Computational Cancer Research: Network-based analysis of cancer data disentangles clinically relevant alterations from molecular measurements

Habilitationsschrift

vorgelegt
der Medizinischen Fakultät Carl Gustav Carus
der Technischen Universität Dresden

von

Dr. rer. nat. Dipl.-Bioinf. Michael Seifert

geboren in Annaberg-Buchholz am 22. Oktober 1981

Dresden 2021

1. Gutachter:	Prof. Dr. Ingo Röder	Institut für Medizinische Informatik und Biometrie Medizinische Fakultät Carl Gustav Carus TU Dresden
2. Gutachter:	Prof. Dr. Lars Kaderali	Universitätsmedizin Greifswald Institut für Bioinformatik Universität Greifswald
3. Gutachter:	Prof. Dr. Holger Fröhlich	Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI Geschäftsfeld Bioinformatik Schloss Birlinghoven

Tag der Abgabe:	11. Mai 2021
Tag der Probevorlesung:	24. Februar 2022
Tag der Verteidigung:	12. Mai 2022

Prof. Dr. Guido Fitze

gez.: _____

Vorsitzende der Habilitationskommission

Contents

In the year 2020 you will be able to go into the drug store, have your DNA sequence read in an hour or so, and given back on a compact disk so you can analyze it.

Walter Gilbert, 1980

1 Introduction and objectives

Cancer is a complex genetic disease that is driven by combinations of mutated genes. These mutations can vary greatly between patients contributing to multiple subtypes with different causes and clinical outcomes. Nowadays, thousands of cancers have been characterized at the genomic, transcriptomic and epigenetic level. Frequently mutated genes and molecular subtypes of different types of cancer have been identified, but it is still extremely challenging to better understand the impact and interplay of mutations to improve prognosis predictions or to design more tailored therapies for individual patients.

Gene expression signatures that distinguish molecular subtypes can be determined in a straightforward manner by using standard statistical approaches, but the identification of major regulators among those genes that control such signatures is still a great challenge. Further, hundreds of genes are typically affected by large DNA copy number alterations, but computational methods that distinguish driver from passenger mutations to reveal which of these genes contribute to cancer development or therapy resistance are widely missing. Moreover, two cancers rarely share identical somatic gene mutation profiles even if they show similar clinical outcome. This means that apart from co-occurring and well-documented frequently mutated genes the vast majority of gene mutations are virtually private for each individual cancer. This raises the question about the role of rarely mutated genes in cancer and how we can predict the impact of all (rare and frequent) gene mutations on clinically relevant characteristics for each individual cancer patient?

Novel computational approaches are urgently needed to disentangle driver genes and molecular mechanisms that contribute to cancer development and therapy failure. A promising strategy is to consider cancer as a disease of combinations of mutated genes that alter cellular pathways and gene regulatory networks. Consequently, the computational analysis of molecular cancer data with the help of gene interaction networks has the great potential to overcome limitations of frequently used standard data analysis tools (e.g. statistical tests, regression methods), which mainly focus on single genes, cannot deal with rare gene mutations, and can only hardly distinguish between

driver and passenger mutations without the usage of prior knowledge.

This habilitation thesis contains a selection of seven publications with a specific focus on the development and application of omics data analysis strategies to address the following highly relevant research objectives:

- Characterization of similarities and differences of different astrocytoma grades to identify gene expression signatures and gene regulatory networks associated with the malignancy of astrocytomas
- Prediction of molecular subtypes of histologically classified oligodendrogliomas to derive characteristic gene expression signatures and associated altered gene regulatory networks
- Identification of molecular subtypes of DNMT3A-mutant acute myeloid leukemia patients to determine gene expression signatures and gene regulatory networks associated with survival differences
- Development of a computational network-based framework to enable a quantification of potential direct and indirect impacts of rare and frequent gene copy number alterations on clinically relevant characteristics
- Implementation of a user-friendly R package to provide the developed network inference and network propagation algorithms along with characteristic case studies to demonstrate the potential of the approaches
- Identification of novel driver gene candidates within the region of the 1p/19q codeletion of oligodendrogliomas
- Prediction of novel marker gene candidates associated with radioresistance and relapse behavior of prostate cancer patients

Besides the identification and characterization of molecular cancer subtypes, the overarching connection between these different studies is the integrative analysis of multiple omics layers by specifically developed algorithms for gene regulatory network inference and network propagation with the goal to identify potential major regulators and to quantify impacts of altered genes on clinically relevant characteristics (Fig. 1.1).

The work on these studies has been a very interesting and highly satisfying journey, which enabled me to work together with excellent researchers from different fields. I was able to develop novel network-based approaches for the integrative analysis of

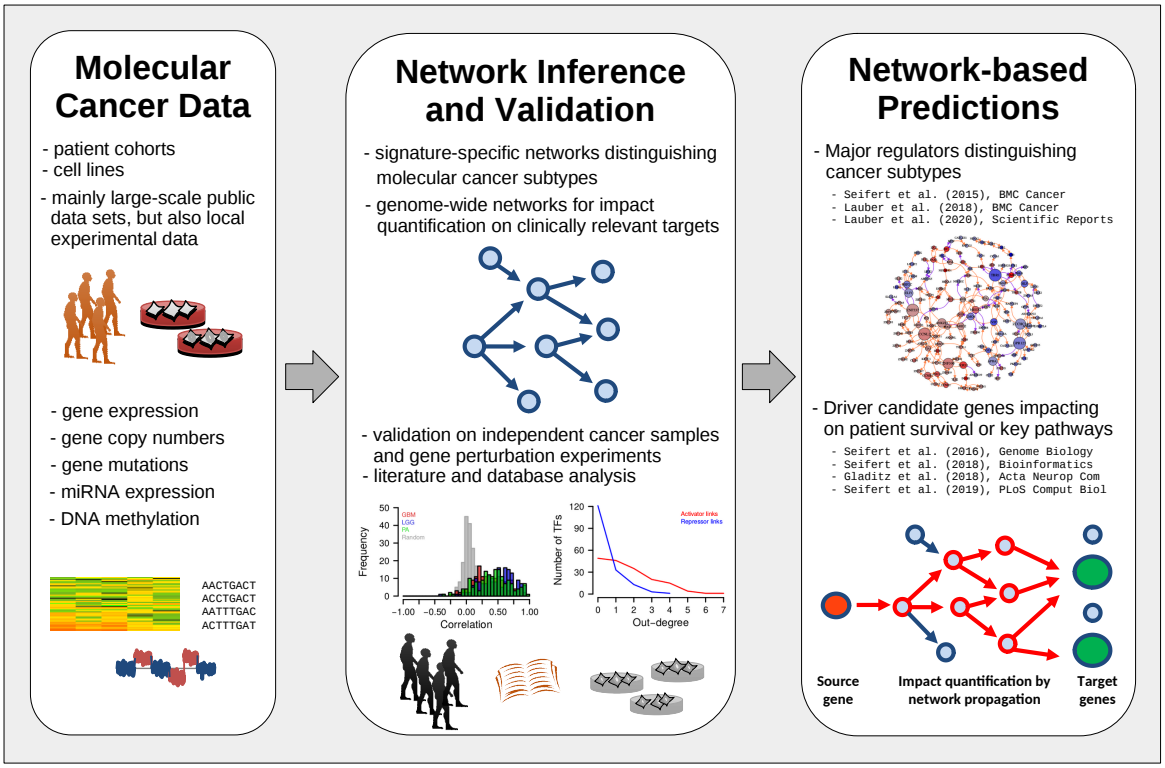

Figure 1.1: General data flow scheme of the network-based analyses that are part of the original works summarized in this habilitation thesis. Different omics data sets formed the basis of each study (left box). These data sets were used to learn signature-specific or genome-wide gene regulatory networks that were validated based on independent data sources (middle box). Resulting networks were used to determine major regulators of molecular signatures that distinguished cancer subtypes (right box: top) or to predict impacts of gene copy number alterations on clinically relevant target genes via network propagation (right box: bottom).

gene copy number and gene expression data filling gaps on the road map of computational tools for the analysis of cancer omics data. My habilitation thesis demonstrates the great potential of these network-based approaches in cancer research.

Outline of the habilitation thesis. In chapter 2, I provide a general overview of the scientific background going from the hallmarks of cancer over the complexity of cancer genomes down to the importance of networks in cancer. This also includes a brief introduction to the mathematical concepts that underlie the network inference and network propagation algorithms developed for the publications that are part of this habilitation thesis. In chapter 3, I briefly motivate and summarize my studies. Chapter 4 contains the publications that I selected for my thesis. The scientific background and results of each publication are introduced and summarized before each publication. In chapter 5, I close the habilitation thesis with a discussion of the results.

2 Scientific background

2.1 Hallmarks of cancer

Cancer is a collective term for related genetic diseases that are characterized by uncontrolled cell proliferation and spread of cancer cells into the surrounding tissue or to distant organs. About two decades ago major discoveries of the past 25 years of cancer research were summarized to define the hallmarks of cancer in a seminal review manuscript by Hanahan and Weinberg (2000). The authors initially defined six essential physiological alterations (self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential, sustained angiogenesis, evasion of apoptosis) that jointly contribute to the malignant growth of cancer cells. These six hallmarks have helped to better structure the complexity of numerous interrelated properties of cancer, but this integrative reductionist view was also not free of critique (e.g. Lazebnik (2010)). Many new additional insights into tumorigenesis were gained in the first decade after the publication of the initial hallmarks.

This has led to a revised review on the hallmarks of cancer by Hanahan and Weinberg (2011). Two new emerging hallmarks (avoiding immune destruction, deregulating cellular energetics) and two enabling traits (genome instability and mutation, tumor-promoting inflammation) were added by the authors to their initial model of cellular alterations that contribute to cancer development. This hallmark model represents an excellent summary of cellular components and processes that are altered in cancer, but it does not answer when and why those alterations arise. The variation in cancer risk among different tissues is mainly attributable to different rates of stem cell divisions, which increases the risk of developing cancer with increasing age (Tomasetti and Vogelstein (2015)).

Consequently, the hallmarks of cancer can be embedded into an evolutionary transformation process with underlying clonal selection of acquired traits that drive the transition of normal to malignant cells (Fouad and Anaei (2017)). This is illustrated in Fig. 2.1. Nowadays, we have a clear notion of the basic principles of cancer devel-

opment, but we are still at the beginning to understand the critical details of individual cancers.

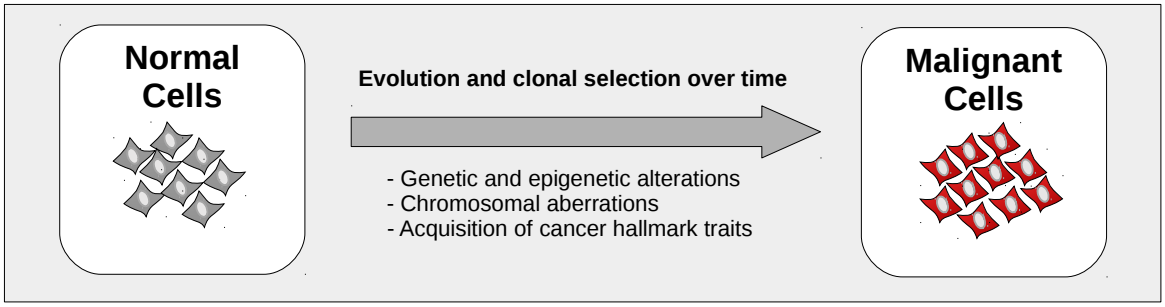

Figure 2.1: Transformation of normal cells to malignant cells. Normal cells are continuously exposed to environmental factors that contribute to DNA damage or are faced with DNA replication errors during cell division. An accumulation of such DNA damages due to DNA repair deficiencies can initiate a cyclic evolutionary process of clonal selection that can transform normal cells to malignant cells over time.

2.2 Complexity of cancer genomes

Fast progress in the development of high-throughput experimental technologies over the last two decades enabled to measure molecular data of different types of cancer with unprecedented detail. Multi-omics analyses of thousands of cancer patients by The Cancer Genome Atlas (TCGA) revealed an enormous complexity of cancer genomes, transcriptomes and methylomes within and between different types of cancer (e.g. The Cancer Genome Atlas Research Network (2008, 2011, 2012c,a,b, 2013a,b, 2014a,b, 2015)).

The systematic analysis of cancer genomes revealed that a single tumor can carry several hundred up to even thousands of somatic mutations depending on the type of cancer (Vogelstein et al. (2013)). Interestingly, major driver genes frequently mutated in specific cancer types were found, but the potential role of almost all infrequently mutated genes is still largely unknown (Vogelstein et al. (2013); Lawrence et al. (2014)). This mutational heterogeneity suggests that many different subtypes of each specific type of cancer exist. This is supported by the observation that individual tumors of the same type of cancer can show strikingly different gene mutations (Mardis (2014)). Still, tumors with largely different gene mutation patterns can show comparable clinical outcomes (Hofree et al. (2013)). A widely accepted explanation for this mutational

heterogeneity of tumors is that cancer arises due to alterations of cellular hallmark pathways. Typically, only a single gene of a cancer-relevant signaling pathway is mutated per tumor most likely because additional mutations of genes of the same pathway do not provide further advantages (Ciriello et al. (2012)).

Further, also DNA copy number alterations and chromosomal instability are a hallmark of cancer (Hanahan and Weinberg (2011); Ciriello et al. (2013); Zack et al. (2013)). Recurrently occurring deletions or duplications of whole chromosomes or chromosomal arms affecting hundreds of genes or infrequently observed deletions and duplications of smaller DNA fragments affecting only very few or a single gene are hard to analyze for driver genes and suffer from the same problem even if the number of altered genes differs. In the first case, the recurrent occurrence of almost identical large chromosomal mutations (e.g. the 1p/19q co-deletion in oligodendrogliomas or the amplification of chromosome 7 in glioblastomas) does not allow to distinguish between driver and passenger genes, because one cannot narrow down specific chromosomal regions. In the second case, the rare occurrence does not allow to obtain robust results, because similar cancer samples are often not available for systematic studies of driver impacts. These challenges cannot be overcome by standard statistical and bioinformatics methods for molecular data analysis.

Overall, the complexity of cancer genomes strongly complicates the identification of driver mutations for individual cancers and puts great challenges to reveal how these mutations alter molecular mechanisms that influence pathogenesis and therapy response. A promising way to overcome this is to consider cancer as a disease of cellular pathways or networks and to integrate this network principle into the development of novel computational strategies to improve the analysis of individual cancer genomes (Krogan et al. (2015)).

2.3 Importance of cellular networks in cancer

Networks generally represent a simplified representation of a complex system to capture basic connections between the components of a system. One can study the individual components of a system and individual connections between components in isolation, but to gain detailed insights on how the full system works it is necessary to study the pattern of connections between all components (Newman (2010)).

This concept can also be transferred to molecular cancer research to analyze heterogeneous molecular omics data sets with the help of network-based approaches.

The components of a cellular network are represented by nodes that constitute genes (or proteins) and connections between nodes are represented by edges that constitute gene-gene (or protein-protein) interactions. This network representation specifies a basic computational model for the analysis of molecular data that is widely considered for its simplicity, generality and potential to identify complex molecular patterns (Barabási and Oltvai (2004)). Instead of analyzing each gene in isolation, we can utilize known or computationally predicted regulatory interactions between genes to analyze how gene mutations or altered transcription levels putatively influence other genes or hallmark pathways driving cancer development and clinical outcomes. Such a network-based analysis of tumor data directly accounts for the fact that cancer is a complex genetic disease that is driven by combinations of mutated genes that greatly vary between individual tumors. This is the central component of computational network medicine for complex human diseases to identify altered molecular modules and pathways and to predict relationships between pathogenic geno- and phenotypes (Barabási et al. (2011)).

