulators per objective (MIPRIP and modularity). The performance of the MIPRIP-Net models was estimated from MIPRIP, but the combination with the modularity influences the regulators which were selected by the MIPRIP model. Plotting the number of selected regulators separately for both objectives, the intercept of both curves defines the optimal weighting factor. At this point the number of regulators selected by the basic MIPRIP approach and by the new modularity-based approach were balanced and the prediction performance was still comparable to MIPRIP alone. The final network models were then computed with the optimized weighting factor and the combination of direct regulators, which was used in most of the MIPRIP models over all cross-validation runs, was selected. For this combination the highly connected indirect regulators were identified. The optimization of the weighting factor has to be performed once per dataset. The current implementation of MIPRIP-Net is limited to the “best” subnetwork, but it can be extended for the identification of several clusters/modules.

All MIPRIP variations can be easily applied to gene expression data to identify the most important regulators of a particular gene in or between different conditions. So far, MIPRIP-Comp together with the basic MIPRIP version (single-mode) is implemented in the R-package ‘MIPRIP2’. ‘MIPRIP2’ together with the generic regulatory networks, a user’s guide and some example data is publicly available at https://github.com/network-modeling/MIPRIP and http://www.leibniz-hki.de/en/miprip.html.

In this thesis, MIPRIP was applied to study the regulation of the telomerase in *S. cerevisiae* as well as different types of human cancers.

4.2 Telomerase regulation in *S. cerevisiae*

Telomere length maintenance is a precisely controlled process in all eukaryotic cells. Its activation is highly important during embryogenesis as well as for cancer cells to enable their replicative immortality. In contrast to somatic cells, stem cells, most of the cancer cells as well as *S. cerevisiae* express the telomerase. As the telomeres are highly conserved between eukaryotes and around 23 % of all yeast genes have human homologs, *S. cerevisiae* is a well suited model system to study

telomere maintenance (Teixeira, 2013). Therefore, I studied the transcriptional regulation of the telomerase in *S. cerevisiae*. In yeast, the telomerase consists of the ever shorter telomere (*EST*) 1-3 genes and the template RNA TLC1. Around 500 *TLM* genes have been identified that showed shorter or longer telomeres compared to the wild-type when mutated (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Shachar *et al.*, 2008; Ungar *et al.*, 2009). *TLM* genes leading to shorter telomeres after deletion are positive regulators of telomere maintenance and their absence may have a direct influence on the expression of the telomerase as well as their activity. Hence, MIPRIP was applied to gene expression data of yeast deletion strains to identify the most important regulators of each *EST* gene by comparing the regulatory processes between yeast deletion strains with short telomeres compared to deletion strains with normal (wild-type) telomere length. The putative regulators of the three *EST* genes were extracted from the generic yeast regulatory network and were mainly based on ChIP-experiments. The MIPRIP analysis of the *EST* genes resulted in 32 regulators which were significantly more often selected in the models of the samples with short telomeres compared to the samples with normal telomere length (controls). For *EST1*, I identified Sum1, Hst1 and Srb2 as the most significant regulators of the short *tlm* deletion strains compared to the controls. Sum1 is a chromatin silencing factor which can build a complex together with Hst1 and Rfm1 (Bedalov *et al.*, 2003; Li *et al.*, 2013; McCord *et al.*, 2003; Zill and Rine, 2008). This complex deacetylases the histones at the promoters and therefore represses the expression of the genes. The MIPRIP results led to the assumption that a complex of Sum1 and Hst1 is involved in *EST1* regulation and maybe also Rfm1 is indirectly involved as it can be part of the complex (McCord *et al.*, 2003). Furthermore, it was shown, that Sum1 can similarly to the sirtuins Sir2 and Hst1 interact with Rap1, indicating that Sum1 is involved in telomere maintenance (Li *et al.*, 2013). However, until now no direct influence of Sum1 or Hst1 on the expression of the *EST* genes has been reported. A deletion of *SUM1*, *HST1* or *SRB2* led to yeast strains with shorter telomeres and the expression of *EST1* was highly upregulated in these deletion strains. This upregulation of *EST1* in response to *SUM1* deletion could be also confirmed by quantitative RT-PCR from our collaboration partner Andre Maicher from the group of Prof. Luke (IMB, Mainz). Hence, Sum1 is a negative regulator of *EST1*. This finding is surprising as the *sum1* deletion strain showed shorter telomeres. As *EST1*

is upregulated in the samples with shorter telomeres compared to samples with wild-type telomere length, there could be some imbalances between the telomerase subunits which limit the activity of the telomerase. Our results suggest that Sum1 and Hst1 are further involved in telomere maintenance besides their interaction with Rap1. If these regulators are not directly involved in the transcription of the *EST* genes, they can also act as chromatin remodellers. As shown for other networks (Fu *et al.*, 2012; Mangan and Alon, 2003; Mangan *et al.*, 2006; Mangan *et al.*, 2003), Sum1 together with Hst1 and *EST1* could form an incoherent feed-forward loop to regulate telomere length maintenance. In this incoherent loop, either Sum1/Hst1 or *EST1* positively regulate telomere length, while *EST1* is negatively regulated by the Sum1/Hst1 complex (Figure 28). Negative feedback mechanisms play a self-regulating role in the cell and are also crucial for telomere maintenance. It was reported that Rap1 together with the Rif-complex represses the elongation of telomeres via a negative feedback loop (Yang *et al.*, 2017). Furthermore, Est1 can be degraded by the proteasome and is not present during G1 phase (Osterhage *et al.*, 2006). This indicates that *EST1* can be part of the negative feedback loop together with the Rap1 interacting proteins Sum1 and Hst1.

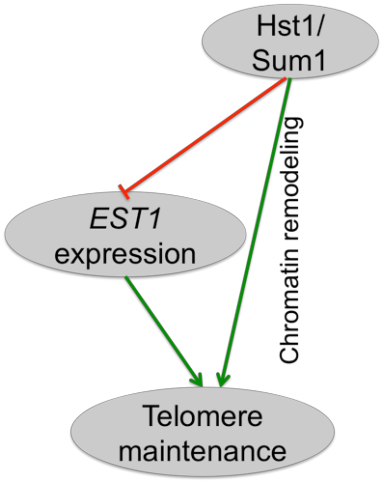

Figure 28. Incoherent feed-forward loop.
Hst1/Sum1 negatively regulates *EST1*, whereby Hst1/Sum1 and *EST1* positively regulate telomere length. Image taken from (Poos *et al.*, 2016).

For *EST2*, only Gln3 was identified as a significant regulator in the MIPRIP analysis. Gln3 has not been described in the context of telomere maintenance before, but a deletion of *GLN3* leads to a strong downregulation of *EST2*. According to our preditions, Sin3, Dig1, Srb2, Hir1 and Ume6 were significant regulators of *EST3*.

For all these regulators no direct link to telomerase regulation was reported so far. Ume6 and Sin3 together repress the expression of meiotic genes (Lardenois *et al.*, 2015). But it was also reported previously that Ume6 can act as a positive regulator (Rubin-Bejerano *et al.*, 1996; Washburn and Esposito, 2001). From these regulators, a deletion of *UME6* is associated with a strong upregulation of *EST3*, while *EST3* is downregulated in the deletion strains *sin3*, *srb2*, *dig1* and *hir1*.

In summary, MIPRIP was applied to study the regulation of the three *EST* genes in yeast deletion strains with shorter telomeres compared to deletion strains with normal telomere length. This resulted in novel regulators of the telomerase holoenzyme and several of these regulators affect histone levels or modifications. Especially Sum1 had a high influence on the expression of *EST1*, which could also be validated in two cell lines.

4.3 Pan-cancer analysis of *TERT* regulation

The activation of telomere maintenance mechanisms is crucial for cancer cells to enable their replicative immortality. Most cancer cells maintain their telomeres by re-activating telomerase, a reverse transcriptase consisting of the catalytic subunit *TERT* and the template RNA *TERC* (Sandin and Rhodes, 2014). *TERC* is consistently expressed, but the expression of *TERT* is the limiting factor of telomerase activity (Feng *et al.*, 1995; Kim *et al.*, 1994) indicating that the regulation of *TERT* is an interesting research topic. Therefore, MIPRIP was applied to study the regulation of *TERT* in 19 different cancer entities. For this gene expression data from TCGA was used together with the generic human regulatory network. In the generic human regulatory network 75 TFs of *TERT* could be identified. The identified TFs of *TERT* from the databases showed a high overlap with 54 potential *TERT* regulators described in a review by Ramlee *et al.* (Ramlee *et al.*, 2016). The TFs described in Ramlee *et al*. were limited to ChIP-experiments or EMSA, while the generic human regulatory network integrates also computationally predictions of TFBS.

For the pan-cancer analysis the multi-mode of MIPRIP-Comp was used to identify (i) the most common *TERT* regulators over all cancer entities and (ii) the significant *TERT* regulators of each cancer type *vs.* all other cancer types. Nearly all cancer

types showed a good performance over all 300 models. Thymoma and testicular germ cell cancer showed the highest *TERT* expression and also the best performance. For melanoma skin cancer the worst performance was observed, even though the expression of *TERT* was not lowest.

4.3.1 Nine regulators were predicted to be involved in *TERT* regulation in all different cancer entities

The multi-mode MIPRIP analysis led to nine regulators of *TERT* which were significant across all cancer types. To validate these regulators *in silico*, I performed a Pubmed query with the regulator symbols in the context of telomerase regulation. This search led to a significant enrichment of Pubmed entries of these regulators together with *TERT*. From the nine common *TERT* regulators, AR, E2F2, E2F4, PAX5 and PAX8 have been described in the literature as regulators of *TERT*. The androgen receptor (AR) is a nuclear receptor and has been reported as a repressor of *TERT* in prostate cancer. Treatment with an AR agonist inhibited the promoter activity of *TERT*, while treatment with an AR antagonist did not result in the same effect. If AR is mutated, the recruitment to the *TERT* promoter was less efficient (Moehren *et al.*, 2008). Besides prostate cancer, AR was also identified as a significant hit in bladder, breast, colorectal and ovary cancer. Several members of the family of E2F TFs were found as significant in 13 out of the 19 different cancer entities. Factors of the E2F TF family are part of the DNA damage response and are also involved in cell cycle. They can bind to the E2 recognition motif (Crowe *et al.*, 2001). E2F2 is an activator of *TERT* and E2F4 was found as a regulator of *TERT* in human B-cell lymphoma (Chebel and Ffrench, 2010; Mani *et al.*, 2008). PAX5 and PAX8 were identified in nearly half of the cancer types (9 out of 19) and are active during early development. They have 2 respectively 4 binding sites at the TSS of the *TERT* promoter to activate the transcription of *TERT* (Bougel *et al.*, 2010; Chen *et al.*, 2008). A siRNA knockdown as well as an over-expression assay of PAX5 in lymphocytes showed that PAX5 is an activator of *TERT* (Bougel *et al.*, 2010). PAX8 is also an activator of *TERT.* In addition, it can also activate *TERC* (Chen *et al.*, 2008). Besides the 5 well characterized *TERT* regulators, I identified the regulators BATF (predicted for 6 out of 19 cancer types), MXI1 (3 out of 19), SMARCB1 (4 out of 19) and TAF1 (5 out of 19) as significant *TERT* regulators of all cancer types. These regulators have not been described in the literature so far. SMARCB1 is part

of the SWI/SNF chromatin-remodeling complex (Wang *et al.*, 2017). In *S. cerevisiae*, it was reported that the Swi/Snf complex is involved in the silencing of genes close to telomeres (Dror and Winston, 2004). In humans, a depletion of the SWI/SNF related gene SMARCAL1 led to an ALT-like phenotype (Poole *et al.*, 2015) suggesting that also other members of the SWI/SNF complex may be involved in telomere maintenance.

In summary, MIPRIP performed well to predict the regulation of *TERT* in different cancer types. Several well-described regulators of *TERT* could be identified, but also new promising regulators of *TERT* that would need further experimental validation.

4.3.2 *TERT* promoter mutations led to different regulatory mechanisms in melanoma

In the pan-cancer analysis of *TERT*, the modeling performance was worst for melanoma. As melanoma skin cancer is one of the cancer entities with a high rate of *TERT* promoter mutations, the melanoma skin cancer dataset was divided into a group of samples showing a *TERT* promoter mutation and a control group with wild-type *TERT* promoter to improve the modeling performance. With these subgroups a MIPRIP dual-mode analysis was performed to identify the regulators which were significantly selected more often in one of the groups compared to the other group. For the melanoma skin cancer samples with *TERT* promoter mutation the regulators AR, E2F1, ETS1 and JUND were most significant. AR was also a significant common *TERT* regulator in the pan-cancer analysis described above. So far, AR has been identified as a *TERT* regulator only in prostate cancer. But AR can also increase cell invasion in melanoma cells (Wang *et al.*, 2017). E2F1 is a repressor of *TERT* (Crowe *et al.*, 2001). It was recently shown that an inhibition of E2F1 can increase the cell death rate in melanoma cells, even for melanoma cells resistant to BRAF-inhibitors (Rouaud *et al.*, 2018). Hence, E2F1 is an interesting therapeutic target in melanoma cells. But from the identified regulators specific for the melanoma samples with a *TERT* promoter mutation ETS1 was the most prominent hit as these mutations create an additional binding site for TFs of the ETS-family (Horn *et al.*, 2013; Huang *et al.*, 2013). *TERT* promoter activity is enhanced by ETS binding and a p52 dependent activation of the non-canonical NF-κB signaling pathway (Li *et al.*, 2015). It was furthermore shown that these *TERT* promoter

mutations can lead to a two- to four-fold higher *TERT* promoter activity in melanoma cells (Horn *et al.*, 2013; Huang *et al.*, 2013). I analyzed publicly available expression data of an ETS1 siRNA knockdown in a melanoma cell line with a *TERT* promoter mutation. *TERT* was downregulated in the knockdown sample compared to controls showing the activating effect of ETS1 on *TERT* expression. To evaluate the performance of MIPRIP, the MIPRIP results of the melanoma case study were compared to results obtained with the well-established tool ISMARA. However, the overlap between both tools was very low. ISMARA was not able to predict ETS1 as an important regulator of *TERT* in the samples with the promoter mutation, but instead GABPA, another member of the ETS-family, was predicted with very low significance. It was shown that GABPA can bind only to the *TERT* promoter mutation at site C228T and not at C250T (Mancini *et al.*, 2018). However, only one third of the mutated samples had the mutation at C228T, while two-third showed a C250T mutation (Cancer Genome Atlas, 2015). Hence, it would be interesting in the future to perform a MIPRIP analysis separately for the C228T and the C250T mutated samples.

For the melanoma skin cancer samples with the wild-type *TERT* promoter, the regulators HMGA2, HIF1, RUNX2 and TAL1 were identified as most significant *TERT* regulators. HMGA2 belongs to the high-mobility group of AT-hook proteins. These proteins are expressed during embryonic development (Chiappetta *et al.*, 1996), in several benign tumors like lipomas (Schoenmakers *et al.*, 1995) and in malignant tumors of the vulva (squamous cell carcinoma and malignant melanoma) (Agostini *et al.*, 2015). Only a few malignant vulva samples showed a *TERT* promoter mutation, but HMGA2 was expressed in nearly all samples (Agostini *et al.*, 2015). This suggests that *TERT* and HMGA2 are involved in tumorigenesis, while the association between *TERT* promoter mutations and HMGA2 expression is still unclear. The model suggests that HMGA2 regulates *TERT* only in the absence of a *TERT* promoter mutation, for which an experimental validation would be interesting.

With the melanoma case study, I showed that splitting up the datasets into distinct subtypes increases the modeling performance and is necessary to identify the exact regulatory mechanisms. Melanoma skin cancer patients with a *TERT* promoter mutation have a decreased survival rate compared to patients without this mutation (Griewank *et al.*, 2014). The subtype specific regulators can be used as biomarkers to improve risk stratification or for targeted therapies.

4.4 Regulatory network explaining *TERT* regulation in prostate cancer

Until now, neither *TERT* promoter mutations nor ALT occurrence have been detected in prostate cancer (Heaphy *et al.*, 2011). Therefore, prostate cancer is well suited to find new regulators of telomerase. The pan-cancer MIPRIP analysis led to 17 significant regulators that were specific for prostate cancer compared to all other cancer entities (Table 10). As there were also some healthy prostate samples available, I compared the regulatory processes between prostate cancer and healthy prostate tissue. This led again to 17 cancer-specific regulators. Although, the *TERT* expression in healthy tissue samples was very low, 12 regulators were overlapping between both MIPRIP analyses. With these 12 prostate cancer specific *TERT* regulators and their most important regulators, I performed a MIPRIP-Net analysis to identify a regulatory network explaining *TERT* regulation. After the optimization of the tradeoff-parameter, the subnetwork, which (i) could best predict the expression of *TERT* and (ii) was highly connected with other regulators, consisted of six significant direct *TERT* regulators as well as fourteen additional regulators. The six significant direct *TERT* regulators were BHLHE40, CTCF, IRF1, MITF, PITX1 and TFAP2D. From the additional regulators, seven regulators can potentially bind to the *TERT* promoter, SMC3, PML and USF1 showed a link to telomere maintenance (based on the TelNet database) and POLR2A is relatively often mutated in prostate cancer patients (8%) which indicates that our subnetwork includes several telomere maintenance relevant hits. To validate the MIPRIP predictions, our collaboration partners at the University Hospital Hamburg-Eppendorf, the group of Guido Sauter/Ronald Simon, performed an IHC-staining of the six significant direct regulators BHLHE40, CTCF, IRF1, MITF, PITX1 and TFAP2D on TMAs of approximately 17,000 patients. Here, we focused on PITX1 as it was the most significant regulator of *TERT* in both MIPRIP analysis.