One of the first studies that demonstrated the great power of the analysis of molecular cancer data with the help of protein interaction networks was the classification of breast cancer into metastatic and non-metastatic tumors by Chuang et al. (2007). Chuang et al. analyzed gene expression data of two independent studies from van't Veer et al. (2002) and Wang et al. (2005), which had initially shown only very poor overlap of individual marker genes correlated with metastatic potential. Early network-based analysis approaches mainly followed the principle of 'guilt by association' that is motivated by the observation that genes or proteins share molecular and phenotypic properties with their direct network interaction partners (Schwikowski et al. (2000)). This concept has been further generalized in different approaches to include the local network neighborhood of a gene or protein to improve the identification of gene clusters or modules (Brohée and van Helden (2006)). These local network neighborhood approaches are nowadays frequently replaced by network propagation algorithms (Cowen et al. (2017)).

Starting with prior information of initially altered nodes (e.g. mutated genes, differentially expressed genes), network propagation algorithms transmit information from each node to its direct neighbor nodes in an iterative manner enabling to predict previously hidden data patterns that emerge from the underlying molecular network. Network propagation has been developed and applied in different research disciplines including statistical physics, electrical engineering, machine learning, data sciences, biology

and medicine (Cowen et al. (2017)). Successful biological and medical applications include gene function prediction (e.g. Noble et al. (2005); Sharan et al. (2007)), module discovery (e.g. Mitra et al. (2013); Leiserson et al. (2015)), disease characterization and disease gene prediction (e.g. Cho et al. (2012); Ideker and Sharan (2008); Ruffalo et al. (2015)), prediction of novel drug targets (e.g. Csermely et al. (2013); Chen et al. (2012); Shnaps et al. (2016)) and patient stratification and subtype discovery (e.g. Vanunu et al. (2010); Hofree et al. (2013); Zhang et al. (2018)). Almost all approaches utilized existing human protein-protein interaction networks such as STRING (Szklarczyk et al. (2017)) or data of biological pathways such as Pathway Commons (Cerami et al. (2011)) as basis for network propagation. These networks usually contain experimentally validated interactions and interactions predicted by computational methods.

Such global pre-existing networks mainly provide a general representation of interactions between genes or proteins across different tissues. But cells of specific tissues or within a specific tissue can largely differ in their gene expression profiles. Consequently, not each reported interaction will exist in each cell. This is critical for network propagation and can lead to inaccurate results due to the usage of interactions that may not be possible or exist in reality. Thus, as an alternative to these general networks, the usage of tissue-specific networks directly learned from molecular data can contribute to better account for the great heterogeneity within and across different types of cancer and further provides the opportunity to develop network propagation algorithms that exploit individual cancer gene expression profiles.

2.4 Inference of cancer-specific gene interaction networks from molecular data

The computational reconstruction of gene regulatory networks (also called reverse engineering) from molecular data represents one of the fundamental challenges in computational biology. The accurate prediction of biological interactions between genes is essential to obtain a systems-based view and detailed insights into molecular mechanisms that drive individual cancers. High-throughput technologies to measure genome-wide gene expression profiles allow to gain snapshots of transcriptomes of different cancer cells that can be used to learn gene regulatory networks. The computational challenge is to predict regulatory dependencies between regulators and their target genes. The aggregation of all predicted interactions between genes comprises the

learned gene regulatory network.

A broad range of more than thirty different network inference methods have been developed over the last two decades as outlined in different reviews and comparison studies (e.g. Li et al. (2008); Marbach et al. (2010, 2012); Chai et al. (2014)). The used computational frameworks include regression, correlation and mutual information approaches, but also other methods such as ANOVA, Boolean networks, Bayesian networks or Random Forest have been studied. Some of the methods were exclusively developed for the inference of gene regulatory networks from gene expression data (De Smet and Marchal (2010); Marbach et al. (2010)), whereas other methods integrate multiple data layers (e.g. Bar-Joseph et al. (2003); Reiss et al. (2006); Jörnsten et al. (2011); Marbach et al. (2012)). These methods differ in their data requirements and large-scale comparison studies revealed that each algorithm has certain strengths and weaknesses and that their results can differ substantially (Marbach et al. (2010, 2012)).

Interestingly, sparse linear regression approaches were among the best performing methods for gene regulatory network inference (Marbach et al. (2012)). Such regression approaches also enable a straightforward interpretation of the predicted links and their corresponding parameters, which is important for the design of specific network propagation algorithms that take the gene expression levels of individual genes into account instead of only considering the existence of regulatory links between genes. Nevertheless, these regression-based methods varied largely in their performances due to the usage of different underlying data resampling strategies (Marbach et al. (2012)). Still, sparse linear models represent a well-suited simplified approach to model the gene expression behavior of the vast majority of genes (Jörnsten et al. (2011)). Limitations of the resembling strategies can be overcome by repeating the network inference several times to later focus only on links that were robustly identified in the majority of networks or to utilize the whole ensemble of regression-based networks for downstream analyses. Such ensemble-based strategies have a long tradition in computational systems biology to derive robust network models from experimental data (Kaltenbach et al. (2009); Marbach et al. (2009, 2010, 2012)).

In this habilitation thesis, I focused on the inference of gene regulatory networks utilizing sparse regression models. The good performance of these models in combination with their simplicity and interpretability also enabled a direct integration of the model parameters into the development of novel network propagation methods that make use of the learned regression models to quantify impacts of altered genes on

clinically relevant downstream targets and cellular pathways. The basics of the underlying network inference algorithm are described in the following.

2.5 Network inference based on sparse regression

Following the detailed descriptions in Seifert et al. (2016) and Seifert and Beyer (2018), we divide the network inference problem into independent gene-specific sub-network inference tasks. We assume that for each gene $i \in \{1, \ldots, N\}$ its expression level e_{id} in a sample $d \in \{1, \ldots, D\}$ is modeled by a linear combination

$$e_{id} = a_{ii} \cdot c_{id} + \sum_{j \neq i} a_{ji} \cdot e_{jd} \tag{2.1}$$

of its gene-specific copy number c_{id} and the expression levels e_{jd} of other potential regulator genes $j \neq i$. The parameters of this gene-specific linear model are defined by $\vec{a}_i := (a_{1i}, \ldots, a_{Ni}) \in \mathbb{R}^N$, where a_{ii} quantifies the direct local gene copy number effect and a_{ji} with $j \neq i$ specifies the contribution of the expression of gene j on the expression of gene i. It has already been shown that linear models represent a reasonable approximation for the modeling of gene expression levels for the majority of genes (Jörnsten et al. (2011)).

We use lasso (least absolute shrinkage and selection operator) regression (Tibshirani (1996)), which realizes variable selection and regularization, to compute a sparse solution for the linear model in (2.1) by minimizing the residual sum of squares to determine an optimal solution

$$\vec{a}_i^{\,*} := \underset{\vec{a}_i}{\operatorname{argmin}} \sum_{d=1}^{D} \left(e_{id} - \left(a_{ii} \cdot c_{id} + \sum_{j \neq i} a_{ji} \cdot e_{jd} \right) \right)^2 + \lambda_i \sum_{j=1}^{N} |a_{ji}| \tag{2.2}$$

for each gene i with respect to a fixed complexity parameter $\lambda_i \geq 0$ that specifies the amount of shrinkage of the individual model parameters in $\vec{a}_i$ toward zero. Larger values of λ_i lead to greater shrinkage enabling to select the most relevant predictors (own gene-specific copy number and expression levels of other genes) that best explain the expression of the response gene i. Irrelevant model parameters are automatically shrunken to zero by lasso.

The obtained model parameters $\vec{a}_i^{\,*}$ depend on the choice of the gene-specific complexity parameter λ_i. We determine the optimal gene-specific complexity parameter along with the corresponding optimal model parameters by cross-validation using the

R package glmnet (Friedman et al. (2010)). We further determine the significance of model parameters when they first enter the lasso model in (2.2) using a significance test for lasso (Lockhart et al. (2014)). In more detail, we first compute the lasso solution paths for all active predictors (model parameters in $\vec{a}_i^*$ that are unequal zero) of the gene-specific linear model for the given data set using the R package lars (Hastie and Efron (2013)). These paths are evaluated using the R package covTest (Lockhart et al. (2013, 2014)) to obtain p-values that quantify the relevance of individual active predictors for the gene-specific linear model. The network inference step for gene i is illustrated in Fig. 2.2.

The obtained optimal parameters of the sparse linear model can be directly interpreted in the context of relationships between genes j and i: $a_{ji}^* < 0$ suggests that the putative regulator j is associated with the inhibition of target i, $a_{ji}^* > 0$ suggests that the putative regulator j is associated with the activation of target i, and $a_{ji}^* = 0$ suggests that no putative regulatory link between j and i exists. It is important to note that a dependency between the genes j and i can either represent direct or indirect causal interactions or only a correlation. The integration of copy number data into the linear model extends pure correlation-based network inference approaches.

Initially, we considered the standard detection limit of the covariance test implementation by Lockhart et al. (2013) to select highly predictive links. This strategy was used in Seifert et al. (2015), Seifert et al. (2016), Lauber et al. (2018), and Seifert et al. (2019), which are all part of this habilitation thesis. Later, in my R package regNet (Seifert and Beyer (2018)), I explicitly modified the R function covTest from Lockhart et al. (2013) to avoid the undocumented implicit rounding of p-values to four decimals. This enabled a correction of p-values for multiple testing by computing false discovery rates (FDRs) (Benjamini and Hochberg (1995)) considering the p-values computed for the active parameters of all gene-specific linear models. These FDR-adjusted p-values were considered to obtain the networks in Seifert and Beyer (2018), Gladitz et al. (2018), and Lauber et al. (2020).

Further, the described network inference approach has been used in different variations within the frame of this habilitation thesis. Genome-wide gene regulatory networks based on gene copy number and expression profiles have been learned in Seifert et al. (2016), Seifert and Beyer (2018), Gladitz et al. (2018), and Seifert et al. (2019). Gene regulatory networks associated with gene expression signatures have been learned in Seifert et al. (2015), Lauber et al. (2018), and Lauber et al. (2020). Study-specific modifications are described in the corresponding manuscripts motivat-

ing restrictions to transcription factor expression levels as predictors instead of considering all genes (Seifert et al. (2015); Lauber et al. (2018)) and the usage of miRNA expression levels as additional omics layer instead of gene copy number profiles (Lauber et al. (2020)). Details to the integration of network instances by focusing on reproducible links or by ensemble-based integrations of network results are given in the publications.

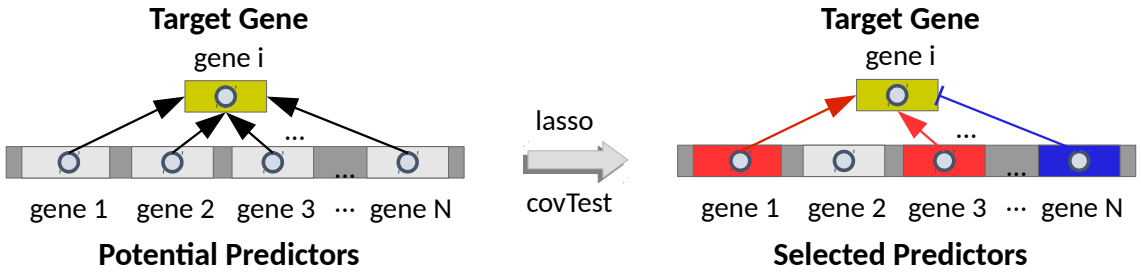

Figure 2.2: Illustration of the network inference step for target gene i. The expression level of target gene i is modeled as a linear combination of potential predictors. Lasso regression in combination with a significance test for lasso is used to select the most relevant predictors (red: activation link, blue: inhibitory link). This network inference step is done for each gene i to obtain a global gene regulatory network.

2.6 Network-based prediction of gene expression levels

Again following the detailed descriptions in Seifert et al. (2016) and Seifert and Beyer (2018), a regression-based gene regulatory network can be used to predict the expression levels of genes in a given data set. We compute the correlation between predicted $\hat{e}_{id}$ and originally measured expression levels e_{id} for each gene $i \in \{1, \ldots, N\}$ across all samples $d \in \{1, \ldots, D\}$ of a given data set to quantify the predictive power of each network. This is done by computing the Pearson correlation coefficient

$$r_i := \frac{\sum_{d=1}^{D} (\hat{e}_{id} - \overline{\hat{e}}_i) \cdot (e_{id} - \overline{e}_i)}{\sqrt{\sum_{d=1}^{D} (\hat{e}_{id} - \overline{\hat{e}}_i)^2 \cdot \sum_{d=1}^{D} (e_{id} - \overline{e}_i)^2}} \tag{2.3}$$

for each gene i. The corresponding mean values of the predicted and the originally measured expression levels of gene i over all samples D are given by $\overline{\hat{e}}_i$ and $\overline{e}_i$, respectively. A correlation $r_i > 0$ implies that the learned network is able to predict the measured expression trend of a gene: the greater positive the correlation r_i the better the network-based prediction.

These correlations can be used to evaluate how good a network is able to predict the expression behavior of individual genes and to analyze the location and shape of the correlation distribution obtained for all genes. This enables comparisons of network instances of different complexities and comparisons to random baseline network models. In addition, gene-specific correlations between predicted and originally measured expression levels provide the basis to integrate the quality of the predictions of individual genes into the impact computations of my newly developed network propagation algorithm, which is described in the next section.

2.7 Network-based propagation of gene expression alterations

Networks can be used to determine impacts of gene perturbations (e.g. gene expression changes due to directly underlying gene copy number alterations) on other genes in the network. The key idea is to propagate these impacts along the network edges from the affected gene to its direct neighbors and from those to their direct neighbors in an iterative manner. This can be realized by network propagation algorithms.

We developed a novel network propagation algorithm in Seifert et al. (2016) that quantifies for each gene pair (j, i) the direct and indirect contribution of gene j on the expression of gene i under consideration of all existing network paths from j to i, the prediction quality of individual genes along the paths, and possibly existing feedback loops. These impacts can be computed over all patients in a cohort or for each individual patient. It is also possible to integrate potential inhibitor or activator contributions. Initial mathematical descriptions of the different variations of the network propagation algorithm have been provided in Seifert et al. (2016) along with in-depth validation studies. Next, I introduce the basic version of the network propagation algorithm that we used for the computation of a cohort-specific impact matrix following the detailed descriptions provided in Seifert et al. (2016) and Seifert and Beyer (2018).

We consider a data set of D samples for which the expression level e_{id} and the

copy number c_{id} of each gene $i \in \{1, \ldots, N\}$ have been measured for each sample $d \in \{1, \ldots, D\}$. We further consider a learned network and its underlying gene-specific linear models specified in (2.1). We denote the optimal parameter vector of the linear model of gene i in (2.2) by $\vec{a}_i^* := (a_{1i}^*, \ldots, a_{Ni}^*)$. First, we compute for each gene i the correlation coefficient r_i between predicted and originally measured expression levels across all samples of the given data set as specified in (2.3). Subsequently, we only consider predictable genes with positive correlations between predicted and observed expression levels ($r_i > 0$). Poorly predictable genes (i.e. genes with small positive r_i) will only contribute very little to the total impact score. Next, we compute the corresponding explained variance $R_i^2 = r_i \cdot r_i$ of each predictable gene covered by its underlying sparse linear model in (2.1). We further set $R_i^2 := 0$ for unpredictable genes $(r_i < 0)$ to exclude those genes from the network propagation.

Let us now consider each regulator gene j of gene i to determine for each regulator its direct contribution to the observed explained variance R_i^2 of gene i. We first compute the average proportion of each regulator j on the prediction of the expression of target gene i by

$$p_{ji} = \frac{1}{D} \sum_{d=1}^{D} \frac{|a_{ji}^* \cdot e_{jd}|}{|a_{ii}^* \cdot c_{id}| + \sum_{v \neq i} |a_{vi}^* \cdot e_{vd}|}$$

and we determine the direct average copy number contribution of target gene i by

$$p_{ii} = \frac{1}{D} \sum_{d=1}^{D} \frac{|a_{ii}^* \cdot c_{id}|}{|a_{ii}^* \cdot c_{id}| + \sum_{v \neq i} |a_{vi}^* \cdot e_{vd}|}$$

under consideration of all D samples of the given data set. The usage of absolute values in the computation of p_{ij} (and p_{ii}) accounts for regulator genes that either act as potential inhibitors or activators of target gene i enabling to reveal the strongest regulators independent of their mode of action. If a gene j is not a direct regulator of gene i (learned $a_{ji}^* = 0$ in the underlying gene-specific linear model of gene i), then $p_{ji} := 0$. In analogy, if target gene i does not have a direct copy number effect (learned $a_{ii} = 0$ in the underlying gene-specific linear model of gene i), then $p_{ii} := 0$. This allows to define a basic network flow matrix

$$F = (f_{ji})_{1 \leq j,i \leq N} := p_{ji} \cdot R_i^2 \tag{2.4}$$

by weighting the explained variance R_i^2 of target gene i with the average proportion p_{ji}

of its direct predictors (gene copy number, regulator genes) j. F quantifies the direct impacts of regulators on target genes. In more detail, each column i of F contains the explained variance of a target gene i split into average proportions according to the contributions of its copy number and its target gene-specific regulators. Since the predictions of expression levels by the underlying gene-specific linear model are not perfect, the explained variance fulfills $0 \leq R_i^2 < 1$. Therewith, the column sum norm of F is strictly less than one. We use this property to compute indirect effects between each pair of genes (i.e. the network flow) via the following equation

$$F^* = \sum_{k=1}^{\infty} F^k \tag{2.5}$$

that sums over the contributions of all network paths of increasing length k. The matrix F^k specifies the k-th matrix power obtained by a k-fold matrix multiplication of F. An element f_{ji}^k of F^k represents the impact of a trans-acting regulator gene j on the explained variance of a target gene i via all directed network paths from j to i of length k. Since the basic network flow matrix F has a column sum norm that is strictly less than one, the network flow F^* will converge to its limit $(I - F)^{-1} - I$ (geometric series of matrix F starting at one), where I is the identity matrix and $(I - F)^{-1}$ specifies the inverse of matrix $I - F$.

However, the computation of the inverse of a large matrix ($I-F$ has dimension $N \times N$) is very time consuming. In addition, due to the sparsity of F (majority of entries are zero because only the most relevant predictors should be included in a network) and its entries in $[0, 1)$, we also know that the values of the elements in F^k should relatively quickly approach zero. Thus, it is more efficient to approximate F^* by only adding an additional F^k if the obtained difference of the sum over F^k up to k and the previous sum up to $k - 1$ is greater than a predefined threshold. We stopped the approximation of F^* if the sum of the differences of the column sums of the current and the previous approximated matrix is less than 10^{-3}. The resulting matrix F^* represents the impact matrix that contains for each gene pair (j, i) the direct and indirect impacts that flow via the underlying network form gene j to gene i. The absolute impact of a gene j on the expression of a gene i is given by the entry f_{ji}^*. The basic idea of impact quantification by network propagation is illustrated in Fig. 2.3.

We considered impact matrices computed by this network propagation approach in Seifert et al. (2016), Seifert and Beyer (2018), Gladitz et al. (2018), and Seifert et al. (2019), which are all part of this habilitation thesis, to analyze downstream impacts of

observed gene copy number or expression alterations on cancer-relevant pathways or clinically relevant signature genes.

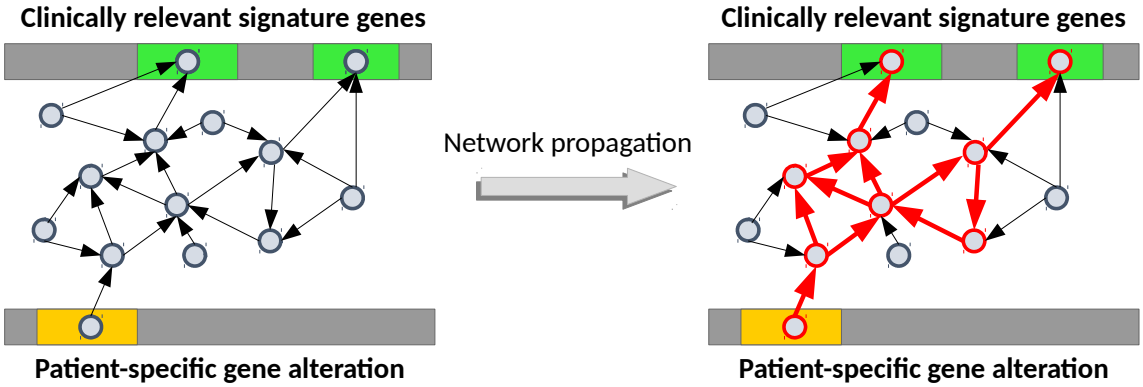

Figure 2.3: Illustration of impact quantification by network propagation. The potential impact of a patient-specific gene alteration (orange box) on clinically relevant signature genes (green boxes) can be quantified with the help of a gene regulatory network. All existing network paths (red arrows) that connect the altered gene to the individual signature genes are considered. This strategy allows to quantify direct or indirect impacts between each pair of genes enabling to analyze the potential impact of each patient-specific gene alteration on clinically relevant target genes by taking all individual patient-specific alterations into account.

3 Motivation and summary of studies

After completing my PhD in bioinformatics with a specific focus on the development of Hidden Markov Models for the analysis of sequential data from high-throughput omics experiments (Seifert (2010)), it became clear to me that such tools in combination with methodological developments in the field of computational systems biology and systems medicine would offer a great chance to continue my career with a strong focus on medical applications. I started to work as a postdoc on network-based approaches for the analysis of cancer omics data in May 2012 in the research group of Prof. Dr. Andreas Beyer at the Biotechnology Center (BIOTEC) TU Dresden. In December 2015, after two more postdoc positions, I started to work as a group leader at the Institute for Medical Informatics and Biometry (IMB) TU Dresden headed by Prof. Dr. Ingo Roeder. I established a Bioinformatics Core Unit at the IMB to support the analysis of molecular high-throughput data and to develop innovative methods and strategies for the integrative analysis of omics data. Over all these years, I had the chance to work together with researchers from different disciplines to lay the ground for this habilitation thesis.

3.1 Molecular stratification and driver gene identification for gliomas

In 2012, I met Dr. Barbara Klink, who was working at the Institute for Medical Genetics of the Faculty of Medicine of the TU Dresden, and we started to exchange ideas and decided to work together on different brain cancer research projects with a specific focus on the analysis of molecular data of astrocytomas and oligodendrogliomas belonging to the group of gliomas (Ohgaki and Kleihues (2013); Louis et al. (2016)).

One of our first key questions was to find out how different grades of astrocytomas differ at the molecular level and which molecular factors contribute to their increasing malignancy? Omics data sets of different astrocytoma grades were publicly available, but studies that compared all four grades did not exist with few exceptions dating back to the time when omics approaches were still in their infancy (Rickman et al. (2001);

Hunter et al. (2002); Rorive et al. (2006)). To address this, we systematically characterized similarities and differences between the astrocytoma grades at the level of single genes, signaling pathways and gene regulatory networks. The results of our study were published as (see also Section 4.1):

- ***Michael Seifert**, Martin Garbe, Betty Friedrich, Michel Mittelbronn and Barbara Klink (2015): **Comparative transcriptomics reveals similarities and differences between astrocytoma grades**, BMC Cancer, 15:952.*

This first success motivated us to further analyze molecular data of oligodendrogliomas, which are closely related to astrocytomas. Oligodendrogliomas were classified purely based on histology for many years (Louis et al. (2007)), but these histological classifications were known to be error-prone and not always consistent between different neuropathologists (Coons et al. (1997); van den Bent (2010)). When we started this project in December 2015, characteristic molecular markers of oligodendrogliomas (1p/19q co-deletion and IDH1/2) had already been identified and had further been shown to improve diagnosis and prediction of treatment response (Cairncross et al. (1998); Jansen et al. (2010); Labussiere et al. (2010)), but almost all publicly available data sets of oligodendrogliomas were still established on the basis of pure histological classifications. We therefore decided to analyze publicly available gene copy number and gene expression profiles of histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) with the goal to identify and characterize molecular subtypes, associated gene regulatory networks and potential major regulators. The results of our study were published as (see also Section 4.2):

- *Chris Lauber, Barbara Klink and **Michael Seifert** (2018): **Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups**, BMC Cancer, 18:399.*

This study enabled us to gain a deep understanding of molecular alterations that characterize oligodendrogliomas, but one of the main challenges still remained. The identification of core driver genes involved in oligodendroglioma development had not made much progress since many years. The 1p/19q co-deletion is most likely caused by an unbalanced translocation (Jenkins et al. (2006)), but no fusion genes that drive the tumor development have been found most likely because the break points are located in a gene-poor heterochromatic region. Further, in-depth searches for inactivating point mutations have identified FUBP1 located on 1p and CIC located on 19q as potential tumor suppressors (Bettegowda et al. (2011); Eisenreich et al. (2013)), but

both mutations are only observed in some or an increased fraction of patients (The Cancer Genome Atlas Research Network (2015)) implying that they are not driving the initial tumor development. Essentially, the challenge is that hundreds of genes on the p-arm of chromosome 1 and on the q-arm of chromosome 19 are affected by the co-deletion of one copy of these chromosomal arms. Since oligodendrogliomas show nearly identical co-deletions, it is not possible to simply overlay the copy number profiles of many oligodendrogliomas to localize chromosomal regions on 1p and 19q that potentially drive tumor development. Further, a differential gene expression analysis of genes of the 1p/19q region will result in hundreds of differentially expressed genes, but one cannot distinguish between drivers and passengers. Thus, the identification of driver gene candidates within the region of the 1p/19q co-deletion is an enormous challenge that can only be addressed by the usage of novel innovative methods.

One promising way to address this was the application of network-based data analysis strategies that I had already established in Seifert et al. (2016). Consequently, we searched for novel driver gene candidates with the help of oligodendroglioma-specific gene regulatory networks to determine how differentially expressed genes within the region of the 1p/19q co-deletion act on altered cancer-relevant signaling and metabolic pathways. The results of our study were published as (see also Section 4.6):

- *Josef Gladitz, Barbara Klink and* ***Michael Seifert*** *(2018):* ***Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion****, Acta Neuropathologica Communications, 6:49.*

These three selected studies represent important contributions to glioma research that I did together with Dr. Barbara Klink. However, our collaboration has been very productive leading to several other publications with researchers from different fields that I did not include in this habilitation thesis (Seifert et al. (2014); Abou-El-Ardat et al. (2017); Alfonso et al. (2017); Klapproth et al. (2018); Mäder et al. (2018); Zakrzewski et al. (2019); Biedermann et al. (2019); Seifert et al. (2020)).

3.2 Survival differences of DNMT3A-mutant acute myeloid leukemia patients

In 2016, I started to work on a project with the goal to identify molecular factors associated with survival differences of DNMT3A-mutant acute myeloid leukemia patients.

This was motivated by the SyTASC (Systems-based Therapy of AML Stem Cells) project funded by the German Cancer Aid in which I was involved to support bioinformatics data analyses for the work package of Prof. Dr. Ingo Roeder.

Acute myeloid leukemia (AML) is a highly malignant cancer of myeloid blood cells that is characterized by a rapid growth of abnormal immature myeloblasts that lost their ability to differentiate leading to the replacement of normal cells in bone marrow and blood (Döhner et al. (2015)). DNMT3A belongs to the most frequently mutated genes in AML (The Cancer Genome Atlas Research Network (2013a)), which encodes a DNA methyltransferase (Shah and Licht (2011)) that is important for normal hematopoiesis (Challen et al. (2011); Yang et al. (2015)). Mutations of DNMT3A have been associated with very poor prognosis (e.g. Ribeiro et al. (2012); Renneville et al. (2012)). Nevertheless, some DNMT3A-mutant patients have shown relatively long survival or even reached a long-term remission (Ploen et al. (2014); Sun et al. (2016)), but molecular differences distinguishing short- and long-lived patients had not been extensively studied.

To fill this gap, we considered molecular data of DNMT3A-mutant AML patients from TCGA (The Cancer Genome Atlas Research Network (2013a)) to search for subgroups with survival differences, to characterize their underlying molecular alterations and associated gene regulatory networks, and to analyze the transfer of our findings to independent cohorts. The results of our study were published as (see also Section 4.3):

- *Chris Lauber, Nádia Correia , Andreas Trumpp , Michael A. Rieger, Anna Dolnik, Lars Bullinger, Ingo Roeder and* ***Michael Seifert*** *(2020):* ***Survival differences and associated molecular signatures of DNMT3A-mutant acute myeloid leukemia patients****, Scientific Reports, 10:12761.*

This study represents the first in-depth computational approach that characterizes molecular factors associated with survival differences of DNMT3A-mutant AML patients, which could contribute to the development of robust markers for an improved patient stratification.

3.3 Impact of rare gene copy number alterations on survival of cancer patients

In 2012, I started to work on the development of computational methods to quantify impacts of gene copy number alterations on clinically relevant characteristics. This work

was motivated by the fact that more and more large-scale omics data sets of different types of cancer were published, but it was still extremely challenging to determine impacts of individual mutations. The mountains and hills of frequently mutated genes had been characterized, but it was still largely unknown which contribution the long tail of rarely mutated genes had (Vogelstein et al. (2013); Lawrence et al. (2014)). Rare mutations could act in combination with frequent mutations or they could independently contribute to tumor development. But essentially, the importance of rare mutations in comparison to frequent mutations was not known.

A major reason for this lack of knowledge was that computational methods to quantify the impacts of rare gene mutations did not exist. A groundbreaking publication by Hofree et al. (2013) showed that one can utilize existing gene or protein interaction networks in combination with network propagation to stratify highly diverse gene mutation profiles of cancer patients into homogeneous subgroups with consistent clinical behavior. A similar approach was used by Leiserson et al. (2015) to identify the impact of rare mutations on cellular pathways and protein complexes. These approaches greatly helped to better characterize the potential impact of small gene mutations (single nucleotide variations and small insertions/deletions), but they are only suboptimal for the analysis of gene copy number alterations, because they do not account for alterations of expression levels of affected genes.

To overcome this, I developed a network inference algorithm to directly learn gene regulatory networks form hundreds of paired gene copy number and expression profiles (see Section 2.5). I used these networks in combination with a specifically designed network propagation algorithm (see Section 2.7) to quantify the impacts of rare and frequent gene copy number alterations on patient survival or signaling and metabolic pathways.