The paired-like homeodomain 1 (PITX1) gene was originally described as a developmental factor. A study in mice showed that the enhancer *Pen* can regulate Pitx1 only in hindlimbs because of the hindlimb's specific 3D chromatin structure (Kragesteen *et al.*, 2018). PITX1 is described as a tumor suppressor (Otsubo *et al.*, 2017), an activator of p53 (Liu and Lobie, 2007), an inhibitor of the RAS pathway

(Kolfschoten *et al.*, 2005) and in hypoxia cells PITX1 promotes proliferation through the HIF1α response (Mudie *et al.*, 2014). All these facts make PITX1 an interesting target for cancer therapy. There are also two studies which showed that PITX1 is a substrate of the protein tyrosine phosphatase 1B (PTP1B), which can be inhibited by Sorafenib in hepatocellular carcinoma (Tai *et al.*, 2016) or Regorafenib in colorectal carcinoma (Teng *et al.*, 2016). In the last one to two years many papers described PITX1 as a potential biomarker in different cancer types, e.g. melanoma (Osaki *et al.*, 2013) and oral epithelial dysplasia (Nakabayashi *et al.*, 2014). A high PITX1 expression is associated with a favorable outcome in osteosarcoma (Kong *et al.*, 2015), colorectal (Knosel *et al.*, 2012) and gastric cancer (Qiao *et al.*, 2018) as well as esophageal squamous cell carcinoma (Otsubo *et al.*, 2017). In lung adenocarcinoma patients, a high PITX1 expression is linked to DNA methylation and a poor prognosis (Song *et al.*, 2018). PITX1 is described to be low expressed in different cancer types, including prostate and bladder cancer (Kolfschoten *et al.*, 2005). This is in high agreement with our data, where only 4 % of all analyzed tumor samples showed a high PITX1 expression. Compared to normal prostate tissue PITX1 is upregulated in around two-thirds of the prostate cancer patients. It turned out that PITX1 is a well suited prognostic marker in prostate cancer, which is positively correlated with *ERG*-fusions as around 80 % of the prostate cancer patients upregulate PITX1 (55 % in the *ERG*-fusion negative group). An upregulation of PITX1 led to a poorer survival in *ERG*-fusion positive and negative patients, a higher cell proliferation and all kind of tumor aggressiveness (advanced tumor stage, high Gleason grade, presence of lymph node metastasis, higher levels of pre-operative PSA and positive surgical margin). Caroline Bauer from the group of Karsten Rippe (DKFZ & BioQuant Heidelberg) is currently investigating the effect of PITX1 on *TERT* expression and telomerase activity in the 5 prostate cancer cell lines DU145, LnCap, C4-2, PC-3 and PC3-AR (provided by the lab of Aria Baniahmad at the University Hospital Jena). So far, PITX1 was described as a suppressor of mtert expression and telomerase activity in melanoma cells (Qi *et al.*, 2011). Qi *et al*. imported a human chromosome 5 into the mouse melanoma cell line B16F10 by using microcell-mediated chromosome transfer. They found that PITX1 can directly bind to the *mtert/hTERT* promoter (1 binding site at *mtert* promoter, 3 at *hTERT*). In a further study they showed that PITX1 is directly regulated by the microRNA-19b (miR-19b). An inhibition of PITX1 expression by miR-19b is

associated with an increased *hTERT* transcription and proliferation in melanoma cells (Ohira *et al.*, 2015). It was furthermore shown that PITX1 binds to the *TERT* promoter in gastric cancer (Qiao *et al.*, 2018). In addition to prostate cancer I identified PITX1 as a significant *TERT* regulator also in head and neck squamous cell carcinoma, ovary and cervical cancer (Poos *et al.*, 2019). For head and neck squamous cell carcinoma, PITX1 was described as a promising biomarker because a DNA hypermethylation of PITX1 leads to a poorer prognosis (Sailer *et al.*, 2018) and PITX1 expression could be used to predict chemotherapy response (Takenobu *et al.*, 2016).

The subnetwork includes also CTCF, IRF1, TFAP2D, BHLHE40 and MITF as significant regulators of *TERT* in prostate cancer. The CCCTC-binding factor (CTCF) is a chromatin-binding factor which can bind to more than 20,000 DNA loci in the human genome (Ohlsson *et al.*, 2001). CTCF is crucial for the organization of the three-dimensional chromatin structure (Ong and Corces, 2014). Furthermore, it is involved in the transcriptional regulation of thousands of genes where it can act as activator, but for most genes CTCF has a repressive effect (Ramlee *et al.*, 2016). CTCF can bind to the first exon of *TERT* suppressing its expression in telomerase-negative non-cancer cells (Renaud *et al.*, 2007), preferentially to unmethylated sites. It was furthermore reported to bind to an enhancer region around 4.5 kb upstream of the TSS of the *TERT* promoter repressing its expression (Eldholm *et al.*, 2014). We identified CTCF as a good prognostic marker for *ERG*-fusion negative prostate cancer patients (Höflmayer *et al.*, 2019). Furthermore, a first IHC staining of IRF1 and TFAP2D showed that both could be potential prognostic markers for prostate cancer. IRF1 is involved in the IFN-gamma signaling repression of both *TERT* expression and telomerase activity (Lee *et al.*, 2003). MITF and BHLHE40 could not be clinically validated, because there was no suitable antibody available. Still, overall four out of the six direct regulators of our subnetwork suit as prognostic marker and a high expression of at least 3 of the 4 markers led to a poorer PSA-progression free survival than one marker alone. This is in line with the co-operative effect of TFs in the MIPRIP model.

CEBPA was also a prostate cancer specific *TERT* regulator in the pan-cancer MIPRIP analysis but was not included in the “best” subnetwork. The CCAAT/Enhancer binding protein alpha (CEBPA) is a leucine zipper transcription factor (Johnson *et al.*, 1987; Landschulz *et al.*, 1988) and is involved in cell

proliferation (Wang *et al.*, 2001), terminal differentiation (Chen *et al.*, 1996; Timchenko *et al.*, 1997), the control of inflammatory and immune response (Bristol *et al.*, 2009; Poli, 1998) as well as maintenance of energy metabolism (Wang *et al.*, 1995). It has been furthermore described as a repressor of *TERT* and its loss is correlated with activation of *TERT* expression during tumorigenesis (Kumar *et al.*, 2013). CEBPA modulates the transcription of androgen responsive genes (Zhang *et al.*, 2010) and two studies showed that alterations of CEBPA expression are associated with higher Gleason grades (Yin *et al.*, 2009; Yin *et al.*, 2006) suggesting an important role of CEBPA in prostate cancer. CEBPA was also validated by our collaboration partners. Here, we found a strong prognostic impact of CEBPA for ERG-fusion positive tumor samples with a PTEN deletion. For this subgroup a loss of CEBPA expression was significantly associated with a poor outcome (p-value = 0.0011), an advanced tumor stage (p-value < 0.0001), a high Gleason grade (p-value < 0.0001), high preoperative levels of PSA (p-value = 0.0066) and positive nodal stage (p- value = 0.0003) (Minner *et al.*, 2019).

Nowadays, the PSA level and the Gleason grading are the best-established parameters for prostate cancer. However, the postoperative determination of the Gleason grade is more precise than the preoperative determination (Epstein *et al.*, 2012). To optimize the grading of the biopsies novel biomarkers can be very important and can be included in treatment decisions or can prevent patients from a radical prostatectomy. So far, the Gleason grade is a quantitative estimation by a pathologist and in biopsies there are sometimes only a small amount of tumor glands present. Thus it is especially important for biopsies to have good prognostic markers, and the here identified candidates suit very well for this. To validate new potential biomarker candidates, it is very helpful to have such a large collection of around 17,000 prostate cancer patients as shown in this study.

An optimized Gleason grading system was developed by our consortium partners (Department of Guido Sauter, University Hamburg-Eppendorf), which is called the quantitative Gleason score (IQ Gleason) (Sauter *et al.*, 2018). In an area under curve (AUC) analysis this IQ Gleason grading was 5 % better than the conventional Gleason grading. This new grading includes the percental fraction of Gleason grade 4 and 5 glands in the probe and if also glands with a Gleason grade 5 are present a value of 10 is added. If more than 20 % of the glands in the probe are of Gleason grade 5 a further value of 7.5 is added. In summary, this means that the IQ-Gleason

ranges from 0-117.5. For instance, a sample with a Gleason score of 4+3=7 and 55 % Gleason 4 has an IQ-Gleason of 55, while the IQ-Gleason of a sample with a Gleason score of 4+5=9 (55 % Gleason 4 and 45 % Gleason 5) is 117.5 (55+45+10+7.5). Most markers today did not improve the stratification compared to the IQ Gleason. For PITX1, there was still a stratification for several groups with the same IQ Gleason. But the sample size was quite low and the effect was not significant.

In summary, I developed a novel method integrating MIPRIP with a modularity-based approach and applied it to investigate the regulatory subnetwork which best explains the regulation of *TERT* in prostate cancer patients. The extension of MIPRIP led to a broader view on *TERT* regulation in prostate cancer because also the regulators indirectly influencing the expression of *TERT* by interacting with other regulators were integrated. Several of the identified *TERT* regulators could be validated as novel biomarkers for prostate cancer. According to our predictions, PITX1 was the most significant *TERT* regulator in prostate cancer. These biomarkers are highly relevant for clinicians to decide at the biopsy level if a radical prostatectomy has to be performed.

4.5 Patient stratification according to telomere maintenance mechanisms

Glioblastoma multiform (GBM) patients have a very poor prognosis with a median survival of less than 9 months and limited treatment options (Hakin-Smith *et al.*, 2003; Sturm *et al.*, 2014). To improve the treatment options, it is necessary to stratify the tumors according to their underlying biological processes. One promising approach is the stratification of GBM patients based on telomere maintenance mechanisms (TMMs). While in most adult GBMs telomerase is re-activated to maintain telomeres, in 44 % of pediatric GBMs an ALT pathway is active (Hakin-Smith *et al.*, 2003; Heaphy *et al.*, 2011; Schwartzentruber *et al.*, 2012). In a recent study 7 pedGBM cell lines with different genetic backgrounds in the ATRX/DAXX/H3.3 axis and 57 pedGBM patient samples were characterized regarding typical TMM features extracted from sequencing data as well as cytological and molecular assays (Deeg *et al.*, 2017).

While some telomerase inhibitors were tested in clinical trials, a promising drug target specific for ALT-positive tumors has not yet been identified. To use TMM as a therapeutical target a reliable TMM classification is crucial. As previously shown ALT-positive GBM patients have a longer survival rate (11.9 months compared to 10.2 months) (Hakin-Smith *et al.*, 2003; Mangerel *et al.*, 2014; McDonald *et al.*, 2010). The presence of C-circles is a well-established marker for ALT (Cesare and Reddel, 2010; Henson and Reddel, 2010). For the C-circle assay only a very little amount of DNA (60 ng) is required. Hence, it would be also an interesting option to integrate this assay into the clinical routine. But the instability of the single-stranded C-circles can lead to false-negative results (Henson *et al.*, 2009). Therefore, a combination with other ALT assays is essential to get a reliable patient stratification. So far, the detection of ultra-bright telomere foci by FISH (Mangerel *et al.*, 2014) is the most frequent used technique to identify ALT. However, for this technique tissue sections are required which is a major disadvantage of this method because patient material is always limited. Furthermore, a quantitative evaluation is difficult (Gunkel *et al.*, 2017). To construct a classifier which can distinguish between ALT and non-ALT samples first a set of training samples with known ALT status had to be determined. A combination of features extracted from DNA-, RNA-seq or DNA methylation assays resulted in a reliable TMM prediction. As sequencing readouts are more and more used in the clinical routine, no additional assays have to be performed. Instead of performing DNA- or RNA-seq it would be also possible to determine the expression level or mutation status of the most reliable features by PCR. One limitation of the classifier here was the low sample size together with the incomplete feature list for most of the samples. This makes the prediction difficult and only comparable for similar sample sizes. The advantage of a decision tree-based approach is that features from different datatypes (e.g. sequencing and experimental data) can be easily integrated. In the present form, the classifier distinguishes only between ALT-positive and ALT-negative samples. More specific subgroups could have been defined if the sample set would have been larger. For example, a discrimination of ALT-negative samples with activated telomerase or the presence of the ever shorter telomere phenotype. Furthermore, the possibilities of both TMMs being active in the same tumor sample or a TMM switch within a tumor sample exist.

The classifier was implemented as a web tool called ‘Predicting ALT IN Tumors’ (PAINT) by Nick Kepper from the group of Prof. Rippe (DKFZ, Heidelberg). At the moment PAINT is only available for pedGBM, however, an extension for other cancer entities, e.g. neuroblastoma or soft tissue sarcoma has been considered. Such an extension would require a sample set with a size of at least 100 samples for training purposes.

A number of studies with attempts of identifying a gene expression signature to distinguish between ALT and non-ALT samples (Barthel *et al.*, 2017; Doyle *et al.*, 2012; Lafferty-Whyte *et al.*, 2009). To my knowledge, no reliable TMM gene signature has been identified. A possible explanation for this could be that tissue and cell type specific effects are much higher than the difference between ALT and non-ALT.

Therefore, there is an unmet need to identify a TMM gene signature for pedGBM. Accordingly, I performed a differential gene expression analysis together with a calculation of regulator activities (as described for the MIPRIP framework). Altogether, I identified 115 genes (p-value < 0.01) and 15 regulators (p-value < 0.05) as differential expressed/active between ALT-positive and ALT-negative samples. A clustering based on the differential expressed genes did not lead to a reliable prediction into ALT and non-ALT. Only the clustering based on the activities of the differential active regulators identified 2 separate clusters. In the ALT-negative samples, 15 regulators showed a significantly higher activity compared to the ALT-positive samples, including known *TERT* regulators: SIN3A, ETS1, PAX5, MXI1, POU2F2 and BCL11A. These regulators were downregulated in all ALT samples and only in 4 out of 20 non-ALT samples. Therefore, it seems that these six regulators were highly important for telomerase activity in the non-ALT samples. The classification based on regulator activities performed much better than a standard gene expression analysis. Thus, regulator activities are a powerful and robust method for many different applications. Furthermore, it would be interesting to test if the regulator signature is also valid for other cancer entities (e.g. neuroblastoma) to distinguish between ALT positive and ALT negative.

As described before H3.3 G34R/V and K27M mutant tumors showed specific gene expression profiles that are distinct from each other and from tumors with wild-type H3.3 (Bender *et al.*, 2013; Schwartzentruber *et al.*, 2012; Sturm *et al.*, 2012). In this study, 24 out of 57 pedGBM patient samples harbored a mutation in H3F3A (15 ALT

and 9 non-ALT). Therefore, it is important to keep in mind that there exist different subtypes (e.g. H3.3 K27M *vs*. G34R/V) when talking about gene signatures. Because of the low sample number, it was not possible to identify a reliable cancer subtype specific expression profile which can be also linked to TMM.
In summary, I developed a reliable classification schema to stratify pedGBM patients into ALT and non-ALT, which has the potential to be used in the clinical routine in order to improve patient stratification and in-line selection of treatment strategies.

4.6 Comparison of the MIPRIP approach with ARACNE and VIPER

To elucidate the advantages and disadvantages of the MIPRIP approach, I compared MIPRIP with the well-established tools ARACNE (Lachmann *et al.*, 2016) and VIPER (Alvarez *et al.*, 2016). VIPER calculates regulator activities based on the expression of the target genes, which are determined from a network calculated with the ARACNE algorithm. ARACNE calculates a *de-novo* network from gene expression data using mutual information (Lachmann *et al.*, 2016). The activities can then be calculated either across multiple samples or separately for each sample (Alvarez *et al.*, 2016). As shown for the melanoma case study, MIPRIP is also compared to ISMARA well suited to predict regulators of *TERT* in melanoma samples with and without *TERT* promoter mutation (see 3.3.3). As *TERT* is usually weakly expressed (Ducrest *et al.*, 2001; Fredriksson *et al.*, 2014), ARACNE was not able in contrast to MIPRIP to detect reliable regulators of *TERT*. Therefore, the comparison between MIPRIP and ARACNE/VIPER was performed with gene expression data of 20 CLL patients and 7 healthy controls instead of using the melanoma case study of *TERT*. The goal here was to identify regulators that significantly change their activity between the CLL and the non-malignant B-cell samples. Because of the low sample size of the CLL-data, a B-cell specific ARACNE network was constructed based on publicly available microarray gene expression data of 267 malignant and non-malignant B-cell samples (Basso *et al.*, 2010). The ARACNE network is not limited to TFs as MIPRIP and includes also co-regulators and signaling proteins. First, I compared the number of target genes of the regulators in the ARACNE B-cell network and the generic human regulatory network. In both networks most of the regulators had less than 500 target genes.