I used these novel computational concepts in a first pioneer study to show that a regulatory network learned from gene expression and gene copy number data of 768 human cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. (2012)) can quantify impacts of patient-specific gene copy number alterations on patient survival. Based on an in-depth analysis of six cancer types, we revealed that rare patient-specific gene copy number alterations often had stronger effects on survival signature genes than frequent gene copy number alterations. This analysis also showed that rare gene copy number alterations are important for the prediction of survival of glioblastoma and skin cancer patients, whereas frequent gene copy number alterations were important to predict survival of lung cancer patients. Moreover, a

comparison to a closely related network-based approach showed that the integration of indirectly acting gene copy number alterations significantly improved the separation of patients into short and long survivors. The results of our study were published as (see also Section 4.4):

- ***Michael Seifert**, Betty Friedrich and Andreas Beyer (2016): **Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis**, Genome Biology, 17:204.*

The great success of this study further motivated me to develop the R package regNet to provide the network inference algorithm along with the different network propagation algorithms as user-friendly tools that enable own analyses and methodological extensions. The initial developments of the source code of regNet date back to the year 2012. Major parts of the source code were already used to learn the networks of the original works that are part of this habilitation thesis. The R package regNet was published as (see also Section 4.5):

- ***Michael Seifert** and Andreas Beyer (2018): **regNet: an R package for network-based propagation of gene expression alterations**, Bioinformatics, 34(2), 308-311.*

This package was also used to perform the driver gene candidate prediction for oligodendrogliomas (see Section 4.6) and to search for potential driver genes in radioresistant prostate cancer cell lines (see Section 4.7).

3.4 Impact of gene copy number alterations on radioresistance of prostate cancer

In 2014, Prof. Dr. Anna Dubrovska from the OncoRay TU Dresden and Dr. Barbara Klink invited me to a joint meeting to ask if I can support the analysis of gene copy number and gene expression profiles of prostate cancer cell lines. The central idea of this project was to identify candidate genes that are involved in the regulation of radioresistance of prostate cancer. It became immediately clear to me that this cannot be realized with standard bioinformatics tools, because they only had two cell lines whose radioresistant cells showed large chromosomal deletions and duplications of DNA segments in comparison to their radiosensitive parental cells. These DNA copy

number alterations were induced by irradiation due to error-prone DNA repair of double strand breaks (Mateo et al. (2017)). Many genes are usually located within these regions and alterations of their copy numbers are frequently transferred to the expression level. This further complicated the situation, because we had not only to deal with very few samples but also with hundreds or thousands of gene copy number alterations and expression changes. Thus, the search for potential driver genes associated with radioresistance of prostate cancer cell lines was comparable to finding the needle in a haystack.

I suggested to address this challenging problem with the help of a prostate-cancer specific gene regulatory network to propagate potential impacts of altered genes on known radioresistance marker genes. This was a perfect application for my network inference and network propagation algorithms (Seifert et al. (2016); Seifert and Beyer (2018)), which were still under development at that time. The results of our study were published as (see also Section 4.7):

- ***Michael Seifert****, Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner, Barbara Klink and Anna Dubrovska (2019):* ***Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse****, PLoS Computational Biology, 15(11):e1007460.*

Our computational approach enabled us to predict 14 potential driver candidates that were able to distinguish irradiated prostate cancer patients into early and late relapse groups. In-depth wet lab validation studies of one driver candidate further confirmed the value of our approach.

4 Original works

This chapter contains the seven original works that I selected for my habilitation thesis. I contributed substantially to each of these works, which is also highlighted by the fact that I am either the first author or the last author of these different studies. I am also the corresponding author of each of these publications.

Each original work is introduced by a brief overview of the bibliographic information followed by a brief motivation of the work in the specific scientific context and a summary of the main results of the publication. This overview is completed by a statement that provides detailed information about my contribution to each publication.

Additional information in the form of *Supplementary Materials* have been prepared for each original work, but these materials are not reproduced within this habilitation thesis for reasons of brevity. These materials are available from the corresponding website of the publisher of each original work or they can also be obtained from me upon request.

4.1 Publication:

Comparative transcriptomics reveals similarities and differences between astrocytoma grades

Journal: BMC Cancer

Received: 17 July 2015; **Accepted**: 1 November 2015; **Published**: 16 December 2015

Citation: Michael Seifert, Martin Garbe, Betty Friedrich, Michel Mittelbronn and Barbara Klink (2015): Comparative transcriptomics reveals similarities and differences between astrocytoma grades. BMC Cancer, 15:952.

Placement and summary of the publication

Astrocytomas represent the most frequently diagnosed primary human brain tumors in the course of life (Ohgaki and Kleihues (2005)). At the time of this study, astrocytomas were classified by the World Health Organization (WHO) brain tumor classification system into four histological grades of increasing malignancy (Louis et al. (2007)). In this study, we focused on a systematic computational comparison of molecular data of pilocytic astrocytomas (PA I), diffuse astrocytomas (AS II), anaplastic astrocytomas (AS III) and glioblastomas (GBM IV) representing the most frequently occurring astrocytomas. PA I represents a well-circumscribed slowly growing astrocytoma that is predominantly occurring in childhood and adolescence, whereas AS II, AS III and GBM IV almost exclusively occur in adults and are further characterized by infiltrative growth into the surrounding brain tissue (Louis et al. (2007); Ohgaki and Kleihues (2009); Tonn et al. (2005); Jones et al. (2012)). Molecular analysis of individual astrocytoma grades already revealed insights into genetic, transcriptomic and epigenetic alterations (e.g. Jones et al. (2013); The Cancer Genome Atlas Research Network (2008); Verhaak et al. (2010); Noushmehr et al. (2010); Deshmukh et al. (2011); Wang et al. (2013)), but studies that compared all four astrocytoma grades did not exist with few exceptions performed in the early 2000s (Rickman et al. (2001); Hunter et al. (2002); Rorive et al. (2006)). This provided an excellent basis

for our comprehensive study to systematically compare all four astrocytoma grades.

Our computational study revealed similarities and differences between astrocytoma grades at the level of individual genes, signaling pathways and regulatory networks. In comparison to normal brain, the number of differentially expressed genes generally increased with the astrocytoma grade with one major exception. Interestingly, the cytokine receptor pathway showed nearly the same number of differentially expressed genes in PA I and GBM IV. Further analyses revealed a strong exclusive overexpression of CX3CL1 (fractalkine) and its receptor CX3CR1 in PA I that possibly contributes to the absence of invasive growth. Moreover, surprisingly, we found that PA I was significantly associated with the mesenchymal subtype typically observed for GBM IV. Expression of endothelial and mesenchymal markers (e.g. ANG2, CHI3L1, THBD) indicated a stronger contribution of the micro-environment to the manifestation of the mesenchymal subtype than the tumor biology itself. In accordance with this, we further confirmed by immunohistochemistry that PA I and GBM IV showed ANG2-positive endothelial cells in regions with activated blood vessels, a feature that was largely absent in AS II and AS III. We also inferred a transcriptional regulatory network associated with specific expression differences between PA I and AS II, AS III and GBM IV. Major transcriptional regulators were involved in brain development, cell cycle control, proliferation, apoptosis, chromatin remodeling or DNA methylation. Many of these regulators showed directly underlying DNA methylation changes in PA I or gene copy number alterations in AS II, AS III and GBM IV.

Our study identified similarities and differences between all four astrocytoma grades. We confirmed already known characteristics and further revealed novel insights into astrocytoma biology. Therefore, our findings represent a valuable resource for future computational and experimental studies to further disentangle molecular mechanisms that drive the infiltrative growth of diffuse astrocytomas.

Author contribution

I designed the concept of the study and contributed substantially to the computational analysis. I supervised the gene expression analysis and the network inference that were done as part of a master thesis by Martin Garbe. I further supervised the acquisition and curation of the different pathway and gene annotations by Betty Friederich and Martin Garbe. I established the collaboration to Michel Mittelbronn, who did the validation of the Ang2 expression and localization by immunohistochemistry. I discussed all findings with Barbara Klink, who supported the biological interpretation of our findings. I wrote the manuscript, created the figures and performed the revision of the manuscript.

Seifert *et al. BMC Cancer* (2015) 15:952
DOI 10.1186/s12885-015-1939-9

BMC Cancer

RESEARCH ARTICLE Open Access

Comparative transcriptomics reveals similarities and differences between astrocytoma grades

Michael Seifert[1,2,5*], Martin Garbe[1], Betty Friedrich[1,3], Michel Mittelbronn[4] and Barbara Klink[5,6,7]

Abstract

Background: Astrocytomas are the most common primary brain tumors distinguished into four histological grades. Molecular analyses of individual astrocytoma grades have revealed detailed insights into genetic, transcriptomic and epigenetic alterations. This provides an excellent basis to identify similarities and differences between astrocytoma grades.

Methods: We utilized public omics data of all four astrocytoma grades focusing on pilocytic astrocytomas (PA I), diffuse astrocytomas (AS II), anaplastic astrocytomas (AS III) and glioblastomas (GBM IV) to identify similarities and differences using well-established bioinformatics and systems biology approaches. We further validated the expression and localization of Ang2 involved in angiogenesis using immunohistochemistry.

Results: Our analyses show similarities and differences between astrocytoma grades at the level of individual genes, signaling pathways and regulatory networks. We identified many differentially expressed genes that were either exclusively observed in a specific astrocytoma grade or commonly affected in specific subsets of astrocytoma grades in comparison to normal brain. Further, the number of differentially expressed genes generally increased with the astrocytoma grade with one major exception. The cytokine receptor pathway showed nearly the same number of differentially expressed genes in PA I and GBM IV and was further characterized by a significant overlap of commonly altered genes and an exclusive enrichment of overexpressed cancer genes in GBM IV. Additional analyses revealed a strong exclusive overexpression of CX3CL1 (fractalkine) and its receptor CX3CR1 in PA I possibly contributing to the absence of invasive growth. We further found that PA I was significantly associated with the mesenchymal subtype typically observed for very aggressive GBM IV. Expression of endothelial and mesenchymal markers (ANGPT2, CHI3L1) indicated a stronger contribution of the micro-environment to the manifestation of the mesenchymal subtype than the tumor biology itself. We further inferred a transcriptional regulatory network associated with specific expression differences distinguishing PA I from AS II, AS III and GBM IV. Major central transcriptional regulators were involved in brain development, cell cycle control, proliferation, apoptosis, chromatin remodeling or DNA methylation. Many of these regulators showed directly underlying DNA methylation changes in PA I or gene copy number mutations in AS II, AS III and GBM IV.

Conclusions: This computational study characterizes similarities and differences between all four astrocytoma grades confirming known and revealing novel insights into astrocytoma biology. Our findings represent a valuable resource for future computational and experimental studies.

Keywords: Astrocytoma grades, Pilocytic astrocytoma, Diffuse astrocytoma, Anaplastic astrocytoma, Glioblastoma

*Correspondence: michael.seifert@tu-dresden.de
[1]Innovative Methods of Computing, Center for Information Services and High Performance Computing, Dresden University of Technology, Dresden, Germany
[2]Cellular Networks and Systems Biology, University of Cologne, CECAD, Cologne, Germany
Full list of author information is available at the end of the article

Background

Astrocytomas are the most common primary brain tumors in the course of life [1]. Molecular origins of astrocytomas are not fully understood. Different studies have identified tumorigenic cells with stem-cell-like properties suggesting that astrocytomas originate from neural stem cells [2, 3]. Astrocytomas are classified by the World Health Organization (WHO) grading system into four histological grades of increasing malignancy [4]. Here, we focus on a comparative analysis of the most frequently occurring astrocytomas (pilocytic astrocytoma, diffuse astrocytoma, anaplastic astrocytoma, glioblastoma) of different degrees of aggressiveness to assess for similarities and differences at the level of individual genes, signaling pathways, molecular subtypes and regulatory networks. This is highly important to better understand the development of specific astrocytomas.

The pilocytic astrocytoma WHO grade I (PA I) is a very slowly growing benign astrocytoma. PA I is the most commonly diagnosed brain tumor in childhood and adolescence [5]. The ten-year overall survival rate of PA I patients is greater than 95 % [1]. The treatment of choice for PA I is gross total resection, but PA I tumors that are inoperable or only partly accessible by surgery represent a therapeutic challenge often showing a serve clinical course [6, 7]. Recent studies have indicated that PA I is predominantly a single-pathway disease driven by mutations affecting the MAPK pathway [5, 7]. In addition, PA I can also display histological features of glioblastoma (GBM IV) including microvascular proliferation and necrosis, but in contrast to GBM IV, these features are not directly associated with increased malignancy of PA I [8]. In rare cases, progression of PA I to more malignant astrocytomas has been observed [9].

In contrast to PA I, astrocytomas of WHO grade II to IV almost exclusively occur in adults. These astrocytomas are characterized by a diffuse infiltrating growth into the surrounding brain tissue that is absent in PA I. Therefore, AS II, AS III and GBM IV are also referred to as diffuse gliomas.

The diffuse astrocytoma WHO grade II (AS II) is a slowly growing invasive semi-benign astrocytoma. AS II is frequently diagnosed in young adults between 20 and 45 years with an average age of 35 years [10]. The diffuse invasive growth of AS II with no clearly identifiable boarder between tumor and normal tissue makes complete surgical resection almost impossible [11]. Recurrences of tumors are observed in most patients after few years with progression to more malignant AS III or GBM IV in many cases [12–14]. The median survival of AS II patients is between five to eight years [15].

The anaplastic astrocytoma WHO grade III (AS III) is an invasively and faster growing malignant astrocytoma. AS III is characterized by increased mitotic activity and more variable size and shape of tumor cells in comparison to AS II [4]. The average age of patients diagnosed with AS III is 45 years. When possible, surgical resection followed by radiotherapy and/or chemotherapy is the treatment of choice. Similar to AS II, progression of AS III to the most malignant GBM IV is frequently observed [13, 14]. The overall five-year survival rate of AS III patients is 24 % [16] and the median survival is between one to four years [17].

The glioblastoma WHO grade IV (GBM IV) is the most malignant astrocytoma [4]. GBM IV is a very fast invasively growing tumor. In contrast to AS III, GBM IV also shows necrosis and/or vascular proliferation. Two genetically distinct GBM IV classes are known: (i) secondary GBMs that develop progressively over several years from less malignant AS II or AS III, and (ii) primary GBMs that develop within few months without prior occurrences of lower grade astrocytomas [12, 13]. Only about 5 % of GBM IV cases are secondary GBMs [18]. Patients diagnosed with a secondary GBM are on average younger than primary GBM patients (45 vs. 62 years) [12]. Primary and secondary GBMs are histologically indistinguishable. IDH mutations in secondary GBMs enable a distinction from primary GBMs at the molecular level [19]. These IDH1 or IDH2 mutations are already present in less malignant AS II and AS III [20]. The treatment of choice is surgical resection in combination with radiation and chemotherapy. This intensive treatment increases the average survival of GBM IV patients to about 15 months [21] compared to 13 weeks for surgery alone [22]. Less than 5 % of patients survive longer than five years [18].

Over the last years, rapid advances in experimental technologies have enabled detailed molecular analyses of large cohorts of different types of astrocytomas that provided new insights into pathological mechanisms [5, 7, 19, 23, 24], molecular subtypes [25–27], alterations of signaling pathways [23, 24, 28], or activities of transcriptional regulatory networks [29–33]. Other studies have focused on the characterization of differences between astrocytoma grades to better understand pathogenic impacts of molecular alterations. Differential expression of immune defense genes in PA I in comparison to AS II with potential indications toward benign behavior of PA I have been reported [34]. Characteristic expression of anti-migratory genes has been found in PA I in comparison to AS II, AS III and GBM IV putatively contributing to the compact, well-circumscribed growth of PA I in contrast to the infiltrative growth of higher-grade astrocytomas [35]. Further molecular markers distinguishing PA I from AS II, AS III and GBM IV have been reported in [36, 37]. A comparative analysis of AS II, AS III and GBM IV has revealed greater regulatory network dysregulation associated with increasing astrocytoma grade [33]. Additionally, mutational patterns associated with the origin and chemotherapy therapy-driven evolution of recurrent

secondary gliomas have recently been reported [14]. All these and many other studies have greatly contributed to a better understanding of astrocytoma development hopefully contributing to urgently needed new therapeutic strategies in the near future.