But the generic regulatory network contains some master regulators with 10,000 to 15,000 target genes indicating the research intensity of these factors. The generic regulatory network integrates biological knowledge in form of ChIP-experiments, while ARACNE uses just gene expression data to predict regulator to target gene interactions. Furthermore, ARACNE requires large datasets (Ding *et al.*, 2018) and the network construction is computationally very intensive. Second, I compared the regulators that had a significant change in their activity values between the CLL and the healthy samples. Here, I found a high overlap between the significant regulators of both methods. This showed that the activity calculation led to similar results although the regulator to target gene interactions were not similar between both networks. Both approaches seem to have problems with regulators where the activating and repressing effect of the target genes are similar.

In summary, the MIPRIP approach performs well in combination with the well-established tools ARACNE and VIPER. Therefore, MIPRIP is well suited to study the regulatory processes of a particular gene. Furthermore, the high overlap between the regulator activities calculated with the MIPRIP approach and with VIPER indicate that an ARACNE network could be used for the MIPRIP approach instead of the generic regulatory network. This would be an interesting option to include also co-factors and signaling molecules into the MIPRIP model, although this would work only for genes that are higher expressed and have several interacting partners in the network.

Conclusions and perspectives

In this thesis, I developed a new tool called 'Mixed Integer linear Programming based Regulatory Interaction Predictor' (MIPRIP) to identify the most important regulators of a gene of interest. MIPRIP can be used to compare the regulatory processes between two different datasets/conditions (e.g. treatment vs. control) or multiple datasets/conditions (e.g. pan-cancer analysis) (MIPRIP-Comp). Furthermore, MIPRIP was extended with a modularity-based approach to identify a gene regulatory network of TFs regulating the gene of interest (MIPRIP-Net).

Its application on the regulation of telomere maintenance is summarized in Figure 29.

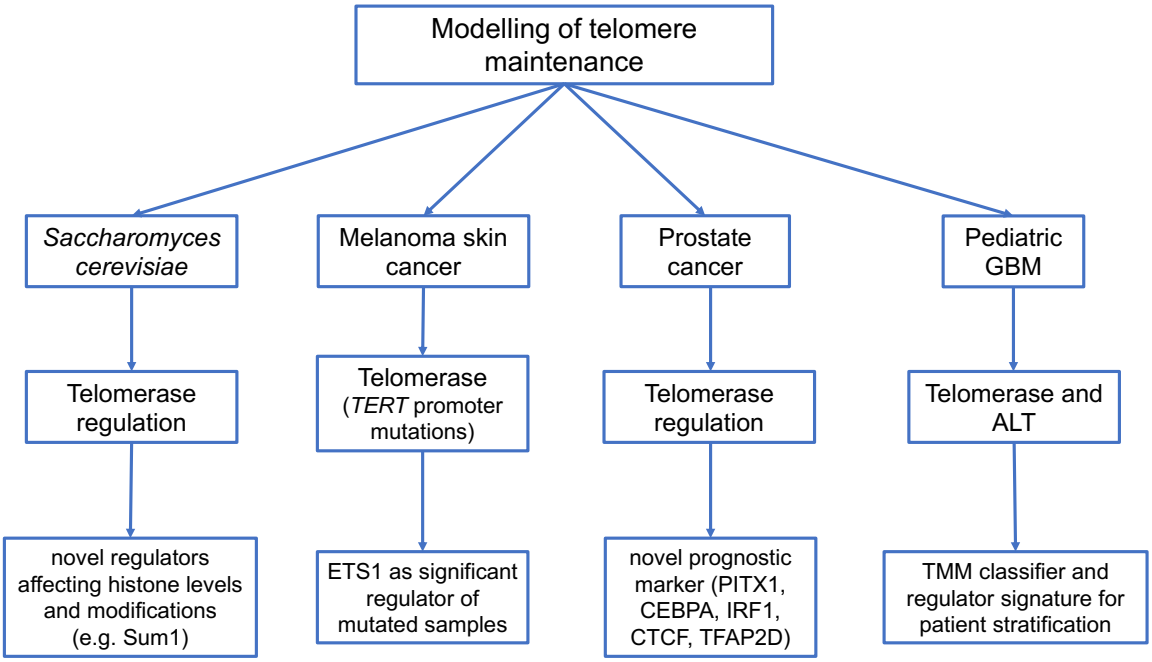

Figure 29. Modeling of telomere maintenance in yeast and different cancer types

First, MIPRIP was applied to identify novel regulators of the telomerase holoenzyme (*EST* genes) in *S. cerevisiae*. Several of the identified regulators affect histone levels or modifications. The most prominent hit was Sum1, which could be validated experimentally as a regulator of *EST1*. Second, I performed a pan-cancer analysis to study the regulation of human telomerase (*TERT*), which identified generic as well as cancer entity specific regulators. Using melanoma skin cancer was as a case study, I showed that MIPRIP was able to identify the well-known *TERT* regulator ETS1 (Horn *et al.*, 2013; Huang *et al.*, 2013) as a significant *TERT* regulator in melanoma samples with a *TERT* promoter mutation. The novel concept of MIPRIP-

Net led to a gene regulatory network of 20 TFs that was predicted to regulate *TERT* expression in prostate cancer. From these 20 regulators, PITX1, CTCF, IRF1, TFAP2D, MITF and BHLHE40 were most significant. An IHC staining of these 6 regulators on TMAs of around 17,000 patients showed that PITX1, CTCF, IRF1 and TFAP2D are novel prognostic markers in prostate cancer. Thus, I could identify TFs whose activity can be determined in biopsy samples to support clinical decisions on radical prostatectomy. These interesting hits are currently validated via knock-down in prostate cancer cell lines.

Third, for pediatric GBM, a cancer entity with a high occurrence of ALT, I constructed a decision tree-based classifier to stratify pedGBM patients according to their active TMM (ALT or non-ALT) by using typical TMM features extracted from sequencing-based readouts as well as from cytological and molecular assays. Here, I could show that a combination of different sequencing-based readouts leads to a highly reliable TMM prediction. After patient stratification into ALT and non-ALT samples, I identified a regulator signature that is downregulated in ALT pedGBM patient samples. Several of these regulators can bind to promoter of *TERT*. It will be interesting to test if the activity of these regulators correlates with clinical parameters.

Last, I compared the MIPRIP framework with the well-established tools ARACNE and VIPER from the Califano lab. Here, I found a high overlap between the TFs showing a significantly different activity between the CLL and the non-malignant B-cell samples calculated with MIPRIP and VIPER. This indicates that a network constructed using the ARACNE algorithm could be combined with MIPRIP to integrate also transcriptional co-factors, chromatin modifiers and other signaling molecules into the regulatory models.

In summary, MIPRIP identified new regulators of the telomerase, several of which could be validated as novel prognostic markers of prostate cancer. Thus, MIPRIP is well suited to study gene regulation.

References

Agostini, A., Panagopoulos, I., Andersen, H.K., Johannesen, L.E., Davidson, B., Trope, C.G., . . . Micci, F. HMGA2 expression pattern and TERT mutations in tumors of the vulva. *Oncology reports* 2015; 33(6):2675-2680.

Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., Conti, D.V., Schumacher, F., Han, Y., . . . Haiman, C.A. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* 2014; 46(10):1103-1109.

Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H. and Califano, A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 2016; 48(8):838-847.

Askree, S.H., Yehuda, T., Smolikov, S., Gurevich, R., Hawk, J., Coker, C., . . . McEachern, M.J. A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length. *Proc Natl Acad Sci U S A* 2004; 101(23):8658-8663.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 2004; 14(3):283-291.

Balwierz, P.J., Pachkov, M., Arnold, P., Gruber, A.J., Zavolan, M. and van Nimwegen, E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res* 2014; 24(5):869-884.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., . . . Garraway, L.A. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 2012; 44(6):685-689.

Barthel, F.P., Wei, W., Tang, M., Martinez-Ledesma, E., Hu, X., Amin, S.B., . . . Verhaak, R.G. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat Genet* 2017; 49(3):349-357.

Basso, K., Saito, M., Sumazin, P., Margolin, A.A., Wang, K., Lim, W.K., . . . Dalla-Favera, R. Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. *Blood* 2010; 115(5):975-984.

Batista, R., Cruvinel-Carloni, A., Vinagre, J., Peixoto, J., Catarino, T.A., Campanella, N.C., . . . Lima, J. The prognostic impact of TERT promoter mutations in glioblastomas is modified by the rs2853669 single nucleotide polymorphism. *Int J Cancer* 2016; 139(2):414-423.

Bauer, T., Eils, R. and Konig, R. RIP: the regulatory interaction predictor--a machine learning-based approach for predicting target genes of transcription factors. *Bioinformatics* 2011; 27(16):2239-2247.

Bedalov, A., Hirao, M., Posakony, J., Nelson, M. and Simon, J.A. NAD+-dependent deacetylase Hst1p controls biosynthesis and cellular NAD+ levels in Saccharomyces cerevisiae. *Molecular and cellular biology* 2003; 23(19):7044-7054.

Beisser, D., Brunkhorst, S., Dandekar, T., Klau, G.W., Dittrich, M.T. and Muller, T. Robustness and accuracy of functional modules in integrated network analysis. *Bioinformatics* 2012; 28(14):1887-1894.

Ben-Aroya, S., Coombes, C., Kwok, T., O'Donnell, K.A., Boeke, J.D. and Hieter, P. Toward a comprehensive temperature-sensitive mutant repository of the essential genes of Saccharomyces cerevisiae. *Mol Cell* 2008; 30(2):248-258.

Ben-Shitrit, T., Yosef, N., Shemesh, K., Sharan, R., Ruppin, E. and Kupiec, M. Systematic identification of gene annotation errors in the widely used yeast mutation collections. *Nat Methods* 2012; 9(4):373-378.

Bender, S., Tang, Y., Lindroth, A.M., Hovestadt, V., Jones, D.T., Kool, M., . . . Pfister, S.M. Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. *Cancer Cell* 2013; 24(5):660-672.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 1995; 57(1):289-300.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995; 57:289-300.

Bianchi, A. and Shore, D. Early replication of short telomeres in budding yeast. *Cell* 2007; 128(6):1051-1062.

Bianchi, A. and Shore, D. Increased association of telomerase with short telomeres in yeast. *Genes Dev* 2007; 21(14):1726-1730.

Blackburn, E.H. Telomeres: structure and synthesis. *J Biol Chem* 1990; 265(11):5919-5921.

Board., P.S.a.P.E. Prostate Cancer Screening (PDQ(R)): Patient Version. In, *PDQ Cancer Information Summaries*. Bethesda (MD); 2002.

Bougel, S., Renaud, S., Braunschweig, R., Loukinov, D., Morse, H.C., 3rd, Bosman, F.T., . . . Benhattar, J. PAX5 activates the transcription of the human telomerase reverse transcriptase gene in B cells. *J Pathol* 2010; 220(1):87-96.

Bovolenta, L.A., Acencio, M.L. and Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 2012; 13:405.

Braun, D.M., Chung, I., Kepper, N., Deeg, K.I. and Rippe, K. TelNet - a database for human and yeast genes involved in telomere maintenance. *BMC Genet* 2018; 19(1):32.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018.

Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 2004; 573(1-3):83-92.

Breslow, D.K., Cameron, D.M., Collins, S.R., Schuldiner, M., Stewart-Ornstein, J., Newman, H.W., . . . Weissman, J.S. A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* 2008; 5(8):711-718.

Bristol, J.A., Morrison, T.E. and Kenney, S.C. CCAAT/enhancer binding proteins alpha and beta regulate the tumor necrosis factor receptor 1 gene promoter. *Mol Immunol* 2009; 46(13):2706-2713.

Butler, K.S., Hines, W.C., Heaphy, C.M. and Griffith, J.K. Coordinate regulation between expression levels of telomere-binding proteins and telomere length in breast carcinomas. *Cancer Med* 2012; 1(2):165-175.

Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* 2015; 161(7):1681-1696.

Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 2015; 163(4):1011-1025.

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., . . . Pfister, S.M. DNA methylation-based classification of central nervous system tumours. *Nature* 2018; 555(7697):469-474.

Castelo-Branco, P., Choufani, S., Mack, S., Gallagher, D., Zhang, C., Lipman, T., . . . Tabori, U. Methylation of the TERT promoter and risk stratification of childhood brain tumours: an integrative genomic and molecular study. *Lancet Oncol* 2013; 14(6):534-542.

Cerone, M.A., Burgess, D.J., Naceur-Lombardelli, C., Lord, C.J. and Ashworth, A. High-throughput RNAi screening reveals novel regulators of telomerase. *Cancer Res* 2011; 71(9):3328-3340.

Cesare, A.J. and Griffith, J.D. Telomeric DNA in ALT cells is characterized by free telomeric circles and heterogeneous t-loops. *Mol Cell Biol* 2004; 24(22):9948-9957.

Cesare, A.J. and Reddel, R.R. Alternative lengthening of telomeres: models, mechanisms and implications. *Nat Rev Genet* 2010; 11(5):319-330.

Chebel, A. and Ffrench, M. Transcriptional regulation of the human telomerase reverse transcriptase: new insights. *Transcription* 2010; 1(1):27-31.

Chen, L., Wu, G. and Ji, H. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics* 2011; 27(10):1447-1448.

Chen, P.L., Riley, D.J., Chen, Y. and Lee, W.H. Retinoblastoma protein positively regulates terminal adipocyte differentiation through direct interaction with C/EBPs. *Genes Dev* 1996; 10(21):2794-2804.

Chen, Q., Ijpma, A. and Greider, C.W. Two survivor pathways that allow growth in the absence of telomerase are generated by distinct telomere recombination events. *Mol Cell Biol* 2001; 21(5):1819-1827.

Chen, Y.J., Campbell, H.G., Wiles, A.K., Eccles, M.R., Reddel, R.R., Braithwaite, A.W. and Royds, J.A. PAX8 regulates telomerase reverse transcriptase and telomerase RNA component in glioma. *Cancer Res* 2008; 68(14):5724-5732.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., . . . Gerstein, M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research* 2012; 22(9):1658-1667.

Chi, M.H. and Shore, D. SUM1-1, a dominant suppressor of SIR mutations in Saccharomyces cerevisiae, increases transcriptional silencing at telomeres and HM mating-type loci and decreases chromosome stability. *Molecular and cellular biology* 1996; 16(8):4281-4294.

Chiappetta, G., Avantaggiato, V., Visconti, R., Fedele, M., Battista, S., Trapasso, F., . . . Fusco, A. High level expression of the HMGI (Y) gene during embryonic development. *Oncogene* 1996; 13(11):2439-2446.

Chiappori, A.A., Kolevska, T., Spigel, D.R., Hager, S., Rarick, M., Gadgeel, S., . . . Schiller, J.H. A randomized phase II study of the telomerase inhibitor imetelstat as maintenance therapy for advanced non-small-cell lung cancer. *Ann Oncol* 2015; 26(2):354-362.

Chin, L., Artandi, S.E., Shen, Q., Tam, A., Lee, S.L., Gottlieb, G.J., . . . DePinho, R.A. p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. *Cell* 1999; 97(4):527-538.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007; 3:140.

Chuang, H.Y., Rassenti, L., Salcedo, M., Licon, K., Kohlmann, A., Haferlach, T., . . . Kipps, T.J. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood* 2012; 120(13):2639-2649.

Chudasama, P., Mughal, S.S., Sanders, M.A., Hubschmann, D., Chung, I., Deeg, K.I., . . . Frohling, S. Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat Commun* 2018; 9(1):144.

Chung, I., Osterwald, S., Deeg, K.I. and Rippe, K. PML body meets telomere: the beginning of an ALTernate ending? *Nucleus* 2012; 3(3):263-275.

Clynes, D., Jelinska, C., Xella, B., Ayyub, H., Scott, C., Mitson, M., . . . Gibbons, R.J. Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nat Commun* 2015; 6:7538.

Cong, Y.S., Wen, J. and Bacchetti, S. The human telomerase catalytic subunit hTERT: organization of the gene and characterization of the promoter. *Hum Mol Genet* 1999; 8(1):137-142.

Consortium, E.P., Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., . . . Birney, E. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489(7414):57-74.

Consortium, F., Suzuki, H., Forrest, A.R., van Nimwegen, E., Daub, C.O., Balwierz, P.J., . . . Hayashizaki, Y. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics* 2009; 41(5):553-562.

Cooperberg, M.R., Broering, J.M. and Carroll, P.R. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst* 2009; 101(12):878-887.

Crowe, D.L., Nguyen, D.C., Tsang, K.J. and Kyo, S. E2F-1 represses transcription of the human telomerase reverse transcriptase gene. *Nucleic Acids Res* 2001; 29(13):2789-2794.

d'Adda di Fagagna, F., Reaper, P.M., Clay-Farrace, L., Fiegler, H., Carr, P., Von Zglinicki, T., . . . Jackson, S.P. A DNA damage checkpoint response in telomere-initiated senescence. *Nature* 2003; 426(6963):194-198.

Dagg, R.A., Pickett, H.A., Neumann, A.A., Napier, C.E., Henson, J.D., Teber, E.T., . . . Reddel, R.R. Extensive Proliferation of Human Cancer Cells with Ever-Shorter Telomeres. *Cell Rep* 2017; 19(12):2544-2556.

de Lange, T. T-loops and the origin of telomeres. *Nat Rev Mol Cell Biol* 2004; 5(4):323-329.

de Lange, T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* 2005; 19(18):2100-2110.

de Lange, T. How telomeres solve the end-protection problem. *Science* 2009; 326(5955):948-952.

De Vitis, M., Berardinelli, F. and Sgura, A. Telomere Length Maintenance in Cancer: At the Crossroad between Telomerase and Alternative Lengthening of Telomeres (ALT). *Int J Mol Sci* 2018; 19(2).

Deeg, K.I., Chung, I., Bauer, C. and Rippe, K. Cancer Cells with Alternative Lengthening of Telomeres Do Not Display a General Hypersensitivity to ATR Inhibition. *Front Oncol* 2016; 6:186.

Deeg, K.I., Chung, I., Poos, A.M., Braun, D.M., Korshunov, A., Oswald, M., . . . Rippe, K. Dissecting telomere maintenance mechanisms in pediatric glioblastoma. 2017:Preprint: *bioRxiv* 129106.

Devereux, T.R., Horikawa, I., Anna, C.H., Annab, L.A., Afshari, C.A. and Barrett, J.C. DNA methylation analysis of the promoter region of the human telomerase reverse transcriptase (hTERT) gene. *Cancer Res* 1999; 59(24):6087-6090.

Ding, H., Douglass, E.F., Jr., Sonabend, A.M., Mela, A., Bose, S., Gonzalez, C., . . . Califano, A. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat Commun* 2018; 9(1):1471.

Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., . . . Weng, Z. Modeling gene expression using chromatin features in various cellular contexts. *Genome biology* 2012; 13(9):R53.

Doyle, K.R., Mitchell, M.A., Roberts, C.L., James, S., Johnson, J.E., Zhou, Y., . . . Broccoli, D. Validating a gene expression signature proposed to differentiate liposarcomas that use different telomere maintenance mechanisms. *Oncogene* 2012; 31(2):265-266; author reply 267-268.

Dror, V. and Winston, F. The Swi/Snf chromatin remodeling complex is required for ribosomal DNA and telomeric silencing in Saccharomyces cerevisiae. *Mol Cell Biol* 2004; 24(18):8227-8235.

Ducrest, A.L., Amacker, M., Mathieu, Y.D., Cuthbert, A.P., Trott, D.A., Newbold, R.F., . . . Lingner, J. Regulation of human telomerase activity: repression by normal chromosome 3 abolishes nuclear telomerase reverse transcriptase transcripts but does not affect c-Myc activity. *Cancer Res* 2001; 61(20):7594-7602.

Dunham, M.A., Neumann, A.A., Fasching, C.L. and Reddel, R.R. Telomere maintenance by recombination in human cells. *Nat Genet* 2000; 26(4):447-450.

Eldholm, V., Haugen, A. and Zienolddiny, S. CTCF mediates the TERT enhancer-promoter interactions in lung cancer cells: identification of a novel enhancer region involved in the regulation of TERT gene. *Int J Cancer* 2014; 134(10):2305-2313.

Elkak, A., Mokbel, R., Wilson, C., Jiang, W.G., Newbold, R.F. and Mokbel, K. hTERT mRNA expression is associated with a poor clinical outcome in human breast cancer. *Anticancer Res* 2006; 26(6C):4901-4904.

Epstein, J.I., Feng, Z., Trock, B.J. and Pierorazio, P.M. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. *Eur. Urol.* 2012; 61(5):1019-1024.

Feng, J., Funk, W.D., Wang, S.S., Weinrich, S.L., Avilion, A.A., Chiu, C.P., . . . et al. The RNA component of human telomerase. *Science* 1995; 269(5228):1236-1241.

Feuerbach, L., Sieverling, L., Deeg, K., Ginsbach, P., Hutter, B., Buchhalter, I., . . . Brors, B. TelomereHunter: telomere content estimation and characterization from whole genome sequencing data. 2016:Preprint: *bioRxiv* 065532. doi: 065510.061101/065532.

Filtz, T.M., Vogel, W.K. and Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci* 2014; 35(2):76-85.

Flynn, R.L., Cox, K.E., Jeitany, M., Wakimoto, H., Bryll, A.R., Ganem, N.J., . . . Zou, L. Alternative lengthening of telomeres renders cancer cells hypersensitive to ATR inhibitors. *Science* 2015; 347(6219):273-277.

Fredriksson, N.J., Ny, L., Nilsson, J.A. and Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* 2014; 46(12):1258-1263.

Frietze, S. and Farnham, P.J. Transcription factor effector domains. *Subcell Biochem* 2011; 52:261-277.

Frink, R.E., Peyton, M., Schiller, J.H., Gazdar, A.F., Shay, J.W. and Minna, J.D. Telomerase inhibitor imetelstat has preclinical activity across the spectrum of non-small cell lung cancer oncogenotypes in a telomere length dependent manner. *Oncotarget* 2016; 7(22):31639-31651.

Frohlich, H. biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics* 2015; 31(20):3290-3298.

Fu, W., Ergun, A., Lu, T., Hill, J.A., Haxhinasto, S., Fassett, M.S., . . . Benoist, C. A multiply redundant genetic switch 'locks in' the transcriptional signature of regulatory T cells. *Nature immunology* 2012; 13(10):972-980.

Fujiwara, C., Muramatsu, Y., Nishii, M., Tokunaka, K., Tahara, H., Ueno, M., . . . Seimiya, H. Cell-based chemical fingerprinting identifies telomeres and lamin A as modifiers of DNA damage response in cancer cells. *Sci Rep* 2018; 8(1):14827.

Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C. and Sladek, R. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* 2009; 10(3):R29.

Gaspar, T.B., Sa, A., Lopes, J.M., Sobrinho-Simoes, M., Soares, P. and Vinagre, J. Telomere Maintenance Mechanisms in Cancer. *Genes (Basel)* 2018; 9(5).

Gatbonton, T., Imbesi, M., Nelson, M., Akey, J.M., Ruderfer, D.M., Kruglyak, L., . . . Bedalov, A. Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet* 2006; 2(3):e35.

Gertz, J., Reddy, T.E., Varley, K.E., Garabedian, M.J. and Myers, R.M. Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Res* 2012; 22(11):2153-2162.

Gilheeney, S.W. and Kieran, M.W. Differences in molecular genetics between pediatric and adult malignant astrocytomas: age matters. *Future Oncol* 2012; 8(5):549-558.

Grassi, E., Zapparoli, E., Molineris, I. and Provero, P. Total Binding Affinity Profiles of Regulatory Regions Predict Transcription Factor Binding and Gene Expression in Human Cells. *PLoS One* 2015; 10(11):e0143627.

Grasso, C.S., Tang, Y., Truffaux, N., Berlow, N.E., Liu, L., Debily, M.A., . . . Monje, M. Functionally defined therapeutic targets in diffuse intrinsic pontine glioma. *Nat Med* 2015; 21(6):555-559.

Greider, C.W. and Blackburn, E.H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* 1985; 43(2 Pt 1):405-413.

Griewank, K.G., Murali, R., Puig-Butille, J.A., Schilling, B., Livingstone, E., Potrony, M., . . . Schadendorf, D. TERT promoter mutation status as an independent prognostic factor in cutaneous melanoma. *J Natl Cancer Inst* 2014; 106(9).

Gunes, C. and Rudolph, K.L. The role of telomeres in stem cells and cancer. *Cell* 2013; 152(3):390-393.

Gunkel, M., Chung, I., Worz, S., Deeg, K.I., Simon, R., Sauter, G., . . . Rippe, K. Quantification of telomere features in tumor tissue sections by an automated 3D imaging-based workflow. *Methods* 2017; 114:60-73.

Gurobi Optimization, L. Gurobi Optimizer Reference Manual. In.; 2018.

Hakin-Smith, V., Jellinek, D.A., Levy, D., Carroll, T., Teo, M., Timperley, W.R., . . . Royds, J.A. Alternative lengthening of telomeres and survival in patients with glioblastoma multiforme. *Lancet* 2003; 361(9360):836-838.

Hanahan, D. and Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* 2011; 144(5):646-674.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., . . . Young, R.A. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004; 431(7004):99-104.

Hayflick, L. The Limited in Vitro Lifetime of Human Diploid Cell Strains. *Exp Cell Res* 1965; 37:614-636.

Hayflick, L. and Moorhead, P.S. The serial cultivation of human diploid cell strains. *Exp Cell Res* 1961; 25:585-621.

Heaphy, C.M., de Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., . . . Meeker, A.K. Altered telomeres in tumors with ATRX and DAXX mutations. *Science* 2011; 333(6041):425.

Heaphy, C.M., Subhawong, A.P., Hong, S.M., Goggins, M.G., Montgomery, E.A., Gabrielson, E., . . . Meeker, A.K. Prevalence of the alternative lengthening of telomeres telomere maintenance mechanism in human cancer subtypes. *The American journal of pathology* 2011; 179(4):1608-1615.

Helpap, B. and Egevad, L. The significance of modified Gleason grading of prostatic carcinoma in biopsy and radical prostatectomy specimens. *Virchows Arch* 2006; 449(6):622-627.

Hemann, M.T., Strong, M.A., Hao, L.Y. and Greider, C.W. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell* 2001; 107(1):67-77.

Henson, J.D., Cao, Y., Huschtscha, L.I., Chang, A.C., Au, A.Y., Pickett, H.A. and Reddel, R.R. DNA C-circles are specific and quantifiable markers of alternative-lengthening-of-telomeres activity. *Nat Biotechnol* 2009; 27(12):1181-1185.

Henson, J.D., Lau, L.M., Koch, S., Martin La Rotta, N., Dagg, R.A. and Reddel, R.R. The C-Circle Assay for alternative-lengthening-of-telomeres activity. *Methods* 2017; 114:74-84.

Henson, J.D. and Reddel, R.R. Assaying and investigating Alternative Lengthening of Telomeres activity in human cells and cancers. *FEBS Lett* 2010; 584(17):3800-3811.

Höflmayer, D., Steinhoff, A., Hube-Magg, C., Simon, R., Kluth, M., Burandt, E., . . . Schroeder, C. High expression of CCCTC-binding factor (CTCF) is linked to poor prognosis in prostate cancer. *Molecular Oncology* 2019; *submitted*.

Horikawa, I., Cable, P.L., Afshari, C. and Barrett, J.C. Cloning and characterization of the promoter region of human telomerase reverse transcriptase gene. *Cancer Res* 1999; 59(4):826-830.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., . . . Kumar, R. TERT promoter mutations in familial and sporadic melanoma. *Science* 2013; 339(6122):959-961.

Hu, H., Zhang, Y., Zou, M., Yang, S. and Liang, X.Q. Expression of TRF1, TRF2, TIN2, TERT, KU70, and BRCA1 proteins is associated with telomere shortening and may contribute to multistage carcinogenesis of gastric cancer. *J Cancer Res Clin Oncol* 2010; 136(9):1407-1414.

Hu, J., Hwang, S.S., Liesa, M., Gan, B., Sahin, E., Jaskelioff, M., . . . DePinho, R.A. Antitelomerase therapy provokes ALT and mitochondrial adaptive mechanisms in cancer. *Cell* 2012; 148(4):651-663.

Hu, Z., Killion, P.J. and Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics* 2007; 39(5):683-687.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L. and Garraway, L.A. Highly recurrent TERT promoter mutations in human melanoma. *Science* 2013; 339(6122):957-959.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002; 18 Suppl 1:S96-104.

Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002; 18 Suppl 1:S233-240.

International Cancer Genome Consortium PedBrain Tumor Project. Recurrent MET fusion genes represent a drug target in pediatric glioblastoma. *Nat Med* 2016; 22(11):1314-1320.

Jiang, P., Freedman, M.L., Liu, J.S. and Liu, X.S. Inference of transcriptional regulation in cancers. *Proc Natl Acad Sci U S A* 2015; 112(25):7731-7736.

Johnson, P.F., Landschulz, W.H., Graves, B.J. and McKnight, S.L. Identification of a rat liver nuclear protein that binds to the enhancer core element of three animal viruses. *Genes Dev* 1987; 1(2):133-146.

Kadosh, D. and Struhl, K. Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. *Cell* 1997; 89(3):365-371.

Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; 28(1):27-30.

Kaul, Z., Cesare, A.J., Huschtscha, L.I., Neumann, A.A. and Reddel, R.R. Five dysfunctional telomeres predict onset of senescence in human cells. *EMBO Rep* 2011; 13(1):52-59.

Kelland, L.R. Overcoming the immortality of tumour cells by telomere and telomerase based cancer therapeutics--current status and future prospects. *Eur J Cancer* 2005; 41(7):971-979.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., . . . Mathelier, A. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* 2018; 46(D1):D260-D266.

Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Jr., . . . Yan, H. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A* 2013; 110(15):6021-6026.

Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L., . . . Shay, J.W. Specific association of human telomerase activity with immortal cells and cancer. *Science* 1994; 266(5193):2011-2015.

Kim, T.H., Kim, Y.E., Ahn, S., Kim, J.Y., Ki, C.S., Oh, Y.L., . . . Chung, J.H. TERT promoter mutations and long-term survival in patients with thyroid cancer. *Endocr Relat Cancer* 2016; 23(10):813-823.

Knosel, T., Chen, Y., Hotovy, S., Settmacher, U., Altendorf-Hofmann, A. and Petersen, I. Loss of desmocollin 1-3 and homeobox genes PITX1 and CDX2 are associated with tumor progression and survival in colorectal carcinoma. *Int J Colorectal Dis* 2012; 27(11):1391-1399.

Kolfschoten, I.G., van Leeuwen, B., Berns, K., Mullenders, J., Beijersbergen, R.L., Bernards, R., . . . Agami, R. A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity. *Cell* 2005; 121(6):849-858.

Kong, G., Liu, Z., Wu, K., Zhang, Y., Deng, Z., Feng, W., . . . Wang, H. Strong expression of paired-like homeodomain transcription factor 1 (PITX1) is associated with a favorable outcome in human osteosarcoma. *Tumour Biol* 2015; 36(10):7735-7741.

Koo, K.C., Park, S.U., Kim, K.H., Rha, K.H., Hong, S.J., Yang, S.C. and Chung, B.H. Predictors of survival in prostate cancer patients with bone metastasis and extremely high prostate-specific antigen levels. *Prostate Int* 2015; 3(1):10-15.

Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schopflin, R., Esposito, A., . . . Andrey, G. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat Genet* 2018; 50(10):1463-1473.

Kranz, A.L., Eils, R. and Konig, R. Enhancers regulate progression of development in mammalian cells. *Nucleic Acids Res* 2011; 39(20):8689-8702.

Kumar, M., Witt, B., Knippschild, U., Koch, S., Meena, J.K., Heinlein, C., . . . Gunes, C. CEBP factors regulate telomerase reverse transcriptase promoter activity in whey acidic protein-T mice during mammary carcinogenesis. *Int J Cancer* 2013; 132(9):2032-2043.

Kupiec, M. Biology of telomeres: lessons from budding yeast. *FEMS Microbiology Reviews* 2014; 38(2):144-171.

Lachmann, A., Giorgi, F.M., Lopez, G. and Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 2016; 32(14):2233-2235.

Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R. and Ma'ayan, A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 2010; 26(19):2438-2444.

Lafferty-Whyte, K., Cairney, C.J., Will, M.B., Serakinci, N., Daidone, M.G., Zaffaroni, N., . . . Keith, W.N. A gene expression signature classifying telomerase and ALT immortalization reveals an hTERT regulatory network and suggests a mesenchymal stem cell origin for ALT. *Oncogene* 2009; 28(43):3765-3774.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., . . . Weirauch, M.T. The Human Transcription Factors. *Cell* 2018; 175(2):598-599.

Landschulz, W.H., Johnson, P.F., Adashi, E.Y., Graves, B.J. and McKnight, S.L. Isolation of a recombinant copy of the gene encoding C/EBP. *Genes Dev* 1988; 2(7):786-800.

Lardenois, A., Stuparevic, I., Liu, Y., Law, M.J., Becker, E., Smagulova, F., . . . Primig, M. The conserved histone deacetylase Rpd3 and its DNA binding subunit Ume6 control dynamic transcript architecture during mitotic growth and meiotic development. *Nucleic Acids Res* 2015; 43(1):115-128.

Lee, M., Teber, E.T., Holmes, O., Nones, K., Patch, A.M., Dagg, R.A., . . . Pickett, H.A. Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Res* 2018; 46(10):4903-4918.

Lee, S.H., Kim, J.W., Lee, H.W., Cho, Y.S., Oh, S.H., Kim, Y.J., . . . Lee, J.H. Interferon regulatory factor-1 (IRF-1) is a mediator for interferon-gamma induced attenuation of telomerase activity and human telomerase reverse transcriptase (hTERT) expression. *Oncogene* 2003; 22(3):381-391.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., . . . Young, R.A. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002; 298(5594):799-804.

Levy, M.Z., Allsopp, R.C., Futcher, A.B., Greider, C.W. and Harley, C.B. Telomere end-replication problem and cell aging. *J Mol Biol* 1992; 225(4):951-960.

Li, B. and Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011; 12:323.

Li, M., Valsakumar, V., Poorey, K., Bekiranov, S. and Smith, J.S. Genome-wide analysis of functional sirtuin chromatin targets in yeast. *Genome biology* 2013; 14(5):R48.