However, most studies have only focused on the identification of differences between astrocytoma grades. This is of course very important to better understand molecular mechanisms associated with aggressiveness of different astrocytoma grades and to reveal novel grade-specific therapeutic targets. On the other hand, still only little is known about commonly altered genes, shared molecular subtypes, common alterations in signaling or metabolic pathways, or activities of major transcriptional regulators. More detailed information about these regulatory mechanisms is also very important to further increase our knowledge about astrocytoma development and may reveal unexpected similarities between astrocytoma grades.

Here, we utilize publicly available molecular data of astrocytomas to systematically characterize similarities and differences of all four astrocytoma grades. In more detail, we characterize transcriptional alterations at the level of individual genes and known molecular pathways. We analyze all four astrocytoma grades for their association with known molecular subtypes and utilize immunohistochemistry to validate Ang2 as a marker gene predicted to distinguish PA I and GBM IV from AS II and AS III. We further determine a regulatory network that distinguishes PA I from AS II, AS III and GBM IV revealing major transcriptional regulators and directly underlying mutations putatively associated with pathobiological differences.

Methods

No ethical approval was required for this study. All utilized public omics data sets were generated by others who obtained ethical approval.

Molecular data of PA I

We considered raw gene expression data of 49 PA I and 9 normal cerebellum reference samples (5 fetal and 4 adult samples) available from Gene Expression Omnibus (GSE44971) [38]. We performed stringent quality controls of all expression arrays by reconstructing the hybridization images. We removed three arrays with slight hybridization artifacts. The remaining samples are listed in Additional file 1: Table S1. All corresponding microarrays were normalized using GCRMA [39] with a design file from BrainArray (HGU133Plus2 version 15.0.0). The resulting PA I gene expression data set comprised 47 PA I samples and 8 corresponding normal cerebellum references for which expression levels were measured for 16,973 genes. We further also downloaded processed DNA methylation profiles available for 38 of the considered PA I samples (GSE44684) analyzed in [38]. Tumor-specific DNA methylation profiles were compared to DNA methylation profiles of normal cerebellum samples from four fetal and two adult probes. We refer to [38] for more details. All PA I tumors were diagnosed in children or young adults (Additional file 2: Figure S1) and fulfill all editorial policies (ethical approval and consent, standards of reporting, data availability).

Molecular data of AS II, AS III and GBM IV

We considered raw gene expression and gene copy number data of AS II, AS III, GBM IV and adult normal brain references from epilepsy patients from the Repository for Molecular Brain Neoplasia Data (Rembrandt, release 1.5.9) [40]. The non-tumor samples from Rembrandt were already used as references for the analysis of AS II, AS III and GBM IV tumors in [41]. We again performed stringent quality controls and removed all patient or reference samples where expression or copy number microarrays had hybridization artifacts. See Additional file 1: Table S1 for considered samples. The remaining gene expression samples were further normalized as previously described for PA I. This resulted in a gene expression data set that comprised 16 AS II, 17 AS III, 45 GBM IV and 21 corresponding normal adult brain references from epilepsy patients for which expression levels were measured for 16,973 genes. Processing of corresponding gene copy number data was more complex (Additional file 2: Text S1). The majority of tumors was diagnosed in older adults. The age at diagnosis tended to increase with the WHO grades of the tumors (Additional file 2: Figure S1). All data sets fulfill the editorial policies (ethical approval and consent, standards of reporting, data availability).

Identification of differentially expressed genes

We performed t-tests to identify under- and overexpressed genes for each type of astrocytoma (PA I, AS II, AS III, GBM IV) under consideration of the corresponding normal brain references. We corrected for multiple testing by computing FDR-adjusted p-values (q-values) for all genes [42] and considered for each type of astrocytoma all genes with q-values below 0.0001 as differentially expressed in tumor compared to normal brain tissue. We further used the sign of the average gene-specific log-ratio of tumor versus normal to specify which of these genes were under- (negative sign) and overexpressed (positive sign) in each specific type of astrocytoma. See Additional file 1: Table S2 for t-test results obtained for all four astrocytoma grades. Further, we note that the considered astrocytoma types represent a heterogeneous group of tumors. PA I is often localized in the cerebellum of children or young adults, whereas AS II, AS III and GBM IV are mainly occurring in the cerebrum of adults. Thus,

it is hard to specify a common normal brain reference that would perfectly fit to all astrocytoma types with respect to their different tumor locations and age incidences. Therefore, we decided to analyze all astrocytomas under consideration of the normal brain references that were used in the corresponding initial publications (see [38] for PA I and [40, 41] for AS II, AS III and GBM IV). With the choice of these references we try to control for the heterogeneity of the astrocytoma grades to identify differences in astrocytoma-specific gene expression in comparison to the surrounding normal brain tissue in which these tumors are typically diagnosed. That is, PA I was analyzed with respect to normal cerebellum. Normal brain references from epilepsy patients were considered for the analysis of AS II, AS III and GBM IV. Note that this choice of references does not exclude that some of the differentially expressed genes that distinguish PA I from AS II, AS III and GBM IV may only occur because of expression differences in the corresponding references. However, considering both references, we found a significant positive correlation between average gene expression levels of normal cerebellum and normal brain from epilepsy patients ($r = 0.874$, $P < 2.2 \times 10^{-16}$). This indicates that the majority of genes has very similar expression profiles in both astrocytoma type-specific references. Thus, the used normal brain references should represent a good compromise to account for the location- and age-specific heterogeneity distinguishing PA I from AS II, AS III and GBM IV.

Molecular subtype classification

We downloaded the Verhaak gene expression signatures of 840 genes (ClaNC840_centroids.xls) available from [25] to determine the similarity of each individual astrocytoma to four known molecular subtypes (neural, proneural, classical, mesenchymal). We identified that 757 of these 840 signature genes were also measured in each of our PA I, AS II, AS III and GBM IV samples. For each of these samples, we first computed for each of the 757 genes its relative expression level ($\log_2$-ratio) in tumor compared to its average expression in normal brain. Next, we computed the correlations of these 757 sample-specific expression levels with the corresponding expression levels of the four molecular subtypes. We further tested if the correlation of an individual sample with a specific subtype was significantly greater than zero (Pearson's product moment correlation test). We finally assigned each astrocytoma sample to the Verhaak-subtype with the greatest significant positive correlation ($P < 0.05$).

Molecular signature distinguishing PA I from AS II, AS III and GBM IV

We determined a molecular gene signature that distinguished PA I from AS II, AS III and GBM IV using the previously identified differentially expressed genes. To realize this, we considered each gene that was (i) underexpressed in PA I but not in AS II, AS III or GBM IV, (ii) unchanged in PA I but not in AS II, AS III or GBM IV, or (iii) overexpressed in PA I but not in AS II, AS III or GBM IV. Then, we considered this reversely and determined each gene that was (iv) underexpressed in AS II, AS III or GBM IV but not in PA I, (v) unchanged in AS II, AS III or GBM IV but not in PA I, or (vi) overexpressed in AS II, AS III or GBM IV but not in PA I. All genes that passed one of these criteria showed characteristic expression differences comparing PA I against AS II, AS III or GBM IV. We further only focused on signature genes with strong expression differences and removed all genes with an average gene expression difference below two comparing both classes. This resulted in 1,089 signature genes distinguishing PA I from AS II, AS III and GBM IV. See Additional file 1: Table S3 for obtained signature genes and their average gene expression log-ratios of tumor versus normal.

Signature-specific regulatory network inference

We considered gene-specific sub-network inference problems to derive a transcriptional regulatory network associated with the expression of molecular signature genes distinguishing PA I from AS II, AS III and GBM IV. Therefore, we focused on the expression levels of $N = 1,089$ signature genes in our data set of in total $D = 125$ astrocytomas. For each signature gene $i \in \{1, \ldots, N\}$, we assumed that its expression level e_{id} in an astrocytoma $d \in \{1, \ldots, D\}$ can be predicted by a linear combination

$$e_{id} = \sum_{j \in \mathtt{TF} \setminus \{i\}} a_{ji} \cdot e_{jd} \quad (1)$$

of the expression levels e_{jd} of transcriptional regulators $j \in \mathtt{TF} \setminus \{i\}$ that were part of the molecular signature that distinguishes PA I from AS II, AS III and GBM IV. Here, $\mathtt{TF}$ defines the subset of genes in the molecular signature that were annotated as TFs (151 of 1,089). The expression level e_{id} of each gene i in an astrocytoma d is given by the $\log_2$-ratio of the expression level of gene i in astrocytoma d in comparison to the expression level of gene i in the corresponding average normal brain reference. The unknown parameters of this signature gene-specific linear model are given by $\vec{a}_i := (a_{ji})_{j \in \mathtt{TF} \setminus \{i\}}$. Each individual parameter $a_{ji} \in \mathbb{R}$ quantifies the impact of the expression level of regulator j on the expression level of signature gene i: (i) $a_{ji} < 0$ specifies that TF j is a putative inhibitor of gene i, (ii) $a_{ji} > 0$ defines that TF j is a putative activator of gene i, and (iii) $a_{ji} = 0$ means that no dependency between j and i exists. We used lasso (least absolute shrinkage and selection operator) regression [43] in combination with a recently developed significance test for lasso [44] to estimate each a_{ji} and its corresponding significance for Eq. (1). This enabled us to select the most relevant putative

regulators of each signature gene (Additional file 1: Table S4, $P < 5 \times 10^{-5}$). Details are provided in Additional file 2: Text S2. We further validated the predictive power of the obtained regulatory network on independent astrocytoma data sets (Additional file 2: Text S4, Figure S7) and we also evaluated the putative proportion of included direct TF-target gene interactions (Additional file 2: Text S5, Figure S8). All these validation studies clearly indicated that the regulatory network included relevant TF-target gene links to predict the expression levels of signature genes based on the expression profiles of TFs.

Gene annotations

We utilized different public resources to create a comprehensive summary of cancer-relevant gene annotations for the analysis of differentially expressed genes. This comprised genes annotated of TFs/cofactors, kinases, phosphatases, signaling pathway genes, metabolic pathway genes, oncogenes, tumor suppressor genes, cancer census genes, and genes essential for cell survival. Details and references are provided in Additional file 1: Table S5. Additional studies of gene functions were done using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and GeneCards (http://www.genecards.org/).

Results and discussion

Transcriptional alterations increase with WHO grade

We first globally analyzed PA I, AS II, AS III and GBM VI and found that the number of differentially expressed genes increased significantly with increasing WHO grade ($r = 0.92$, $P = 0.04$, Pearson's product moment correlation). Corresponding statistics are shown in Fig. 1a for each type of astrocytoma. Compared to PA I known to have the best prognosis, AS II and AS III showed a nearly two-fold increase in differentially expressed genes. A nearly four-fold increase was observed for GBM IV representing the most malignant astrocytoma. We also observed that the number of overexpressed genes in PA I was more than two-fold higher than the number of underexpressed genes. This was much more balanced for AS II and AS III. Similar to PA I, GBM IV also showed clearly more over- than underexpressed genes. The global tendencies remained highly similar but the numbers of differentially expressed genes were clearly reduced when we further restricted the identified genes to those with strong expression changes of absolute $\log_2$-fold-changes greater than two compared to normal brain (Fig. 1a).

Next, we analyzed the identified differentially expressed genes in the context of functional categories or cellular processes known to be involved in cancer. Therefore, we first used data from different public resources to define nine cancer-relevant categories containing genes that are essential for cell survival, oncogenes, tumor suppressor genes, cancer census genes, phosphatases, kinases, metabolome genes, signaling pathway genes, and transcriptional regulators (Additional file 1: Table S5). We then determined for each category the overlap with the differentially expressed genes identified for each type of astrocytoma. Again, we found that the numbers of differentially expressed genes in each category increased significantly with the WHO grades ($r > 0.91$, $P < 0.043$ for all categories, Pearson's product moment correlation). A statistic representing the number of differentially expressed genes in each of these categories for each type of astrocytoma is shown in Fig. 1b. Genes essential for cell survival, phosphatases, and kinases were only significantly overrepresented in AS II, AS III and GBM IV. Oncogenes were enriched in PA I, AS III and GBM IV, whereas tumor suppressor genes were only enriched in AS III and GBM IV. Additionally, cancer census genes [45] and genes that were part of known cancer-relevant signaling pathways were only significantly overrepresented in GBM IV. Although not significantly enriched, we observed several differentially expressed metabolic pathway genes, even more differentially expressed cancer-relevant signaling pathway genes, and many differentially expressed transcriptional regulators in all astrocytoma grades with numbers of affected genes again increasing from PA I to GBM IV (Fig. 1b).

Finally, we further extended the previous analysis to distinguish between under- and overexpressed genes (Additional file 2: Figure S2). No enrichment of underexpressed genes was observed for essential and signaling pathway genes in all four astrocytoma grades. Underexpressed genes annotated as oncogenes, tumor suppressor genes, cancer census genes or transcriptional regulators were significantly enriched in PA I. Phosphatases and kinases were significantly overrepresented among underexpressed genes in AS II, AS III and GBM IV. Underexpressed metabolome genes were only significantly enriched in GBM IV. Further, no significant enrichment of overexpressed genes was observed for phosphatases, kinases and metabolome genes in all four astrocytoma grades. Overexpressed oncogenes were significantly overrepresented in AS II and AS III. Transcriptional regulators, tumor suppressors and cancer census genes were significantly enriched for overexpressed genes in AS II, AS III and GBM IV. Overexpressed signaling pathway genes were significantly enriched in all four astrocytoma grades.

Verhaak classification reveals strong association of PA I with mesenchymal subtype

Classification of astrocytomas according to known molecular subtypes is important to improve treatment decisions and prognosis. Four major subtypes of GBM IV were first revealed in [25] and later also identified in AS II and AS III [27]. This has been widely applied to classify individual

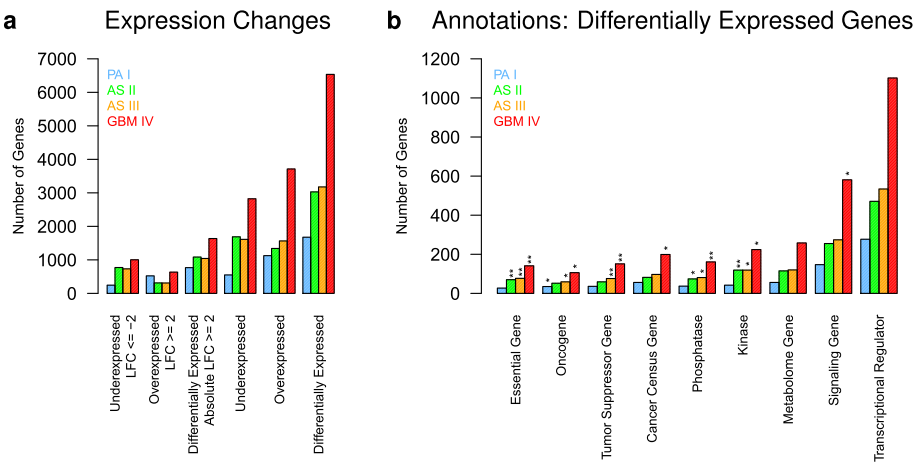

Fig. 1 Expression changes and functional categorization of differentially expressed genes for different astrocytoma grades. **a**, Number of differentially expressed genes identified for each type of astrocytoma in comparison to normal brain references at an FDR of 0.0001. An additional log-fold-change cutoff (LFC) of two was used for the first three categories to focus on genes with strong expression changes. **b**, Number of differentially expressed genes annotated in selected functional categories: essential genes, oncogenes, tumor suppressor genes, cancer census genes, phosphatases, kinases, metabolome pathway genes, signaling pathway genes, and transcriptional regulators (see Methods for details). Significant enrichment of genes in a category within a tumor type is represented by '*' ($P < 0.05$) and '**' ($P < 0.01$) (Fisher's exact test)

AS II, AS III and GBM IV tumors either as neural, proneural, classical or mesenchymal, but so far it has not been tested if one or more of these subtypes are also associated with PA I. Therefore, we used the Verhaak-classifier [25] to compute the correlation between the given signature-specific expression levels of the Verhaak-subtypes and the corresponding gene expression levels of each individual astrocytoma. Correlations of each individual PA I, AS II, AS III and GBM IV tumor with the four Verhaak-subtypes are shown in Fig. 2 and provided in Additional file 1: Table S6.

Interestingly, all PA I tumors showed very homogeneous correlation profiles resulting in a significant association with the mesenchymal subtype (Fig. 2a, $r > 0.14$, $P < 2.14 \times 10^{-5}$ for all PA I, Pearson's product moment correlation). We further confirmed this observation for an independent PA I cohort [46], where again 40 of 41 PA I tumors were significantly correlated with the mesenchymal subtype (Additional file 2: Figure S3, $r > 0.17$, $P < 4.5 \times 10^{-7}$ for all PA I). The mesenchymal subtype was observed to be strongly associated with cultured astroglial cells that showed high expression of microglia markers [25]. Additionally, PA I was reported to show increased microglia proliferation in comparison to AS II, AS III and GBM IV [47]. This indicates that the strong association of PA I with the mesenchymal subtype may at least in part be explained with the role of the microglia. To analyze this, we first identified that 16 microglia/macrophage marker genes from [48] were part of the Verhaak-classifier (Additional file 1: Table S7). Next, we used these genes and found a significant positive correlation between the average expression levels of microglia/macrophage marker genes in PA I and corresponding mesenchymal subtype expression levels from Verhaak ($r = 0.56$, $P < 0.013$). This trend was also observed for AS II, AS III and GBM IV average marker expression profiles ($r > 0.58$, $P < 0.009$) and also for individual AS II, AS III and GBM IV tumors that were not classified as mesenchymal (Additional file 1: Table S7). Thus, additional pathobiological features such as microvascular proliferation and necrosis most likely contribute to the strong association of PA I with mesenchymal subtype.

Microvascular proliferation and necrosis were described as common features of PA I and GBM IV [8]. Also increased necrosis was reported for the mesenchymal subtype [25]. We observed that ANGPT2 (alias ANG2), an endothelial cell marker involved in angiogenesis [49], had significantly higher expression levels in PA I and GBM IV than in AS II or AS III in comparison to normal brain (Additional file 1: Table S2). Interestingly, these astrocytoma grade-specific expression profile of ANGPT2 was highly correlated with that of the endothelial cell marker THBD ($r = 0.86$, $P = 0.07$), which is part of the Verhaak signature. In contrast to THBD, ANGPT2

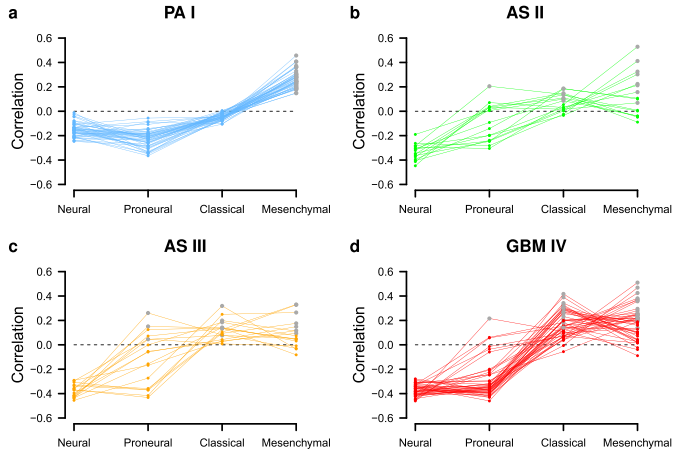

Fig. 2 Classification of individual astrocytoma patients according to known molecular subtypes. Classification of PA I, AS II, AS III and GBM IV patients according to molecular subtypes (Neural, Proneural, Classical, Mesenchymal) defined by Verhaak [25]. Correlations between patient-specific expression levels of Verhaak signature genes and each of the four subtype-specific Verhaak signatures were computed for each patient. Colored curves represent the obtained patient-specific correlations. A grey dot within each patient-specific curve highlights the assigned Verhaak-subtype to which the underlying patient had the strongest positive correlation. The subfigures **a** to **d** show the results for individual PA I, AS II, AS III, and GBM IV patients

is not part of the Verhaak signature, but this positive correlation indicates that microvascular proliferation and necrosis may contribute to the mesenchymal classification obtained for all PA I and many GBM IV tumors. To further test this, we confirmed by immunohistochemistry that PA I and GBM IV showed Ang2-positive endothelial cells (protein expression) in regions with activated blood vessels, a feature that was largely absent in AS II and AS III (Additional file 2: Figure S4, Text S3). We also found that the expression of the mesenchymal marker CHI3L1 [25] was highly correlated with the expression of ANGPT2 ($r = 0.89$, $P < 0.06$). Thus, this all indicates that several different factors contribute to the strong association of PA I with the mesenchymal subtype. In addition, the micro-environment may have a stronger contribution on these subtype-characteristics than the distinct aggressiveness of mostly benign PA I and highly malignant GBM IV tumor cells.

The Verhaak-classification of AS II, AS III and GBM IV was clearly more heterogeneous revealing few proneural, some classical and many mesenchymal astrocytomas in each class (Fig. 2b–d). The neural subtype was clearly underrepresented in the considered cohorts. Only one PA I tumor from [46] was classified as neural with marginally higher significance than for mesenchymal (Additional file 1: Table S6).

The Verhaak-classification scheme has been further refined by a hypermethylator subtype predominantly observed within a subgroup of proneural astrocytomas [26]. A specific mutation of IDH1 frequently found in AS II, AS III and secondary GBM IV has been shown to be a key driver of this subtype [50]. We used the gene expression signature of the hypermethylator subtype (Table 2 in [26]) to determine the correlation of each of our astrocytoma samples with this subtype. As expected, PA I and the majority of our GBM IV tumors, both typically lacking IDH1 mutations, were negatively correlated with the hypermethylator subtype, whereas the majority of AS II and AS II showed positive correlations (Additional file 2: Figure S5).

Specific patterns of differential expression characterize similarities and differences of different astrocytomas

Besides the observed molecular heterogeneity between and within the different astrocytoma types, we next aimed at the identification of core sets of genes that were commonly under- or overexpressed in different astrocytoma subsets. We therefore considered all differentially expressed genes identified for PA I, AS II, AS III and GBM IV and utilized Venn diagrams to quantify the numbers of genes that were exclusively present in specific subsets of these types of astrocytomas (Fig. 3). Expression states of

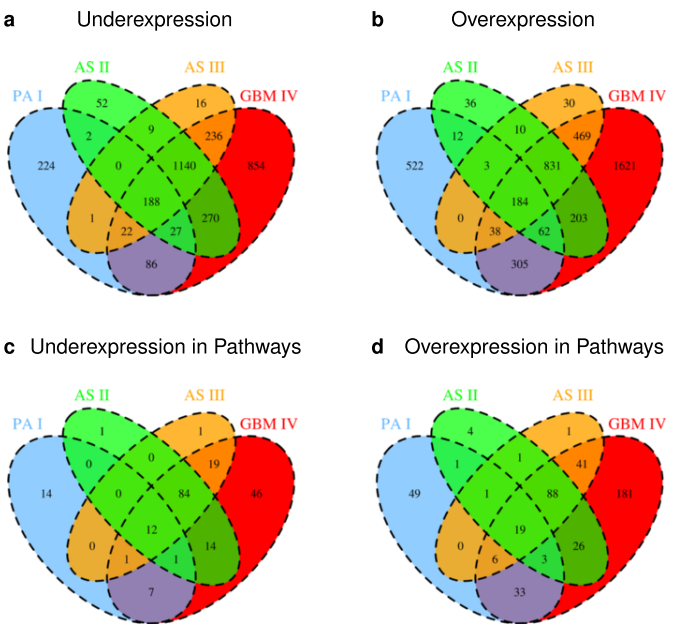

Fig. 3 Comparison of genes with altered expression in different astrocytoma grades. Venn diagrams were used to quantify commonalities and differences between differentially expressed genes identified for each type of astrocytoma in comparison to normal brain. **a**, Underexpressed genes. **b**, Overexpressed genes. **c**, Underexpressed cancer signaling pathway genes. **d**, Overexpressed cancer signaling pathway genes

individual genes for all types of astrocytomas are provided in Additional file 1: Table S2. We observed that the number of commonly under- or overexpressed genes in AS II, AS III and GBM IV were substantially increased in comparison to any intersection of PA I with two more malignant astrocytoma grades (Fig. 3a–b, e.g. 1140 under- and 831 overexpressed genes in common between AS II, AS III and GBM IV vs. 27 under- and 62 overexpressed genes in common between PA I, AS II and GBM IV). Additionally, AS II and AS III alone also shared many more commonly under- or overexpressed genes with GBM IV than with PA I (e.g. 270 under- and 203 overexpressed genes in common between AS II and GBM IV vs. 2 under- and 12 overexpressed genes in common between AS II and PA I). Interestingly, there was a strong exclusive overlap of 86 under- and 305 overexpressed genes in common between PA I and GBM IV that contained substantially more genes than observed between PA I and AS II or PA I and AS III. These different general tendencies were also observed when we exclusively focused on known cancer signaling pathway genes (Fig. 3c–d).

We further analyzed which genes were commonly under- or overexpressed in each of the four specific astrocytoma grades and in different subsets of astrocytoma grades (Fig. 3). We also investigated which molecular processes were regulated by subset-specific genes using GOrilla [51]. Since there were so many transcriptomic changes comparing astrocytomas to normal brain tissue, we only report details for some well-known or potentially interesting genes. We further refer to Additional file 1: Table S2 listing the expression states of all genes in specific astrocytoma subsets. In addition, we have summarized all discussed genes that were exclusively differentially expressed in PA I, AS II, AS III or GBM IV in Table 1.

Selected genes exclusively observed in PA I Considering genes that were exclusively differentially expressed in PA I, we observed several under- (e.g. EN2, EOMES, MEIS1, NEUROD1, ZIC1, ZIC2, ZIC3, ZIC4) and overexpressed (e.g. EGR1, EGR3, OLIG1) TFs involved in brain development. For example, EOMES is involved in neuron division and/or migration [52]. Additionally, three known

Table 1 Selected genes predicted to be differentially expressed in a specific astrocytoma grade

Gene	Chromosome	Band	Expression	Tumor	Annotation
H3F3A	1	q42.12	-	PA I	H3 histone, family 3A
MEIS1	2	p14	-	PA I	Meis homeobox 1
NEUROD1	2	q31.3	-	PA I	neuronal differentiation 1
EOMES	3	p24.1	-	PA I	eomesodermin
ZIC1	3	q24	-	PA I	Zic family member 1
ZIC4	3	q24	-	PA I	Zic family member 4
EGR1	5	q31.2	+	PA I	early growth response 1
EN2	7	q36.3	-	PA I	engrailed homeobox 2
EGR3	8	p21.3	+	PA I	early growth response 3
CDKN2B	9	p21.3	+	PA I	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
NTRK2	9	q21.33	+	PA I	neurotrophic tyrosine kinase, receptor, type 2
HIF1AN	10	q24.31	+	PA I	hypoxia inducible factor 1, alpha subunit inhibitor
SUV420H1	11	q13.2	-	PA I	suppressor of variegation 4-20 homolog 1 (Drosophila)
KRAS	12	p12.1	-	PA I	Kirsten rat sarcoma viral oncogene homolog
ZIC2	13	q32.3	-	PA I	Zic family member 2
SUZ12	17	q11.2	-	PA I	SUZ12 polycomb repressive complex 2 subunit
SUV420H2	19	q13.42	-	PA I	suppressor of variegation 4-20 homolog 2 (Drosophila)
OLIG1	21	q22.11	+	PA I	oligodendrocyte transcription factor 1
OLIG2	21	q22.11	+	PA I	oligodendrocyte lineage transcription factor 2
ATRX	X	q21.1	-	PA I	alpha thalassemia/mental retardation syndrome X-linked
ZIC3	X	q26.3	-	PA I	Zic family member 3
FAM110C	2	p25.3	-	AS II	family with sequence similarity 110, member C
HEY2	6	q22.31	+	AS II	hes-related family bHLH transcription factor with YRPW motif 2
NR2E1	6	q21	-	AS II	nuclear receptor subfamily 2, group E, member 1
EYA1	8	q13.3	+	AS II	EYA transcriptional coactivator and phosphatase 1
GAS2	11	p14.3	-	AS II	growth arrest-specific 2
DLL3	19	q13.2	+	AS II	delta-like 3 (Drosophila)
CDH4	20	q13.33	-	AS II	cadherin 4, type 1, R-cadherin (retinal)
SHROOM2	X	p22.2	-	AS II	shroom family member 2
AP1AR	4	q25	-	AS III	adaptor-related protein complex 1 associated regulatory protein
CDC27	17	q21.32	-	AS III	cell division cycle 27
PPM1D	17	q23.2	+	AS III	protein phosphatase, Mg2+/Mn2+ dependent, 1D
ZNF24	18	q12.2	+	AS III	zinc finger protein 24
TXN2	22	q12.3	+	AS III	thioredoxin 2
AKT3	1	q44	-	GBM IV	v-akt murine thymoma viral oncogene homolog 3
MDM4	1	q32.1	+	GBM IV	MDM4, p53 regulator
PDGFRB	5	q32	+	GBM IV	platelet-derived growth factor receptor, beta polypeptide
VEGFA	6	p21.1	+	GBM IV	vascular endothelial growth factor A
EGFR	7	p11.2	+	GBM IV	epidermal growth factor receptor
FGFR1	8	p11.23	+	GBM IV	fibroblast growth factor receptor 1
FGFR2	10	q26.13	-	GBM IV	fibroblast growth factor receptor 2
BIRC3	11	q22.2	+	GBM IV	baculoviral IAP repeat containing 3
ERRB2	14	q24.3	+	GBM IV	nuclear receptor
NTRK3	15	q25.3	-	GBM IV	neurotrophic tyrosine kinase, receptor, type 3
BRCA1	17	q21.31	+	GBM IV	breast cancer 1, early onset
AKT2	19	q13.2	+	GBM IV	v-akt murine thymoma viral oncogene homolog 2
SMARCA4	19	p13.2	+	GBM IV	SWI/SNF related, matrix associated, actin dependent regulator of chromatin

Summary of discussed genes that were exclusively observed to be under- or overexpressed in a specific type of astrocytoma. The expression state of a gene in tumor is specified by the 'Expression' column with '-' representing underexpression and '+' representing overexpression in comparison to normal brain

chromatin remodelers (SUV420H1, SUV420H2, SUZ12) were underexpressed in PA I. In accordance with a recent study [53], ATRX, a biomarker of adult astrocytomas, was underexpressed in PA I. In contrast to AS III and GBM IV, HIF1AN was strongly overexpressed in PA I. Further, CDKN2B, a tumor suppressor for which overexpression has been reported to inhibit cell proliferation and to cause senescence of glioma cells with intact RB pathway [54], was overexpressed. OLIG2, which has been reported to show increased expression in PA I and high-grade gliomas [55], was overexpressed. NRTK2, which has been reported to be highly expressed in low grade (WHO grade I and II) gliomas [56], was overexpressed. Further, KRAS, which plays an important role in cell cycle regulation, was underexpressed. Additionally, H3F3A, which encodes for a histone variant that is predominantly integrated into chromatin of non-dividing cells, was underexpressed.