Li, Y., Liang, M. and Zhang, Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol* 2014; 10(10):e1003908.

Li, Y., Zhou, Q.L., Sun, W., Chandrasekharan, P., Cheng, H.S., Ying, Z., . . . Tergaonkar, V. Non-canonical NF-kappaB signalling and ETS1/2 cooperatively drive C250T mutant TERT promoter activation. *Nat Cell Biol* 2015; 17(10):1327-1338.

Liu, D.X. and Lobie, P.E. Transcriptional activation of p53 by Pitx1. *Cell Death Differ* 2007; 14(11):1893-1907.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; 15(12):550.

Lovejoy, C.A., Li, W., Reisenweber, S., Thongthip, S., Bruno, J., de Lange, T., . . . Consortium, A.L.T.S.C. Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. *PLoS Genet* 2012; 8(7):e1002772.

Luke-Glaser, S., Poschke, H. and Luke, B. Getting in (and out of) the loop: regulating higher order telomere structures. *Frontiers in oncology* 2012; 2:180.

Lundberg, G., Sehic, D., Lansberg, J.K., Ora, I., Frigyesi, A., Castel, V., . . . Gisselsson, D. Alternative lengthening of telomeres--an enhanced chromosomal instability in aggressive non-MYCN amplified and telomere elongated neuroblastomas. *Genes Chromosomes Cancer* 2011; 50(4):250-262.

Lundblad, V. Telomere maintenance without telomerase. *Oncogene* 2002; 21(4):522-531.

Mallm, J.P., Iskar, M., Ishaque, N., Klett, L.C., Kugler, S.J., Muino, J.M., . . . Rippe, K. Linking aberrant chromatin features in chronic lymphocytic leukemia to deregulated transcription factor networks. *Mol Syst Biol* 2019:under revision.

Mancini, A., Xavier-Magalhaes, A., Woods, W.S., Nguyen, K.T., Amen, A.M., Hayes, J.L., . . . Costello, J.F. Disruption of the beta1L Isoform of GABP Reverses Glioblastoma Replicative Immortality in a TERT Promoter Mutation-Dependent Manner. *Cancer Cell* 2018; 34(3):513-528 e518.

Mangan, S. and Alon, U. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* 2003; 100(21):11980-11985.

Mangan, S., Itzkovitz, S., Zaslaver, A. and Alon, U. The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. *Journal of molecular biology* 2006; 356(5):1073-1081.

Mangan, S., Zaslaver, A. and Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of molecular biology* 2003; 334(2):197-204.

Mangerel, J., Price, A., Castelo-Branco, P., Brzezinski, J., Buczkowicz, P., Rakopoulos, P., . . . Tabori, U. Alternative lengthening of telomeres is enriched in, and impacts

survival of TP53 mutant pediatric malignant brain tumors. *Acta Neuropathologica* 2014; 128(6):853-862.

Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R. and Califano, A. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* 2008; 4:169.

Marcand, S., Gilson, E. and Shore, D. A protein-counting mechanism for telomere length regulation in yeast. *Science* 1997; 275(5302):986-990.

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

Martinez, P. and Blasco, M.A. Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nat Rev Cancer* 2011; 11(3):161-176.

Martinez, P. and Blasco, M.A. Replicating through telomeres: a means to an end. *Trends Biochem Sci* 2015; 40(9):504-515.

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

McCord, R., Pierce, M., Xie, J., Wonkatal, S., Mickel, C. and Vershon, A.K. Rfm1, a novel tethering factor required to recruit the Hst1 histone deacetylase for repression of middle sporulation genes. *Molecular and cellular biology* 2003; 23(6):2009-2016.

McDonald, K.L., McDonnell, J., Muntoni, A., Henson, J.D., Hegi, M.E., von Deimling, A., . . . Royds, J.A. Presence of alternative lengthening of telomeres mechanism in patients with glioblastoma identifies a less aggressive tumor type with longer survival. *J Neuropathol Exp Neurol* 2010; 69(7):729-736.

Mender, I. and Shay, J.W. Telomerase Repeated Amplification Protocol (TRAP). *Bio Protoc* 2015; 5(22).

Mender, I. and Shay, J.W. Telomere Restriction Fragment (TRF) Analysis. *Bio Protoc* 2015; 5(22).

Meng, F.L., Hu, Y., Shen, N., Tong, X.J., Wang, J., Ding, J. and Zhou, J.Q. Sua5p a single-stranded telomeric DNA-binding protein facilitates telomere replication. *EMBO J* 2009; 28(10):1466-1478.

Min, J., Wright, W.E. and Shay, J.W. Alternative lengthening of telomeres can be maintained by preferential elongation of lagging strands. *Nucleic Acids Res* 2017; 45(5):2615-2628.

Minner, S., Lutz, J., Hube-Magg, C., Kluth, M., Simon, R., Hoflmayer, D., . . . Schroeder, C. Loss of CCAAT-enhancer-binding protein alpha (CEBPA) is linked to poor prognosis in PTEN deleted and TMPRSS2:ERG fusion type prostate cancers. *Prostate* 2019; 79(3):302-311.

Moehren, U., Papaioannou, M., Reeb, C.A., Grasselli, A., Nanni, S., Asim, M., . . . Baniahmad, A. Wild-type but not mutant androgen receptor inhibits expression of the hTERT telomerase subunit: a novel role of AR mutation for prostate cancer development. *FASEB J* 2008; 22(4):1258-1267.

Molineris, I., Grassi, E., Ala, U., Di Cunto, F. and Provero, P. Evolution of promoter affinity for transcription factors in the human lineage. *Mol Biol Evol* 2011; 28(8):2173-2183.

Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., . . . Wu, J.R. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the

telomeres of human chromosomes. *Proc Natl Acad Sci U S A* 1988; 85(18):6622-6626.

Mozdy, A.D., Podell, E.R. and Cech, T.R. Multiple yeast genes, including Paf1 complex genes, affect telomere length via telomerase RNA abundance. *Molecular and cellular biology* 2008; 28(12):4152-4161.

Mudie, S., Bandarra, D., Batie, M., Biddlestone, J., Moniz, S., Ortmann, B., . . . Rocha, S. PITX1, a specificity determinant in the HIF-1alpha-mediated transcriptional response to hypoxia. *Cell Cycle* 2014; 13(24):3878-3891.

Nabetani, A. and Ishikawa, F. Alternative lengthening of telomeres pathway: recombination-mediated telomere maintenance mechanism in human cells. *J Biochem* 2011; 149(1):5-14.

Nakabayashi, M., Osaki, M., Kodani, I., Okada, F., Ryoke, K., Oshimura, M., . . . Kugoh, H. PITX1 is a reliable biomarker for predicting prognosis in patients with oral epithelial dysplasia. *Oncol Lett* 2014; 7(3):750-754.

Napier, C.E., Huschtscha, L.I., Harvey, A., Bower, K., Noble, J.R., Hendrickson, E.A. and Reddel, R.R. ATRX represses alternative lengthening of telomeres. *Oncotarget* 2015; 6(18):16543-16558.

Newman, M.E. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 2006; 103(23):8577-8582.

Ng, L.J., Cropley, J.E., Pickett, H.A., Reddel, R.R. and Suter, C.M. Telomerase activity is associated with an increase in DNA methylation at the proximal subtelomere and a reduction in telomeric transcription. *Nucleic Acids Res* 2009; 37(4):1152-1159.

O'Reilly, M., Teichmann, S.A. and Rhodes, D. Telomerases. *Curr Opin Struct Biol* 1999; 9(1):56-65.

O'Sullivan, R.J., Arnoult, N., Lackner, D.H., Oganesian, L., Haggblom, C., Corpet, A., . . . Karlseder, J. Rapid induction of alternative lengthening of telomeres by depletion of the histone chaperone ASF1. *Nat Struct Mol Biol* 2014; 21(2):167-174.

O'Sullivan, R.J. and Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat Rev Mol Cell Biol* 2010; 11(3):171-181.

Ohira, T., Naohiro, S., Nakayama, Y., Osaki, M., Okada, F., Oshimura, M. and Kugoh, H. miR-19b regulates hTERT mRNA expression through targeting PITX1 mRNA in melanoma cells. *Sci Rep* 2015; 5:8201.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 2001; 17(9):520-527.

Okamoto, K. and Seimiya, H. Revisiting Telomere Shortening in Cancer. *Cells* 2019; 8(2).

Oliveira, A.P., Patil, K.R. and Nielsen, J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC systems biology* 2008; 2:17.

Ong, C.T. and Corces, V.G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014; 15(4):234-246.

Orth, J.D., Thiele, I. and Palsson, B.O. What is flux balance analysis? *Nat Biotechnol* 2010; 28(3):245-248.

Osaki, M., Chinen, H., Yoshida, Y., Ohhira, T., Sunamura, N., Yamamoto, O., . . . Kugoh, H. Decreased PITX1 gene expression in human cutaneous malignant melanoma and its clinicopathological significance. *Eur J Dermatol* 2013; 23(3):344-349.

Osterhage, J.L., Talley, J.M. and Friedman, K.L. Proteasome-dependent degradation of Est1p regulates the cell cycle-restricted assembly of telomerase in Saccharomyces cerevisiae. *Nat Struct Mol Biol* 2006; 13(8):720-728.

Osterwald, S., Deeg, K.I., Chung, I., Parisotto, D., Worz, S., Rohr, K., . . . Rippe, K. PML induces compaction, TRF2 depletion and DNA damage signaling at telomeres and promotes their alternative lengthening. *J Cell Sci* 2015; 128(10):1887-1900.

Otsubo, T., Yamada, K., Hagiwara, T., Oshima, K., Iida, K., Nishikata, K., . . . Kawamura, Y.I. DNA hypermethyation and silencing of PITX1 correlated with advanced stage and poor postoperative prognosis of esophageal squamous cell carcinoma. *Oncotarget* 2017; 8(48):84434-84448.

Palm, W. and de Lange, T. How shelterin protects mammalian telomeres. *Annu Rev Genet* 2008; 42:301-334.

Pascolo, E., Wenz, C., Lingner, J., Hauel, N., Priepke, H., Kauffmann, I., . . . Schnapp, A. Mechanism of human telomerase inhibition by BIBR1532, a synthetic, non-nucleosidic drug candidate. *Journal of Biological Chemistry* 2002; 277(18):15566-15572.

Peifer, M., Hertwig, F., Roels, F., Dreidax, D., Gartlgruber, M., Menon, R., . . . Fischer, M. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* 2015; 526(7575):700-704.

Penney, K.L., Stampfer, M.J., Jahn, J.L., Sinnott, J.A., Flavin, R., Rider, J.R., . . . Fiorentino, M. Gleason grade progression is uncommon. *Cancer Res* 2013; 73(16):5163-5168.

Perrem, K., Bryan, T.M., Englezou, A., Hackl, T., Moy, E.L. and Reddel, R.R. Repression of an alternative mechanism for lengthening of telomeres in somatic cell hybrids. *Oncogene* 1999; 18(22):3383-3390.

Pezaro, C., Woo, H.H. and Davis, I.D. Prostate cancer: measuring PSA. *Intern Med J* 2014; 44(5):433-440.

Pickrell, J.K., Gaffney, D.J., Gilad, Y. and Pritchard, J.K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* 2011; 27(15):2144-2146.

Poli, V. The role of C/EBP isoforms in the control of inflammatory and native immunity functions. *J Biol Chem* 1998; 273(45):29279-29282.

Poole, L.A., Zhao, R., Glick, G.G., Lovejoy, C.A., Eischen, C.M. and Cortez, D. SMARCAL1 maintains telomere integrity during DNA replication. *Proc Natl Acad Sci U S A* 2015; 112(48):14864-14869.

Poos, A.M., Kordass, T., Kolte, A., Ast, V., Oswald, M., Rippe, K. and Koenig, R. Modelling TERT regulation across 19 different cancer types based on the MIPRIP 2.0 gene regulatory network approach. *bioRxiv* 2019; 513259.

Poos, A.M., Maicher, A., Dieckmann, A.K., Oswald, M., Eils, R., Kupiec, M., . . . Konig, R. Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. *Nucleic Acids Res* 2016; 44(10):e93.

Poremba, C., Heine, B., Diallo, R., Heinecke, A., Wai, D., Schaefer, K.L., . . . Boecker, W. Telomerase as a prognostic marker in breast cancer: high-throughput tissue microarray analysis of hTERT and hTR. *J Pathol* 2002; 198(2):181-189.

Preto, A., Singhrao, S.K., Haughton, M.F., Kipling, D., Wynford-Thomas, D. and Jones, C.J. Telomere erosion triggers growth arrest but not cell death in human cancer cells retaining wild-type p53: implications for antitelomerase therapy. *Oncogene* 2004; 23(23):4136-4145.

Pryde, F.E. and Louis, E.J. Limitations of silencing at native yeast telomeres. *EMBO J* 1999; 18(9):2538-2550.

Qi, D.L., Ohhira, T., Fujisaki, C., Inoue, T., Ohta, T., Osaki, M., . . . Kugoh, H. Identification of PITX1 as a TERT suppressor gene located on human chromosome 5. *Mol Cell Biol* 2011; 31(8):1624-1636.

Qiao, F., Gong, P., Song, Y., Shen, X., Su, X., Li, Y., . . . Fan, H. Downregulated PITX1 Modulated by MiR-19a-3p Promotes Cell Malignancy and Predicts a Poor Prognosis of Gastric Cancer by Affecting Transcriptionally Activated PDCD5. *Cell Physiol Biochem* 2018; 46(6):2215-2231.

Ramlee, M.K., Wang, J., Toh, W.X. and Li, S. Transcription Regulation of the Human Telomerase Reverse Transcriptase (hTERT) Gene. *Genes (Basel)* 2016; 7(8).

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. and Vilo, J. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* 2016; 44(W1):W83-89.

Reimand, J., Vaquerizas, J.M., Todd, A.E., Vilo, J. and Luscombe, N.M. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res* 2010; 38(14):4768-4777.

Renaud, S., Loukinov, D., Abdullaev, Z., Guilleret, I., Bosman, F.T., Lobanenkov, V. and Benhattar, J. Dual role of DNA methylation inside and outside of CTCF-binding regions in the transcriptional regulation of the telomerase hTERT gene. *Nucleic Acids Res* 2007; 35(4):1245-1256.

Roth, A., Harley, C.B. and Baerlocher, G.M. Imetelstat (GRN163L)--telomerase-based cancer therapy. *Recent Results Cancer Res* 2010; 184:221-234.

Rouaud, F., Hamouda-Tekaya, N., Cerezo, M., Abbe, P., Zangari, J., Hofman, V., . . . Rocchi, S. E2F1 inhibition mediates cell death of metastatic melanoma. *Cell Death Dis* 2018; 9(5):527.

Rubin-Bejerano, I., Mandel, S., Robzyk, K. and Kassir, Y. Induction of meiosis in Saccharomyces cerevisiae depends on conversion of the transcriptional represssor Ume6 to a positive regulator by its regulated association with the transcriptional activator Ime1. *Mol Cell Biol* 1996; 16(5):2518-2526.

Ruden, M. and Puri, N. Novel anticancer therapeutics targeting telomerase. *Cancer Treat Rev* 2013; 39(5):444-456.

Rusinek, D., Pfeifer, A., Krajewska, J., Oczko-Wojciechowska, M., Handkiewicz-Junak, D., Pawlaczek, A., . . . Czarniecka, A. Coexistence of TERT Promoter Mutations and the BRAF V600E Alteration and Its Impact on Histopathological Features of Papillary Thyroid Carcinoma in a Selected Series of Polish Patients. *Int J Mol Sci* 2018; 19(9).

Saeboe-Larssen, S., Fossberg, E. and Gaudernack, G. Characterization of novel alternative splicing sites in human telomerase reverse transcriptase (hTERT): analysis of expression and mutual correlation in mRNA isoforms from normal and tumour tissues. *BMC Mol Biol* 2006; 7:26.

Sailer, V., Charpentier, A., Dietrich, J., Vogt, T.J., Franzen, A., Bootz, F., . . . Schroeck, A. Intragenic DNA methylation of PITX1 and the adjacent long non-coding RNA C5orf66-AS1 are prognostic biomarkers in patients with head and neck squamous cell carcinomas. *PLoS ONE* 2018; 13(2):e0192742.

Sandin, S. and Rhodes, D. Telomerase structure. *Curr Opin Struct Biol* 2014; 25:104-110.

Saraiva, J.P., Oswald, M., Biering, A., Roll, D., Assmann, C., Klassert, T., . . . Konig, R. Fungal biomarker discovery by integration of classifiers. *BMC Genomics* 2017; 18(1):601.

Sauter, G., Clauditz, T., Steurer, S., Wittmer, C., Buscheck, F., Krech, T., . . . Schlomm, T. Integrating Tertiary Gleason 5 Patterns into Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens. *Eur Urol* 2018; 73(5):674-683.

Schacht, T., Oswald, M., Eils, R., Eichmuller, S.B. and Konig, R. Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics* 2014; 30(17):i401-407.

Schneider, T.D. and Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990; 18(20):6097-6100.

Schoenmakers, E.F., Wanschura, S., Mols, R., Bullerdiek, J., Van den Berghe, H. and Van de Ven, W.J. Recurrent rearrangements in the high mobility group protein gene, HMGI-C, in benign mesenchymal tumours. *Nat Genet* 1995; 10(4):436-444.