Selected genes exclusively observed in AS II In comparison to PA I and GBM IV, less genes were found to be exclusively differentially expressed in AS II (Fig. 3a–b). FAM110C, which has been reported to be part of a stem cell-related self-renewal signature associated with resistance to chemotherapy [57] and for which overexpression has been shown to promote cell cycle arrest in rats [58], was underexpressed. CDH4, which encodes for a cell-adhesion protein involved in brain segmentation and neural outgrowth, was underexpressed. Underexpression of CDH4 is known to play a role in early tumor progression of colorectal and gastric cancer [59]. NR2E1 (TLX), which is involved in anterior brain differentiation, was underexpressed. Underexpression of NR2E1 has been associated with cancer stem cell death and longer survival of G-CIMP glioma patients [60]. Further, SHROOM2 involved in cell spreading and GAS2 involved in apoptosis were both underexpressed. The transcription factor HEY2 and the Notch ligand DLL3 both known for their functions in neurogenesis and implicated in glioma biology [61] were overexpressed. EYA1, which encodes for a phosphatase and transcriptional coactivator that is involved in DNA repair and which has been associated with glioma tumorigenesis [62], was overexpressed.

Selected genes exclusively observed in AS III Like for AS II, only relatively few genes were exclusively differentially expressed in AS III. Interestingly, PPM1D, which is involved in p53-mediated cell cycle arrest, was overexpressed. PPM1D gain-of-function mutations have been reported for brain stem gliomas [63]. Additionally, a PPM1D knock-down has been reported to inhibit proliferation and invasion of glioma cells [64]. Further, AP1AR, which negatively regulates cell spreading, size and motility, was underexpressed. CDC27 (APC3), which is part of the anaphase promoting complex and which is involved in timing of mitosis, was underexpressed. Downregulation of a related component (APC7) of the anaphase promoting complex has been observed in breast cancer with poor prognosis [65]. TXN2, which has been identified to play an important role in the protection of osteosarcomas against oxidant-induced apoptosis [66], was overexpressed. Also ZNF24, which is involved in the maintenance of progenitor cell states in the developing central nervous system, was overexpressed. ZNF24 has further been reported to be involved in the negative regulation of angiogenesis [67].

Selected genes exclusively observed in GBM IV Many known cancer genes (e.g. BIRC3, BRCA1, EGFR, ERRB2, PDGFRB, VEGFA) were overexpressed in GBM IV. EGFR signaling has been reported to contribute to radiation and chemotherapy resistance of gliomas [68]. In line with VEGFA overexpression, PDGFRB, which has been reported to enhance glioma angiogenesis in tumor endothelia by promoting pericyte recruitment [69, 70], was overexpressed. Further, MDM4, which has been observed to inhibit a p53-dependent growth control [71, 72], was overexpressed. AKT2, for which underexpression has been reported to induce apoptosis and for which overexpression has been associated with cell survival and invasion of more aggressive gliomas [73, 74], was overexpressed. FGFR1, which has been reported for its increased expression and association with autocrine growth signaling in GBM IV [75], was overexpressed. Further, SMARCA4, which has been observed to have increased expression in gliomas and which is potentially involved in controlling of cell proliferation, migration and invasion [76], was overexpressed. PKG1, which has been reported to promote radioresistance of glioma cells [77, 78], was overexpressed. Further, AKT3, which has recently been reported to inhibit vascular tumor growth [79], was underexpressed. FGFR2, which is frequently found to be underexpressed in primary GBM IV and which has been associated with a poor clinical outcome [80], was underexpressed. NTRK3, which has been reported to show reduced expression in high-grade gliomas due to underlying DNA methylation changes [81], was underexpressed.

Selected genes in the intersection of PA I, AS II, AS III and GBM IV Genes commonly under- or overexpressed in PA I, AS II, AS III and GBM VI were involved in cell cycle regulation, differentiation, apoptosis and cell migration. We found that the cyclin-dependent kinase inhibitor CDKN2D was underexpressed and CD44, HIF1A and MAPKAPK3 were overexpressed in all four astrocytoma grades. CD44 is a well-known stem cell marker that has been reported to represent a potential therapeutic target for glioblastoma [82]. HIF1A encodes the alpha subunit of

the TF hypoxia-inducible factor-1 (HIF-1), which is one of the master regulators of hypoxia response promoting glioma growth and angiogenesis [83]. MAPKAPK3 is a central integrator of mitogen and stress responses in different MAPK pathways [84]. Interestingly, RB1, a know tumor suppressor controlling the progression through G1 into the S-phase of the cell cycle [85], was overexpressed. Induction of wild-type RB1 has been reported to inhibit tumor growth and tumorigenicity [86]. On the other hand, inactivating mutations affecting the RB pathway have frequently been observed in higher-grade gliomas [85]. This potentially indicates that an overexpression of wild-type RB1 in PA I may contribute to a reduced tumor growth, whereas an exclusive overexpression of CDK4 in concert with RB1 observed for AS II, AS III and GBM IV may counteract the inhibition of tumor growth (see next section for more details to CDK4).

Selected genes in the intersection of AS II, AS III and GBM IV but not in PA I Genes commonly under- or overexpressed in AS II, AS III and GBM IV were enriched for cell-cell signaling, cell cycle, differentiation, DNA repair, apoptosis and metabolism. Several known oncogenes (e.g. ABL1, AKT1, MYC, NRAS) and tumor suppressor genes (e.g. ATM, BCL10, TP53) were overexpressed in all three astrocytoma types. AKT1 has been found to enhance proliferation and invasion of glioma cells [87]. Overexpression of NRAS that increased with glioma grade was observed in [88]. Overexpression and different cellular locations of TP53 have been reported for primary and secondary glioblastomas impacting on vasculature control and tumorigenesis [89]. Overexpression of TP53 has also been associated with shorter progression free survival in malignant gliomas [90]. Further, also CDK4 and RAF1 were overexpressed. CDK4 overexpression has been reported to induce hyperploidy and to counteract senescence of cultured mouse astrocytes [91]. Astrocyte-specific overexpression of CDK4 in transgenic mouse lines has been observed to provide cell growth advantages in concert with TP53 pathway alterations [92]. Consecutive RAF1 activation has been reported to induce glioma formation in mice [93]. Moreover, also IDH1 was overexpressed. Interestingly, the overexpression of IDH1 in gliomas has recently been reported to have different impacts on chemotherapy response. Wild-type IDH1 was associated with resistance, whereas mutant-IDH1 showed enhanced sensitivity to therapy [94]. MAP2K4, which has been reported to inhibit tumor cell invasion in lung cancer [95], was strongly underexpressed. Further, also MAP2K1, which is involved in the regulation of many cellular processes including proliferation, differentiation and apoptosis, and also MKRN1, which has been observed to stimulate apoptosis under stress conditions [96], were both underexpressed.

Selected genes in the intersection of AS III and GBM IV but not in PA I and AS II Genes commonly under- or overexpressed in AS III and GBM IV were involved in cell migration, cell cycle, DNA repair, chromatin organization, angiogenesis and metabolism. HIF1AN (FIH-1), an inhibitor of the previously reported HIF-1, was underexpressed. HIF1AN is involved in hypervascularization and survival of glioma cells under hypoxic conditions and may represent a potential therapeutic target [97]. EZH2, a member of the polycomb-group family involved in the control of DNA methylation [98] and histone H3K27 trimethylation [99] over cell generations, was overexpressed. Also VEGFB involved in blood vessel survival [100] and CDC20 contributing to survival of glioma initiating cells [101] were overexpressed. Further, SOX2, a marker for undifferentiated and proliferating cells observed to show expression levels that increase with the glioma grade [102] and reported to regulate genes and pathways associated with malignancy of stem-like and differentiated glioma cells [103], was overexpressed. TACC3, a potential oncogene overexpressed in a grade-specific manner [104] and observed as fusion partner of FGFR3 in glioblastomas [105], was overexpressed. Moreover, IDH2 was overexpressed. Interestingly, another study has associated the overexpression of a point-mutated IDH2 (IDH2R172K) with increased radio sensitivity, reactive oxygen metabolism, suppression of tumor growth and migration in glioma cell lines compared to wild-type IDH2 [106]. Thus, the underlying mutational status of IDH2 may influence tumor aggressiveness of AS III and GBM IV.

Transcriptional alterations of individual signaling pathways typically increase with WHO grade

Next, we focused on individual cancer-relevant signaling pathways and determined corresponding differentially expressed genes for each type of astrocytoma. Figure 4 shows the numbers of overexpressed genes in known cancer signaling pathways representing major differences and some similarities between individual astrocytoma types. We observed strong differences in the number of overexpressed genes for nearly all pathways with gradual increases from PA I to GBM IV. This trend was also observed for the majority of signaling pathways considering underexpressed genes, except for the DNA replication pathway and all DNA repair pathways that both only showed very few or no underexpressed genes in all four astrocytoma grades (Additional file 2: Figure S6). Focusing on overexpression (Fig. 4), especially genes involved in cell cycle, PI3K-AKT, TGF-Beta, focal adhesion, notch, DNA replication and DNA repair pathways were significantly affected by overexpression in AS II, AS III or GBM IV. Genes involved in the regulation of apoptosis were enriched in all four astrocytoma types.

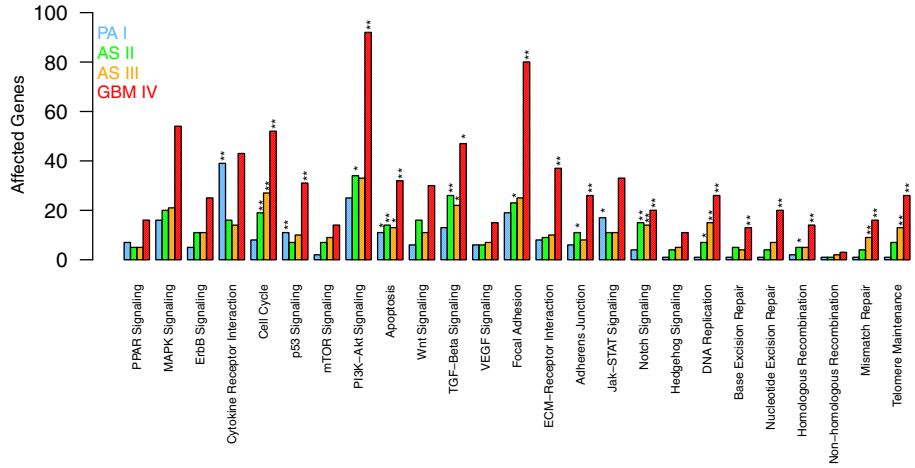

Fig. 4 Characteristic patterns of overexpression in signaling pathways distinguishing different astrocytoma grades. Number of overexpressed genes in each known cancer-relevant signaling pathway are shown for each type of astrocytoma (PA I, AS II, AS III, GBM IV). Significant enrichment of overexpressed genes in a pathway within a tumor type is highlighted by '*' ($P < 0.05$) and '**' ($P < 0.01$) (Fisher's exact test)

Interestingly, the cytokine-cytokine receptor interaction pathway did not not follow the general trend that the numbers of overexpressed genes systematically increased from PA I to GBM IV. This pathway showed nearly the same proportion of overexpressed genes in PA I as in GBM IV, whereas the proportions of overexpressed genes in AS II and AS III were consistently only approximately half as large as for PA I and GBM IV (Fig. 4). This atypical behavior also strongly contributed to significant exclusive overlaps between PA I and GBM IV comparing under- and overexpressed genes (purple subsets in Fig. 3c–d: 7 underexpressed genes with $P < 6.2 \times 10^{-6}$ and 33 overexpressed genes with $P < 1.7 \times 10^{-8}$, Fisher's exact test). We additionally note that the p53 pathway and the Jak-STAT pathway showed both a very similar behavior comparable to those of the cytokine-cytokine receptor pathway (Fig. 4).

Highly overlapping expression patterns of cytokine-cytokine receptor interaction pathway between PA I and GBM IV, but only GBM IV is enriched for known cancer genes

We observed similar proportions of overexpressed genes in the cytokine-cytokine receptor interaction pathway for PA I and GBM IV (Fig. 4). Cytokines are intracellular signaling proteins that are important regulators of immune response, cell growth, differentiation, metastasis, apoptosis and angiogenesis [107–109]. Some alterations of expression levels of specific cytokines, their corresponding receptors and links to their potential role in brain tumor development have already been reported for benign and malignant astrocytomas more than a decade ago [110–112]. In addition, different chemokines and chemokine receptors were found to contribute to glioma cell survival, migration and invasion [113–118]. We therefore focused on individual genes in the cytokine-cytokine receptor interaction pathway to provide a comprehensive overview of differentially expressed genes comparing PA I and GBM IV. A representation of the cytokine-cytokine receptor interaction pathway highlighting exclusively affected and commonly altered genes is shown in Fig. 5. We found a significant overlap of commonly observed under- and overexpressed genes in the cytokine-cytokine receptor interaction pathway comparing PA I and GBM IV (overlap: 20 genes, 1 underexpressed, 19 overexpressed genes, $P < 2.5 \times 10^{-42}$, Fisher's exact test). We further identified genes that were only differentially expressed in PA I (1 under- and 20 overexpressed genes) or in GBM IV (5 under- and 24 overexpressed genes) alone. Only genes that were exclusively overexpressed in GBM IV were significantly enriched for known cancer genes [45] ($P < 4.2 \times 10^{-5}$, Fisher's exact test). These genes were mainly assigned to the CXC chemokine, hematopoietin, PDGF or TGF-Beta pathway subfamilies of the cytokine-cytokine receptor interaction pathway (Fig. 5). This included genes such as EGFR, PDGFRB,

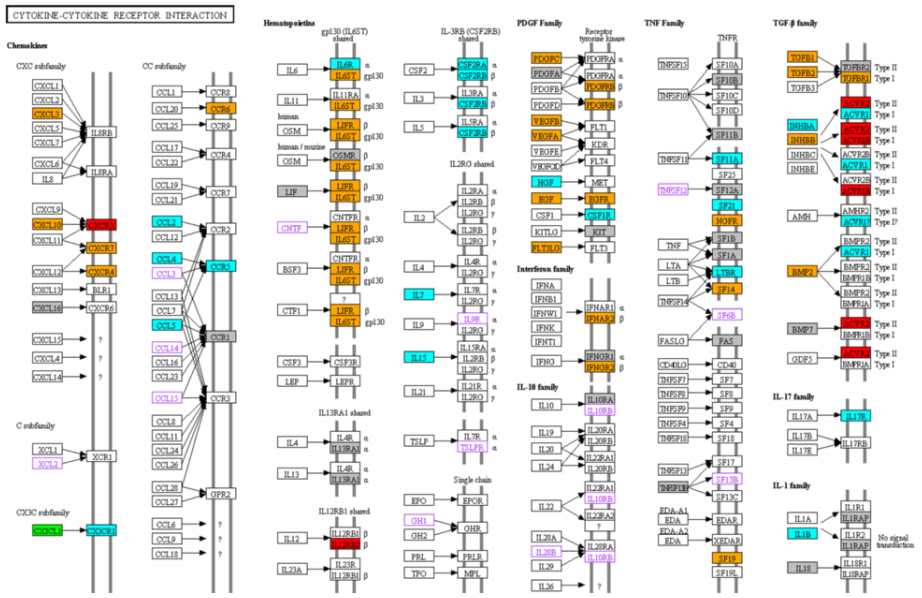

Fig. 5 Transcriptional alterations of cytokine-cytokine receptor interaction pathway genes comparing PA I and GBM IV. Representation of under- and overexpressed cytokine-cytokine receptor interaction pathway genes identified in PA I and GBM IV in comparison to normal brain tissue. The basic pathway map (hsa:04060) was generated using KEGG [130] and manually modified. Genes are colored according to their observed expression level (i) genes only overexpressed in PA I (*light blue*), (ii) genes only underexpressed in GBM IV (*red*), (iii) genes only overexpressed in GBM IV (*orange*), (iv) genes commonly overexpressed in PA I and GBM IV (*grey*), except for KIT that was underexpressed in both tumor types, (v) CX3CL1 overexpressed in PA I and underexpressed in GBM IV (*green*), (vi) genes that were not present on the microarray (*purple*), and (vii) genes with unchanged expression levels (*white*)

TNFRSF14 or VEGFA previously associated with aggressiveness, invasion and poor outcome of GBM IV [25, 119, 120].