Schramm, G., Wiesberg, S., Diessl, N., Kranz, A.L., Sagulenko, V., Oswald, M., . . . Konig, R. PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics* 2010; 26(9):1225-1231.

Schwartzentruber, J., Korshunov, A., Liu, X.Y., Jones, D.T., Pfaff, E., Jacob, K., . . . Jabado, N. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* 2012; 482(7384):226-231.

Setty, M., Helmy, K., Khan, A.A., Silber, J., Arvey, A., Neezen, F., . . . Leslie, C.S. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol* 2012; 8:605.

Shachar, R., Ungar, L., Kupiec, M., Ruppin, E. and Sharan, R. A systems-level approach to mapping the telomere length maintenance gene circuitry. *Mol Syst Biol* 2008; 4:172.

Shain, A.H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., . . . Bastian, B.C. The Genetic Evolution of Melanoma from Precursor Lesions. *N Engl J Med* 2015; 373(20):1926-1936.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., . . . Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; 13(11):2498-2504.

Shay, J.W. and Wright, W.E. Telomeres and telomerase: three decades of progress. *Nat Rev Genet* 2019.

Shore, D. and Bianchi, A. Telomere length regulation: coupling DNA end processing to feedback regulation of telomerase. *EMBO J* 2009; 28(16):2309-2322.

Sieverling, L., Hong, C., Koser, S.D., Ginsbach, P., Kleinheinz, K., Hutter, B., . . . Feuerbach, L. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat Commun, in press* 2019:Preprint: *bioRxiv* 157560.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., . . . Kasprzyk, A. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 2015; 43(W1):W589-598.

Sobinoff, A.P. and Pickett, H.A. Alternative Lengthening of Telomeres: DNA Repair Pathways Converge. *Trends Genet* 2017; 33(12):921-932.

Song, X., Zhao, C., Jiang, L., Lin, S., Bi, J., Wei, Q., . . . Wei, M. High PITX1 expression in lung adenocarcinoma patients is associated with DNA methylation and poor prognosis. *Pathol Res Pract* 2018.

Spitz, F. and Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 2012; 13(9):613-626.

Steurer, S., Mayer, P.S., Adam, M., Krohn, A., Koop, C., Ospina-Klinck, D., . . . Schlomm, T. TMPRSS2-ERG fusions are strongly linked to young patient age in low-grade prostate cancer. *Eur Urol* 2014; 66(6):978-981.

Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* 2000; 16(1):16-23.

Stormo, G.D. and Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* 2010; 11(11):751-760.

Sturm, D., Bender, S., Jones, D.T., Lichter, P., Grill, J., Becher, O., . . . Pfister, S.M. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nature Reviews Cancer* 2014; 14(2):92-107.

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.A., Jones, D.T., Konermann, C., . . . Pfister, S.M. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* 2012; 22(4):425-437.

Sun, Z.W. and Hampsey, M. A general requirement for the Sin3-Rpd3 histone deacetylase complex in regulating silencing in Saccharomyces cerevisiae. *Genetics* 1999; 152(3):921-932.

Taggart, A.K. and Zakian, V.A. Telomerase: what are the Est proteins doing? *Curr Opin Cell Biol* 2003; 15(3):275-280.

Tai, W.T., Chen, Y.L., Chu, P.Y., Chen, L.J., Hung, M.H., Shiau, C.W., . . . Chen, K.F. Protein tyrosine phosphatase 1B dephosphorylates PITX1 and regulates p120RasGAP in hepatocellular carcinoma. *Hepatology* 2016; 63(5):1528-1543.

Takai, H., Smogorzewska, A. and de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr Biol* 2003; 13(17):1549-1556.

Takakura, M., Kyo, S., Kanaya, T., Hirano, H., Takeda, J., Yutsudo, M. and Inoue, M. Cloning of human telomerase catalytic subunit (hTERT) gene promoter and identification of proximal core promoter sequences essential for transcriptional activation in immortalized and cancer cells. *Cancer Res* 1999; 59(3):551-557.

Takenobu, M., Osaki, M., Fujiwara, K., Fukuhara, T., Kitano, H., Kugoh, H. and Okada, F. PITX1 is a novel predictor of the response to chemotherapy in head and neck squamous cell carcinoma. *Mol Clin Oncol* 2016; 5(1):89-94.

Teixeira, M.T. Saccharomyces cerevisiae as a Model to Study Replicative Senescence Triggered by Telomere Shortening. *Frontiers in oncology* 2013; 3:101.

Teng, H.W., Hung, M.H., Chen, L.J., Chang, M.J., Hsieh, F.S., Tsai, M.H., . . . Chen, K.F. Protein tyrosine phosphatase 1B targets PITX1/p120RasGAP thus showing therapeutic potential in colorectal carcinoma. *Sci Rep* 2016; 6:35308.

Teng, S.C., Chang, J., McCowan, B. and Zakian, V.A. Telomerase-independent lengthening of yeast telomeres occurs by an abrupt Rad50p-dependent, Rif-inhibited recombinational process. *Mol Cell* 2000; 6(4):947-952.

Timchenko, N.A., Harris, T.E., Wilde, M., Bilyeu, T.A., Burgess-Beusse, B.L., Finegold, M.J. and Darlington, G.J. CCAAT/enhancer binding protein alpha regulates p21 protein and hepatocyte proliferation in newborn mice. *Mol Cell Biol* 1997; 17(12):7353-7361.

Tokutake, Y., Matsumoto, T., Watanabe, T., Maeda, S., Tahara, H., Sakamoto, S., . . . Furuichi, Y. Extra-chromosomal telomere repeat DNA in telomerase-negative immortalized cell lines. *Biochem Biophys Res Commun* 1998; 247(3):765-772.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., . . . Chinnaiyan, A.M. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; 310(5748):644-648.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., . . . Zhu, Z. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005; 23(1):137-144.

Trescher, S., Munchmeyer, J. and Leser, U. Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. *BMC Syst Biol* 2017; 11(1):41.

Ungar, L., Harari, Y., Toren, A. and Kupiec, M. Tor complex 1 controls telomere length by affecting the level of Ku. *Curr Biol* 2011; 21(24):2115-2120.

Ungar, L., Yosef, N., Sela, Y., Sharan, R., Ruppin, E. and Kupiec, M. A genome-wide screen for essential yeast genes that affect telomere length maintenance. *Nucleic Acids Res* 2009; 37(12):3840-3849.

Uziel, O., Yosef, N., Sharan, R., Ruppin, E., Kupiec, M., Kushnir, M., . . . Lahav, M. The effects of telomere shortening on cancer cells: a network model of proteomic and microRNA analysis. *Genomics* 2015; 105(1):5-16.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009; 10(4):252-263.

Vaziri, H., West, M.D., Allsopp, R.C., Davison, T.S., Wu, Y.S., Arrowsmith, C.H., . . . Benchimol, S. ATM-dependent telomere loss in aging human diploid fibroblasts and DNA damage lead to the post-translational activation of p53 protein involving poly(ADP-ribose) polymerase. *EMBO J* 1997; 16(19):6018-6033.

Villa, R., Folini, M., Lualdi, S., Veronese, S., Daidone, M.G. and Zaffaroni, N. Inhibition of telomerase activity by a cell-penetrating peptide nucleic acid construct in human melanoma cells. *FEBS Lett* 2000; 473(2):241-248.

Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., . . . Soares, P. Frequency of TERT promoter mutations in human cancers. *Nat Commun* 2013; 4:2185.

Wang, F., Podell, E.R., Zaug, A.J., Yang, Y., Baciu, P., Cech, T.R. and Lei, M. The POT1-TPP1 telomere complex is a telomerase processivity factor. *Nature* 2007; 445(7127):506-510.

Wang, H., Iakova, P., Wilde, M., Welm, A., Goode, T., Roesler, W.J. and Timchenko, N.A. C/EBPalpha arrests cell proliferation through direct inhibition of Cdk2 and Cdk4. *Mol Cell* 2001; 8(4):817-828.

Wang, L., Hurley, D.G., Watkins, W., Araki, H., Tamada, Y., Muthukaruppan, A., . . . Print, C.G. Cell cycle gene networks are associated with melanoma prognosis. *PLoS One* 2012; 7(4):e34247.

Wang, N.D., Finegold, M.J., Bradley, A., Ou, C.N., Abdelsayed, S.V., Wilde, M.D., . . . Darlington, G.J. Impaired energy homeostasis in C/EBP alpha knockout mice. *Science* 1995; 269(5227):1108-1112.

Wang, X., Lee, R.S., Alver, B.H., Haswell, J.R., Wang, S., Mieczkowski, J., . . . Roberts, C.W. SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. *Nat Genet* 2017; 49(2):289-295.

Wang, Y., Ou, Z., Sun, Y., Yeh, S., Wang, X., Long, J. and Chang, C. Androgen receptor promotes melanoma metastasis via altering the miRNA-539-3p/USP13/MITF/AXL signals. *Oncogene* 2017; 36(12):1644-1654.

Washburn, B.K. and Esposito, R.E. Identification of the Sin3-binding site in Ume6 defines a two-step process for conversion of Ume6 from a transcriptional repressor to an activator in yeast. *Mol Cell Biol* 2001; 21(6):2057-2069.

Weirauch, M.T. and Hughes, T.R. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell Biochem* 2011; 52:25-73.

Weischenfeldt, J., Simon, R., Feuerbach, L., Schlangen, K., Weichenhan, D., Minner, S., . . . Schlomm, T. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* 2013; 23(2):159-170.

Wellinger, R.J. and Zakian, V.A. Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: beginning to end. *Genetics* 2012; 191(4):1073-1105.

Wick, M., Zubov, D. and Hagen, G. Genomic organization and promoter characterization of the gene encoding the human telomerase reverse transcriptase (hTERT). *Gene* 1999; 232(1):97-106.

Wilkinson, A.C., Nakauchi, H. and Gottgens, B. Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. *Cell Syst* 2017; 5(4):319-331.

Williams, S.C. No end in sight for telomerase-targeted cancer drugs. *Nat Med* 2013; 19(1):6.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., . . . Davis, R.W. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* 1999; 285(5429):901-906.

Wright, J.H., Gottschling, D.E. and Zakian, V.A. Saccharomyces telomeres assume a non-nucleosomal chromatin structure. *Genes Dev* 1992; 6(2):197-210.

Wright, W.E., Brasiskyte, D., Piatyszek, M.A. and Shay, J.W. Experimental elongation of telomeres extends the lifespan of immortal x normal cell hybrids. *EMBO J* 1996; 15(7):1734-1741.

Wunderlich, Z. and Mirny, L.A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* 2009; 25(10):434-440.

Xin, H., Liu, D., Wan, M., Safari, A., Kim, H., Sun, W., . . . Songyang, Z. TPP1 is a homologue of ciliate TEBP-beta and interacts with POT1 to recruit telomerase. *Nature* 2007; 445(7127):559-562.

Yang, C.W., Tseng, S.F., Yu, C.J., Chung, C.Y., Chang, C.Y., Pobiega, S. and Teng, S.C. Telomere shortening triggers a feedback loop to enhance end protection. *Nucleic Acids Res* 2017; 45(14):8314-8328.

Yang, J.H., Li, J.H., Jiang, S., Zhou, H. and Qu, L.H. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res* 2013; 41(Database issue):D177-187.

Yeager, T.R., Neumann, A.A., Englezou, A., Huschtscha, L.I., Noble, J.R. and Reddel, R.R. Telomerase-negative immortalized human cells contain a novel type of promyelocytic leukemia (PML) body. *Cancer Res* 1999; 59(17):4175-4179.

Yin, H., Lowery, M. and Glass, J. In prostate cancer C/EBPalpha promotes cell growth by the loss of interactions with CDK2, CDK4, and E2F and by activation of AKT. *Prostate* 2009; 69(9):1001-1016.

Yin, H., Radomska, H.S., Tenen, D.G. and Glass, J. Down regulation of PSA by C/EBPalpha is associated with loss of AR expression and inhibition of PSA promoter activity in the LNCaP cell line. *BMC Cancer* 2006; 6:158.

Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppin, E. and Sharan, R. Toward accurate reconstruction of functional protein networks. *Mol Syst Biol* 2009; 5:248.

Yu, H. and Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* 2006; 103(40):14724-14731.

Zhang, J., Gonit, M., Salazar, M.D., Shatnawi, A., Shemshedini, L., Trumbly, R. and Ratnam, M. C/EBPalpha redirects androgen receptor signaling through a unique bimodal interaction. *Oncogene* 2010; 29(5):723-738.

Zhang, Q., Kim, N.K. and Feigon, J. Architecture of human telomerase RNA. *Proc Natl Acad Sci U S A* 2011; 108(51):20325-20332.

Zill, O.A. and Rine, J. Interspecies variation reveals a conserved repressor of alpha-specific genes in Saccharomyces yeasts. *Genes Dev* 2008; 22(12):1704-1716.

Appendix

Table S1. Corresponding genes of the investigated regulator (R) deletion strains of the dataset of Reimand and coworkers (Reimand *et al.*, 2010).

R	Group	R	Group	R	Group	R	Group
CDC73	short	***ARG81***	control	***GAT1***	control	***LEU3***	control
CST6	short	***ARG82***	control	***GAT3***	control	***MAC1***	control
GAL11	short	***ARO80***	control	***GCN4***	control	***MAL13***	control
HFI1	short	***ARR1***	control	***GCR1***	control	***MAL33***	control
HST1	short	***ASH1***	control	***GCR2***	control	***MBF1***	control
MOT2	short	***ASK10***	control	***GLN3***	control	***MBP1***	control
MOT3	short	***AZF1***	control	***GTS1***	control	***MCM1***	control
OPI1	short	***BAS1***	control	***GZF3***	control	***MDS3***	control
PGD1	short	***BDF2***	control	***HAA1***	control	***MET28***	control
RPN4	short	***CAC2***	control	***HAC1***	control	***MET31***	control
RSC2	short	***CAD1***	control	***HAL9***	control	***MET32***	control
RTF1	short	***CAF17***	control	***HAP2***	control	***MGA1***	control
SIF2	short	***CAF4***	control	***HAP3***	control	***MGA2***	control
SIN3	short	***CAT8***	control	***HAP4***	control	***MIG1***	control
SIR3	short	***CBF1***	control	***HAP5***	control	***MIG2***	control
SRB2	short	***CHA4***	control	***HAT1***	control	***MSI1***	control
SRB5	short	***CIN5***	control	***HAT2***	control	***MSN1***	control
SUM1	short	***CRZ1***	control	***HDA1***	control	***MSN2***	control
DIG1	long	***CSE2***	control	***HIR1***	control	***MSN4***	control
HCM1	long	***CUP2***	control	***HIR2***	control	***MSS11***	control
MET18	long	***CUP9***	control	***HIR3***	control	***MTH1***	control
NUT1	long	***DAL80***	control	***HMS1***	control	***NDT80***	control
RAP1	long	***DAL81***	control	***HMS2***	control	***NGG1***	control
REB1	long	***DAL82***	control	***HOG1***	control	***NOT3***	control
RIF1	long	***DAT1***	control	***HPA2***	control	***NRG1***	control
RIF2	long	***DOT5***	control	***HSF1***	control	***OAF1***	control
SRB8	long	***DOT6***	control	***HST3***	control	***PDR1***	control
SSN2	long	***ECM22***	control	***HST4***	control	***PDR3***	control
SSN3	long	***ESC2***	control	***IME1***	control	***PDR8***	control
ABF1	control	***FKH1***	control	***INO2***	control	***PHD1***	control
ACA1	control	***FKH2***	control	***INO4***	control	***PHO2***	control
ACE2	control	***FLO8***	control	***ISW1***	control	***PHO23***	control
ADA2	control	***FZF1***	control	***ISW2***	control	***PHO4***	control
ADR1	control	***GAL3***	control	***IXR1***	control	***PIB2***	control
AFT2	control	***GAL4***	control	***KAR4***	control	***PIP2***	control
ARG80	control	***GAL80***	control	***KSS1***	control	***POP2***	control

R	Group	R	Group	R	Group	R	Group
PPR1	control	***SIN4***	control	***SUT2***	control	***YDR520C***	control
PUT3	control	***SIP3***	control	***SWI3***	control	***YER028C***	control
RCS1	control	***SIP4***	control	***SWI4***	control	***YER051W***	control
RDR1	control	***SIR1***	control	***SWI5***	control	***YER130C***	control
RDS1	control	***SIR2***	control	***SWI6***	control	***YER184C***	control
RDS2	control	***SKN7***	control	***TAF14***	control	***YFL044C***	control
RFX1	control	***SKO1***	control	***TBS1***	control	***YFL052W***	control
RGM1	control	***SMK1***	control	***TEC1***	control	***YGL131C***	control
RGT1	control	***SMP1***	control	***THI2***	control	***YGR067C***	control
RIC1	control	***SNF1***	control	***TIS11***	control	***YGR089W***	control
RIM101	control	***SNF11***	control	***TOS8***	control	***YHP1***	control
RIS1	control	***SNF2***	control	***TUP1***	control	***YIL130W***	control
RLF2	control	***SNF5***	control	***TYE7***	control	***YJL103C***	control
RLM1	control	***SNF6***	control	***UGA3***	control	***YJL206C***	control
RLR1	control	***SOK2***	control	***UME1***	control	***YKL005C***	control
RME1	control	***SPS18***	control	***UME6***	control	***YKL222C***	control
ROX1	control	***SPT10***	control	***UPC2***	control	***YKR064W***	control
RPD3	control	***SPT2***	control	***WAR1***	control	***YLR278C***	control
RPH1	control	***SPT20***	control	***WTM1***	control	***YML081W***	control
RPI1	control	***SPT23***	control	***WTM2***	control	***YMR075W***	control
RSC1	control	***SPT3***	control	***XBP1***	control	***YNR063W***	control
RTG1	control	***SPT4***	control	***YAP1***	control	***YOX1***	control
RTG3	control	***STB1***	control	***YAP3***	control	***YPL230W***	control
RTT107	control	***STB2***	control	***YAP5***	control	***YPR022C***	control
SAS3	control	***STB3***	control	***YAP6***	control	***YPR196W***	control
SAS4	control	***STB4***	control	***YAP7***	control	***YRR1***	control
SAS5	control	***STB5***	control	***YBL054W***	control	***ZAP1***	control
SDS3	control	***STB6***	control	***YBR033W***	control	***ZDS1***	control
SEF1	control	***STP1***	control	***YBR239C***	control	***ZMS1***	control
SET2	control	***STP2***	control	***YDR026C***	control		
SFL1	control	***STP4***	control	***YDR049W***	control		
SFP1	control	***SUT1***	control	***YDR266C***	control		

The telomere phenotype was annotated from (Askree *et al.*, 2004; Ben-Shitrit *et al.*, 2012; Gatbonton *et al.*, 2006; Shachar *et al.*, 2008; Ungar *et al.*, 2009).

Table S2. Putative regulators of the *EST* genes (taken from YEASTRACT).

Regulator	*EST*s	Regulator	*EST*s
Msn4	*EST1, EST2, EST3*	**Mbp1**	*EST1*
Sfp1	*EST1, EST2, EST3*	**Mga1**	*EST3*
Ste12	*EST1, EST2, EST3*	**Mig1**	*EST1*
Abf1	*EST2, EST3*	**Mig3**	*EST3*
Ace2	*EST2, EST3*	**Nrg1**	*EST2*
Cst6	*EST1, EST2*	**Nrg2**	*EST2*
Fhl1	*EST1, EST2*	**Pdr1**	*EST2*
Gcn4	*EST1, EST2*	**Pdr3**	*EST2*
Gln3	*EST2, EST3*	**Rfx1**	*EST1*
Ixr1	*EST1, EST3*	**Rgt1**	*EST2*
Msn2	*EST2, EST3*	**Rme1**	*EST2*
Sas3	*EST2, EST3*	**Rsc1**	*EST2*
Sin3	*EST1, EST3*	**Rtg3**	*EST2*
Sin4	*EST1, EST3*	**Sko1**	*EST3*
Srb2	*EST1, EST3*	**Snf1**	*EST2*
Swi5	*EST2, EST3*	**Snf2**	*EST1*
Tec1	*EST1, EST2*	**Snf6**	*EST1*
Arg81	*EST2*	**Sok2**	*EST3*
Cbf1	*EST1*	**Spt10**	*EST3*
Cdc73	*EST2*	**Spt20**	*EST1*
Cin5	*EST3*	**Spt4**	*EST3*
Cse2	*EST3*	**Sum1**	*EST1*
Cup2	*EST1*	**Swi3**	*EST1*
Dig1	*EST3*	**Swi4**	*EST1*
Hap2	*EST3*	**Tup1**	*EST1*
Hir1	*EST3*	**Ume6**	*EST3*
Hsf1	*EST1*	**Yrm1**	*EST2*
Hst1	*EST1*		

Table S3. List of transcription factors putatively regulating *TERT*, from the generic human gene regulatory network.

TF	Edge strength score	TF	Edge strength score	TF	Edge strength score
AP-2	2.00	**HMGA2**	2.00	**PITX1**	2.00
AR	2.00	**HNRNPK**	2.00	**POLR2A**	0.50
BATF	0.25	**IKZF1**	2.00	**POU2F2**	0.25
BCL11A	0.25	**IRF1**	2.00	**RAD21**	0.50
BHLHE40	2.00	**JUND**	2.00	**RELA**	2.00
CEBPA	2.00	**KLF2**	2.00	**REST**	0.25
CTCF	0.50	**MAX**	2.50	**RUNX2**	2.00
CTCFL	2.00	**MAZ**	2.00	**SIN3A**	0.50
E2F1	2.00	**MEN1**	2.00	**SIN3AK20**	0.50
E2F2	2.00	**MITF**	2.00	**SMAD3**	2.00
E2F4	2.25	**MXD1**	2.00	**SMARCB1**	0.25
E2F5	2.00	**MXI1**	0.50	**SP3**	2.00
E2F6	2.25	**MYB**	2.00	**TAF1**	0.25
EGR1	2.50	**MYC**	3.75	**TAF9**	2.00
EPAS1	2.00	**MYCN**	2.00	**TAL1**	2.00
ESR1	2.00	**MZF1**	2.00	**TCF12**	0.25
ESR2	2.00	**NFAT5**	2.00	**TCF7**	2.00
ETS1	2.00	**NFATC2**	2.00	**TFAP2A**	2.00
ETS2	2.00	**NF.KB**	2.00	**TFAP2B**	2.00
GLI1	2.00	**NFKB1**	1.00	**TFAP2C**	2.00
GLI2	2.00	**NFKB.P50.P65**	2.00	**TFAP2D**	2.00
GRHL2	2.00	**NFX1**	2.00	**TP53**	2.00
HEY1	0.25	**NR2F2**	2.00	**TP73**	2.00
HIF.1	2.00	**PAX5**	2.00	**WT1**	2.00
HIF1A	2.00	**PAX8**	2.00	**ZBTB48**	2.00

Table S4. Specific *TERT* regulators of BLCA from the pan-cancer MIPRIP analysis.

TF	*p*-value
PAX8	1.92 E-14
AR	4.23 E-12
PAX5	5.33 E-11
E2F6	2.50 E-09
NF.KB	1.87 E-05
E2F4	2.54 E-04
SIN3A	3.10 E-04
BATF	4.11 E-03
GLI2	1.73 E-02
HIF.1	1.94 E-02
IKZF1	2.97 E-02
NFATC2	3.95 E-02
MYCN	3.95 E-02

Table S5. Specific *TERT* regulators of BRCA from the pan-cancer MIPRIP analysis.

TF	*p*-value
MYCN	1.17 E-17
TAF9	5.51 E-17
TCF12	5.47 E-13
E2F2	5.80 E-07
AR	5.49 E-06
SMARCB1	9.03 E-06
MYC	1.09 E-05
POU2F2	4.33 E-05
PAX8	9.28 E-05
MAX	3.79 E-04
TAF1	5.60 E-04
E2F4	1.42 E-03
BHLHE40	4.15 E-03
CTCF	5.62 E-03
ESR1	6.76 E-03
HIF1A	4.73 E-02

Table S6. Specific *TERT* regulators of CESC from the pan-cancer MIPRIP analysis.

TF	*p*-value
TAL1	2.76 E-15
BCL11A	6.74 E-08
HIF.1	6.74 E-08
MXI1	2.04 E-07
HIF1A	8.40 E-07
NF.KB	1.53 E-05
POLR2A	3.23 E-05
E2F5	3.77 E-05
ETS1	2.73 E-04
MXD1	3.93 E-04
TP73	8.11 E-04
PITX1	1.19 E-03
TAF9	1.74 E-03
BATF	2.12 E-03
AP-2	2.23 E-03
TCF12	6.16 E-03
TP53	4.13 E-02

Table S7. Specific *TERT* regulators of COADREAD from the pan-cancer MIPRIP analysis.

TF	*p*-value
RAD21	2.23 E-17
AR	2.23 E-17
BATF	4.61 E-13
SP3	8.36 E-11
TAF1	1.26 E-09
EPAS1	4.10 E-08
RUNX2	5.14 E-08
SMARCB1	2.58 E-07
HEY1	2.58 E-07
GLI1	3.07 E-06
CTCF	4.81 E-06
TCF12	2.57 E-05
E2F6	7.50 E-04
E2F4	8.24 E-03
PAX5	8.71 E-03

Table S8. Specific *TERT* regulators of ESCA from the pan-cancer MIPRIP analysis.

TF	*p*-value
MXD1	2.23 E-19
NR2F2	2.54 E-19
PAX8	4.35 E-12
GLI1	2.56 E-08
KLF2	3.75 E-08
GLI2	1.96 E-07
TFAP2A	3.52 E-05
ETS2	3.55 E-05
RUNX2	5.54 E-05
MAX	3.58 E-04
NF.KB	5.19 E-04
HMGA2	6.42 E-03
CEBPA	2.10 E-02

Table S9. Specific *TERT* regulators of GBM from the pan-cancer MIPRIP analysis.

TF	*p*-value
ETS2	9.21 E-26
MEN1	7.77 E-25
WT1	7.16 E-20
CTCFL	1.39 E-09
EGR1	1.55 E-07
HMGA2	1.31 E-06
SP3	1.33 E-05
ESR1	7.59 E-05
TP73	2.99 E-04
HIF.1	5.01 E-04
RUNX2	7.10 E-03
TFAP2D	3.69 E-02

Table S10. Specific *TERT* regulators of HNSC from the pan-cancer MIPRIP analysis.

TF	*p*-value
PITX1	1.46 E-16
ESR1	1.39 E-15
TAF1	1.07 E-13
HNRNPK	1.20 E-11
RAD21	7.91 E-11
MYC	2.93 E-08
ETS1	2.36 E-06
AP-2	3.07 E-05
IRF1	3.64 E-05
MZF1	2.91 E-04
GRHL2	9.20 E-04
HEY1	9.20 E-04
MAZ	2.98 E-03
RELA	1.58 E-02
MXI1	1.58 E-02

Table S11. Specific *TERT* regulators of LAML from the pan-cancer MIPRIP analysis.

TF	*p*-value
SMAD3	9.40 E-30
ESR2	9.37 E-29
IKZF1	8.63 E-27
NFAT5	6.67 E-26
CEBPA	4.29 E-16
MZF1	2.34 E-10
TCF7	9.38 E-06
SP3	3.17 E-04

Table S12. Specific *TERT* regulators of LIHC from the pan-cancer MIPRIP analysis.

TF	*p*-value
HMGA2	4.66 E-24
MITF	2.51 E-17
TFAP2D	3.01 E-15
SIN3A	6.33 E-12
ZBTB48	7.01 E-11
WT1	1.54 E-10
HNRNPK	7.73 E-06
MYC	1.29 E-05
REST	2.32 E-05
SP3	2.48 E-04
MAZ	3.60 E-04
POU2F2	4.26 E-03
E2F2	4.26 E-03
JUND	1.11 E-02
GRHL2	1.44 E-02
KLF2	3.83 E-02

Table S13. Specific *TERT* regulators of LUAD from the pan-cancer MIPRIP analysis.

TF	*p*-value
PAX5	3.61 E-15
E2F4	9.01 E-14
SMARCB1	5.89 E-12
NFKB.P50.P65	1.44 E-10
AP-2	1.44 E-10
RELA	2.25 E-09
NFATC2	9.83 E-09
ESR2	3.44 E-07
PAX8	4.44 E-03
HIF.1	5.25 E-03
POU2F2	2.93 E-02
BCL11A	2.93 E-02
BHLHE40	4.75 E-02
JUND	4.75 E-02
TAF1	4.75 E-02

Table S14. Specific *TERT* regulators of LUSC from the pan-cancer MIPRIP analysis.

TF	*p*-value
MYB	4.50 E-18
POU2F2	4.50 E-18
TCF7	1.10 E-14
BATF	3.30 E-14
EGR1	9.67 E-12
IRF1	1.07 E-11
EPAS1	3.32 E-08
RUNX2	9.25 E-06
E2F2	1.86 E-04
ZBTB48	5.60 E-04
E2F1	1.48 E-03
BHLHE40	7.94 E-03
MITF	8.25 E-03
TP73	1.10 E-02
ESR1	1.50 E-02
NFX1	4.18 E-02

Table S15. Specific *TERT* regulators of OV from the pan-cancer MIPRIP analysis.

TF	*p*-value
ESR2	6.81 E-11
MAX	9.02 E-11
E2F1	5.57 E-10
MEN1	2.37 E-07
HMGA2	6.38 E-07
POLR2A	1.78 E-04
NR2F2	1.01 E-03
AR	1.04 E-03
NFAT5	3.61 E-03
BCL11A	6.46 E-03
MAZ	1.10 E-02
PAX8	1.53 E-02
PITX1	2.21 E-02
TFAP2A	2.80 E-02
GRHL2	3.29 E-02

Table S16. Specific *TERT* regulators of PAAD from the pan-cancer MIPRIP analysis.

TF	*p*-value
TAL1	1.82 E-16
JUND	5.39 E-15
SIN3AK20	1.00 E-14
SIN3A	3.03 E-08
NFKB1	1.11 E-07
TFAP2B	9.40 E-07
NFKB.P50.P65	1.44 E-06
TCF12	9.19 E-06
IRF1	1.72 E-04
E2F6	2.19 E-03
CEBPA	4.72 E-03
REST	2.32 E-02

Table S17. Specific *TERT* regulators of SKCM from the pan-cancer MIPRIP analysis.

TF	*p*-value
E2F5	3.66 E-27
GLI2	2.98 E-19
MYCN	3.55 E-07
TP53	4.70 E-06
GRHL2	6.41 E-04
TCF7	4.15 E-03
CTCF	4.71 E-03
TFAP2C	1.24 E-02
HIF1A	1.67 E-02
REST	2.65 E-02
PAX5	4.07 E-02

Table S18. Specific *TERT* regulators of STAD from the pan-cancer MIPRIP analysis.

TF	*p*-value
MZF1	4.97 E-23
SIN3AK20	8.34 E-20
EGR1	5.48 E-19
PAX8	2.21 E-11
MYCN	6.15 E-11
MXI1	1.87 E-09
TFAP2C	3.47 E-07
E2F4	6.42 E-04
IKZF1	5.73 E-03
NR2F2	6.93 E-03
MYC	3.26 E-02
MXD1	3.63 E-02

Table S19. Specific *TERT* regulators of TGCT from the pan-cancer MIPRIP analysis.

TF	*p*-value
CTCFL	1.72 E-21
MYC	3.98 E-20
HEY1	4.73 E-18
TFAP2D	1.12 E-14
ZBTB48	1.12 E-14
JUND	7.67 E-09
MYCN	5.11 E-07
MAX	6.38 E-06
HIF.1	1.66 E-04
EPAS1	5.46 E-04
TAF1	5.46 E-04
HIF1A	6.06 E-03
TCF12	1.53 E-02
NFATC2	2.48 E-02

Table S20. Specific *TERT* regulators of THYM from the pan-cancer MIPRIP analysis.

TF	*p*-value
SMAD3	2.66 E-19
MXD1	6.21 E-14
IKZF1	1.12 E-10
RELA	1.09 E-09
MAX	2.95 E-07
ETS1	2.96 E-07
BATF	4.08 E-04
TP53	5.74 E-04
MZF1	1.72 E-03
SIN3AK20	5.50 E-03
E2F6	6.74 E-03
E2F4	7.62 E-03
E2F2	1.80 E-02
NFAT5	2.33 E-02
BHLHE40	4.04 E-02

Table S21. Specific *TERT* regulators of UCEC from the pan-cancer MIPRIP analysis.

TF	*p*-value
HNRNPK	3.04 E-27
NFKB1	2.43 E-17
PAX5	1.38 E-15
BCL11A	6.40 E-15
E2F4	7.88 E-14
SMARCB1	1.85 E-12
RUNX2	1.33 E-05
MITF	4.82 E-04
NFATC2	9.68 E-04
E2F5	2.10 E-03
JUND	2.60 E-03
BATF	4.05 E-02

Table S22. *TERT* Regulators significant for healthy prostate tissue compared to prostate cancer samples

Regulators normal	Frequency tumor	Frequency normal	*p*-value
TAF9	3	53	1.60 E-12
AP-2	1	42	2.15 E-11
ETS2	0	31	2.81 E-09
HIF1A	0	31	2.81 E-09
E2F5	0	29	1.06 E-08
HNRNPK	1	33	1.06 E-08
EPAS1	0	27	4.33 E-08
TP73	0	24	3.69 E-07
CTCFL	3	33	4.91 E-07
TFAP2B	2	30	5.57 E-07
SMAD3	0	21	2.62 E-06
MXD1	2	27	3.59 E-06
MYCN	3	29	5.13 E-06
ESR1	2	26	6.36 E-06
NFAT5	3	27	1.62 E-05
RUNX2	1	20	4.50 E-05
RELA	3	25	5.07 E-05
TP53	4	27	5.80 E-05
SP3	3	24	8.50 E-05
TAL1	6	29	1.75 E-04
EGR1	1	17	2.60 E-04
E2F4	7	29	4.41 E-04
HMGA2	7	28	7.23 E-04
NFKB.P50.P65	3	20	7.79 E-04
HIF.1	4	21	1.41 E-03
GRHL2	1	14	1.60 E-03
ZBTB48	8	28	1.60 E-03
NFX1	11	31	3.62 E-03
MZF1	2	14	6.35 E-03
PAX8	2	14	6.35 E-03
NFATC2	3	16	6.38 E-03
GLI1	1	11	9.37 E-03
E2F6	13	31	1.12 E-02
HEY1	9	24	1.70 E-02
TCF7	2	12	1.79 E-02
IKZF1	7	20	2.44 E-02
ESR2	4	15	2.51 E-02
NF.KB	1	9	2.90 E-02
JUND	13	28	3.12 E-02
WT1	2	10	4.83 E-02

Table S23. Association between PITX1 immunostaining results and prostate cancer phenotype in all tumors (data provided by AG Sauter/Simon)

Parameter	n evaluable	negative (%)	low (%)	high (%)	p value
All cancers	15,011	38.3	57.7	4.0	
Tumor stage					
pT2	9,555	41.5	55.5	3.0	<0.0001
pT3a	3,366	34.6	60.4	5.0	
pT3b-pT4	2,030	30.0	63.0	7.0	
Gleason grade					
≤3+3	2,794	41.8	55.3	2.8	
3+4	7,971	40.2	56.5	3.3	
3+4 Tert.5	720	38.9	57.6	3.5	<0.0001
4+3	1,479	30.6	62.8	6.6	
4+3 Tert.5	1,056	31.3	63.5	5.2	
≥4+4	867	28.7	61.5	9.8	
Lymph node metastasis					
N0	9,067	37.7	58.0	4.3	<0.0001
N+	1,121	30.2	63.2	6.6	
Preop. PSA level (ng/ml)					
<4	1,815	35.5	59.7	4.8	<0.0001
4-10	8,860	39.3	57.4	3.4	
10-20	3,167	38.3	57.2	4.5	
>20	1,081	36.9	56.7	6.4	
Surgical margin					
negative	11,973	39.2	57.1	3.7	<0.0001
positive	2,985	35.1	59.8	5.1	

Table S24. PedGBM samples and features for training the classifier (Deeg *et al.*, 2017)

ICGC-ID	CHROMO-THRIPSIS	*TP53*	*TERTP* MUTATION	*TERTP* METHYLATION	*TERT* RPKM	*ATRX*	*H3F3A*	ATRX IHC	TELOMERE CONTENT	ULTRA-BRIGHT TELOMERE FOCI	C-CIRCLE ASSAY	TMM
GBM005	1	1	0	0	0	1	1	1	0	0	1	ALT
GBM006	0	1	0	0	0	1	0	0	NA	1	1	ALT
GBM011	1	1	0	1	0	0	0	1	1	0	1	ALT
GBM015	1	1	0	0	0	1	0	0	NA	1	1	ALT
GBM017	1	1	0	0	0	1	1	0	1	1	1	ALT
GBM019	1	1	0	0	0	1	0	0	0	1	1	ALT
GBM027	0	1	0	0	0	0	1	0	NA	1	1	ALT
GBM028	0	1	0	0	1	1	0	0	NA	1	1	ALT
GBM050	1	1	NA	0	0	1	0	0	NA	NA	1	ALT
GBM056	1	1	0	0	0	1	0	NA	1	NA	1	ALT
GBM062	0	1	NA	0	0	1	1	NA	NA	NA	1	ALT
GBM067	1	1	0	0	0	1	0	NA	1	NA	1	ALT
GBM095	0	0	0	0	NA	1	0	NA	1	NA	1	ALT
GBM001	1	0	0	1	1	0	0	1	0	0	0	non-ALT
GBM002	1	1	0	0	0	0	1	1	0	0	0	non-ALT
GBM004	1	0	0	1	1	0	0	1	NA	0	0	non-ALT
GBM007	1	1	0	1	1	0	0	1	NA	0	0	non-ALT
GBM023	1	0	0	1	0	0	0	1	0	0	0	non-ALT
GBM032	1	1	0	1	1	0	0	1	NA	0	0	non-ALT
GBM034	1	0	0	0	0	0	0	1	NA	0	0	non-ALT
GBM036	0	0	0	0	0	0	0	1	NA	0	0	non-ALT
GBM038	1	1	0	0	1	1	0	1	0	0	0	non-ALT
GBM049	1	0	0	1	1	0	0	1	NA	0	0	non-ALT
GBM052	1	0	0	1	0	0	0	1	1	0	0	non-ALT
GBM053	1	1	0	1	1	0	0	1	0	0	0	non-ALT
GBM058	0	1	1	1	1	0	0	NA	0	NA	0	non-ALT*
GBM086	0	1	1	0	NA	0	0	NA	0	NA	0	non-ALT*
SF188	1	1	0	1	1	0	0	1	0	NA	0	non-ALT
KNS42	1	1	1	1	1	0	1	1	0	NA	0	non-ALT*
SJ-G2	1	1	0	1	0	1	0	0	1	NA	1	ALT
MGBM1	0	1	0	0	0	1	1	0	1	NA	1	ALT
NEM157	0	1	0	0	0	1	1	0	1	NA	1	ALT
NEM165	1	1	0	NA	0	0	1	1	1	NA	1	ALT
NEM168	1	1	0	NA	0	0	1	1	1	NA	1	ALT

* prediction based on *TERT* promoter mutation

Table S25. Predicted TMM in the remaining pedGBM patient samples

ICGC-ID	CHROMOTHRIPSIS	*TP53*	*TERTP* MUTATION	*TERTP* METHYLATION	*TERT* RPKM	*ATRX*	*H3F3A*	ATRX IHC	TELOMERE CONTENT	ULTRA-BRIGHT TELOMERE FOCI	C-CIRCLE ASSAY	PREDICTED TMM	ACCURACY
GBM012	0	1	0	0	NA	1	1	NA	NA	NA	0	ALT	0.906
GBM016	1	1	0	0	0	1	1	0	NA	1	NA	ALT	0.9
GBM018	0	0	0	1	1	0	1	0	NA	0	NA	non-ALT	0.9
GBM022	0	1	0	0	NA	1	1	0	NA	1	NA	ALT	0.9
GBM024	0	0	0	0	0	1	1	1	NA	0	NA	non-ALT	0.9
GBM025	0	1	0	0	0	0	0	NA	NA	NA	NA	non-ALT	0.9
GBM026	1	0	0	1	0	0	1	NA	NA	NA	0	non-ALT	0.9
GBM033	0	1	0	0	0	0	1	0	NA	1	0	ALT	0.9
GBM042	0	0	1	0	NA	0	1	NA	NA	NA	NA	non-ALT*	
GBM043	1	1	0	1	NA	1	1	NA	NA	NA	NA	ALT	0.906
GBM044	0	1	0	0	NA	1	1	NA	NA	NA	NA	ALT	0.906
GBM045	0	1	0	1	NA	1	0	NA	NA	NA	NA	ALT	0.906
GBM046	0	0	0	0	NA	0	0	NA	NA	NA	NA	non-ALT	0.906
GBM048	1	0	0	1	NA	0	0	NA	NA	NA	0	non-ALT	0.906
GBM057	0	0	NA	1	1	0	0	NA	0	NA	0	non-ALT	0.9
GBM059	1	1	1	1	1	0	0	NA	0	NA	0	non-ALT	0.9
GBM060	1	1	0	0	0	0	1	NA	1	NA	0	non-ALT	0.9
GBM061	0	0	0	0	NA	0	0	NA	NA	NA	0	non-ALT	0.906
GBM063	1	1	NA	1	1	1	1	NA	NA	NA	NA	ALT	0.9
GBM065	1	1	0	1	1	0	1	NA	0	NA	0	non-ALT	0.9
GBM066	1	1	0	1	0	0	0	1	0	NA	0	non-ALT	0.846
GBM071	1	1	0	1	NA	0	0	1	NA	NA	0	non-ALT	0.923
GBM079	0	1	NA	1	NA	1	1	NA	NA	NA	NA	ALT	0.906
GBM082	0	1	0	1	NA	0	0	NA	NA	NA	NA	non-ALT	0.906
GBM083	0	0	NA	1	NA	0	1	NA	NA	NA	NA	non-ALT	0.906
GBM084	0	0	0	1	NA	0	0	NA	NA	NA	NA	non-ALT	0.906
GBM085	1	1	0	1	NA	1	1	NA	NA	NA	NA	ALT	0.906
GBM096	0	1	0	0	NA	0	1	NA	1	NA	0	non-ALT	0.906
GBM098	0	1	0	NA	NA	1	1	NA	NA	NA	NA	ALT	0.853
GBM100	0	0	0	NA	NA	1	1	NA	NA	NA	NA	ALT	0.853

* prediction based on *TERT* promoter mutation

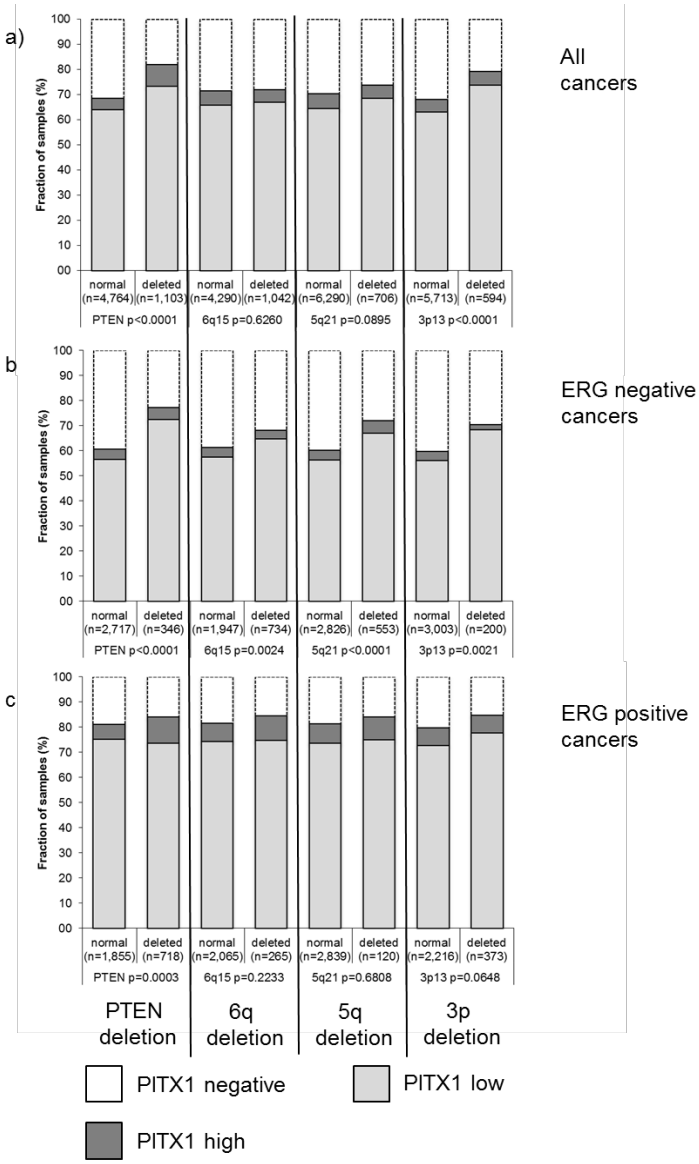

Figure S1. Association between PITX1 and PTEN, 6q15, 5q21 and 3p13 deletions

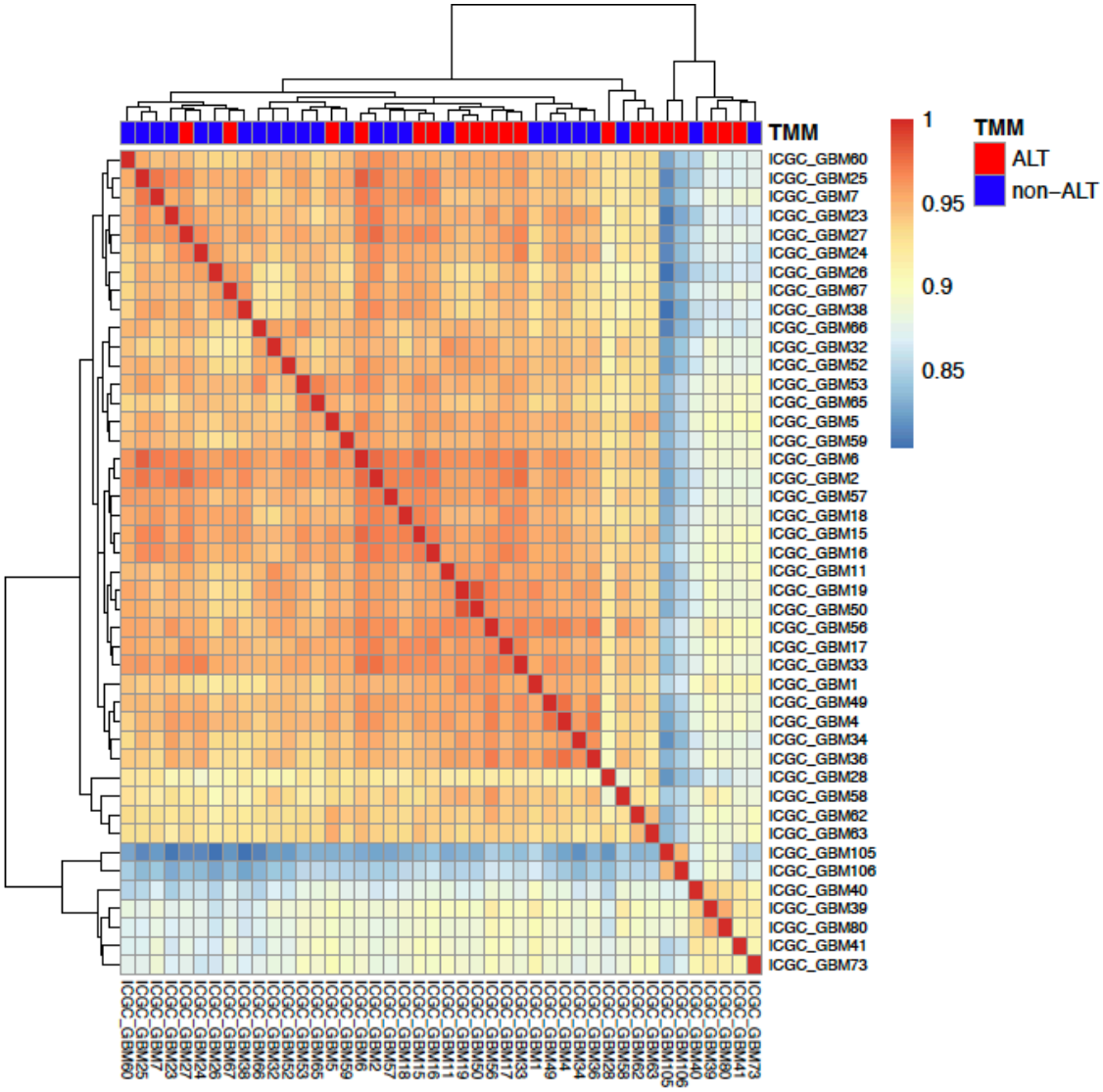

Figure S2. Clustering of pedGBM patient samples and cell lines based on gene expression data.

The cell lines are grouped in a separate cluster and were hence excluded by the differential gene expression and regulator activity analysis.

Acknowledgements

In particular I would like to thank Prof. Dr. Rainer König for providing me the opportunity to perform my PhD in his group, for his supervision and support throughout the years. Special thanks also go to Prof. Dr. Karsten Rippe for offering me to join his group, for his co-supervision, fruitful discussions and support.
I would also like to thank my thesis advisory committee members Prof. Dr. Franziska Matthäus and Dr. Carl Hermann for their time, support and fruitful discussions.
Many thanks to Prof. Dr. Ursula Kummer and Dr. Sevin Turcan for agreeing to be examiners on my defense committee.
Furthermore, I would like to especially thank all current and former members of the AG König and the AG Rippe for fruitful discussions, support and the great time in the lab/office, at retreats and lab outings as well as several evenings. It was a pleasure for me to work with you all!
Special thanks to Delia Braun, Verena Körber and Ashwini Kumar Sharma for critically reading my thesis, Caroline Bauer for performing the PITX1 experiments and the current and former members of the “Telo-Team” (Inn, Katha, Delia, Caro, Armin and Lukas) for intensive discussions about the world of telomeres. Special thanks also to Marcus Oswald for all the mathematical support, Volker Ast for handling all my bureaucratic stuff in Jena, Theresa Kordaß for constructing the generic human regulatory network, Amol Kolte and Joao Saraiva for taking care of the MIPRIP website as well as Daniela Röll and Neeraja Jaishankar for performing the validation experiments.
I would also like to thank all members of the CancerTelSys consortium, especially Manuel Gunkel, Roman Spilger, Thomas Wollmann, Holger Erfle as well as Karl Rohr for imaging and data analysis of the TMA slide; Andre Maicher and Brian Luke for performing the Sum1 validation experiment; and Ronald Simon, Cornelia Schröder, Jan Meiners and all people in the department of Guido Sauter who did the TMA staining and evaluation of our identified *TERT* regulators.
And a big thank-you to my friends Laura and Lisa at home as well as my friends in Heidelberg and abroad for all the the great support and nice things we did together.
Finally, particular thanks to my family, especially my parents for all their love and support over all the years!