Differences in CX3CL1 expression between PA I and AS II, AS III and GBM IV may contribute to absence or presence of glioma cell invasion

Interestingly, CX3CL1 (also known as fractalkine or neurotactin), a member of the cytokine-cytokine receptor interaction pathway (Fig. 5) encoding for a chemokine, showed a characteristic expression pattern distinguishing PA I from AS II, AS III and GBM IV (Additional file 1: Table S2). The soluble form of the CX3CL1 protein is a potent chemoattractant of T-cells and monocytes, while the cell-surface-bound form promotes strong adhesion of those leukocytes [121]. CX3CL1 implements its adhesive and migratory functions by interacting with the chemokine receptor CX3CR1 [122]. The roles of CX3CL1 and CX3CR1 in glioma invasion and progression have been reviewed for malignant astrocytomas in [117]. Potential contributions of both genes to suppress an invasive phenotype in PA I have not been studied so far.

We found that CX3CL1 and CX3CR1 were overexpressed in PA I in comparison to normal brain tissue, whereas we further observed strong underexpression of CX3CL1 and unchanged expression of CX3CR1 in AS II, AS III and GBM IV (Additional file 1: Table S2). In accordance, CX3CR1 has been reported to be expressed in gliomas [123], and CX3CL1 has been reported to reduce neuronal migration by increasing cell adhesion [124]. Potentially, a similar CX3CL1-induced mechanism in PA I may contribute to the absence of infiltrative growth typically observed for AS II, AS III and GBM IV [125]. This hypothesis is supported by the finding that the inhibition of CX3CL1 strongly increased glioma cell invasion suggesting that functionally active CX3CL1 counteracts an invasive phenotype [116]. Additionally, they also reported that TGFB1 negatively influenced the expression of CX3CL1 facilitating glioma cell detachment and dispersion. In agreement with our hypothesis, the expression

levels of CX3CL1 were negatively correlated with those of TGFB1 ($r = -0.98$, $P < 0.01$). We observed overexpression of TGFB1 in AS II, AS III and GBM IV, whereas TGFB1 expression was unchanged in PA I in comparison to normal brain tissue (Additional file 1: Table S2).

A transcriptional signature distinguishes PA I from AS II, AS III and GBM IV

Besides some similarities, our previous studies clearly indicated the existence of systematic differences between PA I and AS II, AS III and GBM IV supporting the finding that both classes represent different pathobiological entities [126]. To further investigate this, we determined a molecular signature comprising 1,089 differentially expressed genes distinguishing PA I from AS II, AS III and GBM IV (Fig. 6, Additional file 1: Table S3). This signature included all under- and overexpressed genes from PA I that did not show the same expression state in AS II, AS III or GBM IV. Vice versa, this signature also included each gene that was identified as under- or overexpressed in AS II, AS III or GBM IV but which did not show the same expression state in PA I. Clusters of genes that were under- or overexpressed in one class but not in the other are clearly visible characterizing differences between PA I and AS II, AS III and GBM IV (Fig. 6). A gene annotation analysis (Additional file 1: Table S3) further revealed that nearly 14 % of the signature genes were annotated as TFs (151 of 1,089), about 10 % were part of known cancer-relevant signaling pathways (111 of 1,089), about 5 % were known cancer genes (55 of 1,089) and about 3 % were part of metabolic pathways (34 of 1,089).

A regulatory network is associated with expression differences between PA I and AS II, AS III and GBM IV

Next, we used the 151 differentially expressed TFs from the molecular signature (Fig. 6, Additional file 1: Table S3)

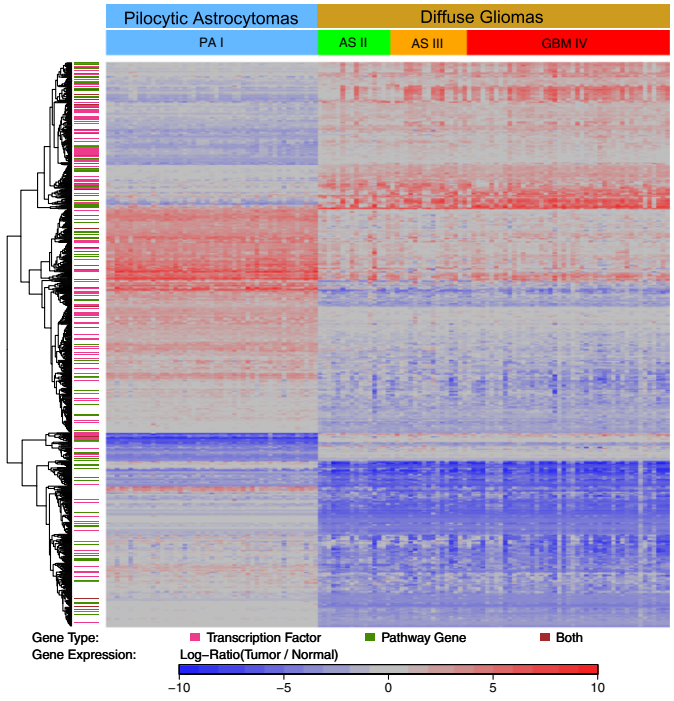

Fig. 6 Molecular signature distinguishing PA I from AS II, AS III and GBM IV. The heatmap shows the expression levels of 1,089 genes (*rows*) with strong expression differences between PA I and AS II, AS III and GBM IV for individual tumor patients (*columns*). Expression levels are displayed as log-ratios comparing gene expression levels in tumor to normal brain. Underexpressed genes are displayed in blue, unchanged expressed genes in grey, and overexpressed genes are displayed in red. Genes were clustered according to their similarity of expression levels across all tumor samples. The color code at the left side highlights genes that are TFs (*purple*), signaling pathway genes (*green*), or both (*brown*)

to learn a transcriptional regulatory network that best explained expression changes of all signature genes distinguishing PA I from AS II, AS III and GBM IV (Fig. 7, Additional file 1: Table S4). This network contained for each individual signature gene those TFs that may act as putative regulators of this gene. The regulatory network was extremely sparse containing only 1,558 out of 164,439 theoretically possible regulatory links from TFs to signature genes. We observed more than three times more activator than repressor links in the network (1,195 vs. 363). Nine TFs did not have any outgoing regulatory links to other signature genes, and no putative regulators were identified for 83 signature genes.

Still, as expected, the obtained regulatory network was highly predictive for the expression levels of signature genes in our astrocytoma data set used to learn the network (Additional file 2: Figure S7a). We further used the obtained regulatory network to predict expression changes of signature genes in three independent brain tumor cohorts (41 PA I from [46], 465 low grade gliomas including 50 AS II and 104 AS III from TCGA LGG, 553 GBM IV from TCGA GBM [23], see Additional file 2: Text S4 for details). We observed that the regulatory network was very predictive for the vast majority of signature genes (Additional file 2: Figures S7b–d). We also analyzed the proportion of putative direct TF-target gene interactions by comparing predicted target genes of TFs in the regulatory network to target genes predicted by TF-based motif search in promoter sequences of signature genes (see Additional file 2: Text S5 for details). We observed significant overlaps of network- and motif-based target genes for many TFs, but there were also TFs with only little or no overlaps (Additional file 2: Figure S8). All these tests indicated that the regulatory network contained relevant TF-target gene links to enable the prediction of signature gene expression levels.

Expression changes of hub regulators characterize differences between PA I and AS II, AS III and GBM IV

We next utilized the obtained signature-specific regulatory network to identify central hub TFs with many outgoing links to other signature genes. These hub regulators are represented by large nodes in Fig. 7. The majority of these TFs had on average lower expression levels in AS II, AS III and GBM IV than in PA I (blue nodes). A smaller proportion of hub TFs had higher expression levels in AS II, AS III and GBM IV than in PA I (red nodes). Many of these hub TFs were part of three major functional categories: (i) TFs involved in apoptosis, cell proliferation, cell cycle and DNA repair (CCNA2, CCNB1, CCNB2, CDC20, CHD5, GPR123, MEF2C, NEUROD1, VIP, ZNF365), (ii) TFs involved in chromatin remodeling, histone modifications and DNA methylation (CHD5, DNMT1, EZH2, JARID2), and (iii) TFs involved in brain development and differentiation (ARNT2, CHD5, DNMT1, ELF1, EOMES, HLF, JARID2, LHX1, MEF2C, NEUROD1, OLIG1, SOX10, SOX11, THRB, TBR1, VIP, ZIC1, ZIC3).

Next, we studied the hierarchy of TFs in the regulatory network to identify signature-specific hub TFs that had many regulatory links to other TFs. We found that several TFs had clearly increased numbers of outgoing links to other TFs (Additional file 2: Figure S9). Six TFs had more than five outgoing regulatory links to other TFs (CCNL2, GPR123, ZCCHC24, TBR1, ZNF300, ZNF337). CCNL2 encodes for a cyclin involved in the regulation of splicing, apoptosis and cell growth [127]. GPR123 is a member of the adhesion family of G-protein coupled receptors mutated in leukemia [128]. TBR1 encodes for a T-box TF required for normal brain development expressed in post-mitotic cells [129]. Nothing was known in the literature about the functions of ZCCHC24, ZNF300 and ZNF337 so far. We analyzed their network-target genes to learn more about their putative functions. This suggested that ZCCHC24 is involved in the regulation of the cell cycle and of cell-cell interactions. ZNF300 might act on developmental processes impacting on DNA and histone methylation patterns. ZNF337 might contribute to genomic and epigenomic integrity.

Mutations affecting TFs contribute to differences between PA I and AS II, AS III and GBM IV

To further characterize how genomic and epigenomic mutations may have contributed to expression differences of TFs between PA I and AS II, AS III and GBM IV, we analyzed the individual signature-specific TFs for alterations of DNA methylation levels or gene copy number mutations in comparison to normal tissue. Gene copy number mutations are typically absent in PA I, but changes of DNA methylation patterns within gene bodies or up- and downstream of transcription start sites have been reported [38]. In contrast to PA I, deletions and amplifications of individual genes are typically present in AS II, AS III and GBM IV [40]. DNA methylation profiles were available for the majority of our PA I tumors (38 of 47) and gene copy number profiles were available for all our AS II, AS III and GBM IV tumors. We therefore analyzed the expression of individual signature-specific TFs in relation to directly underlying mutations (Fig. 8).

We found for PA I that TFs with altered expression and/or altered DNA methylation levels were part of three major functional categories (Fig. 8a): (i) TFs involved in development and differentiation (e.g. EN2, EOMES, DMRT2, NR0B1), (ii) TFs involved in cell cycle control, proliferation and apoptosis (e.g. CDC20, NFKBIZ, ZCCHC24), and (iii) TFs involved in chromatin remodeling and DNA methylation (ESRRG, L3MBTL4, SATB2). Several TFs were strongly under- or overexpressed in

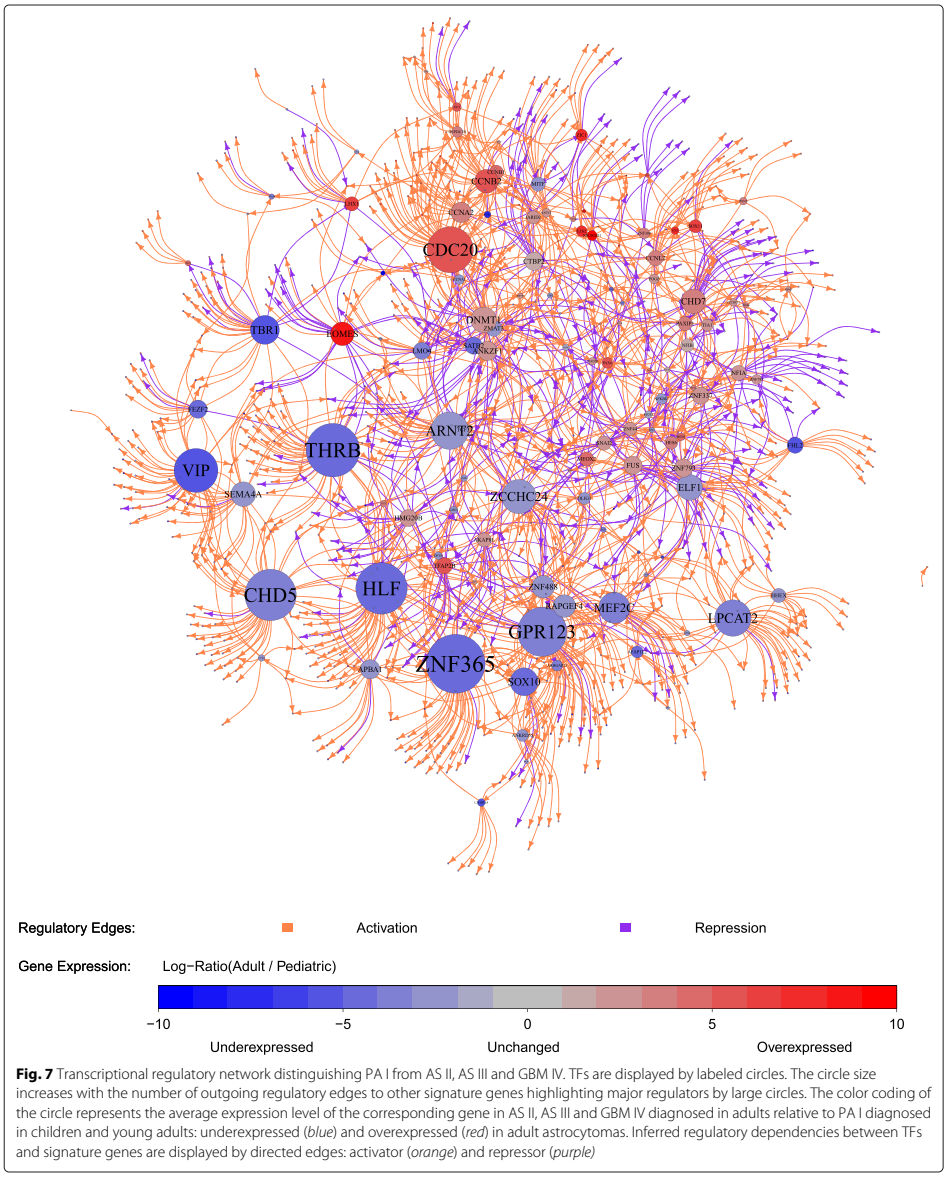

Fig. 7 Transcriptional regulatory network distinguishing PA I from AS II, AS III and GBM IV. TFs are displayed by labeled circles. The circle size increases with the number of outgoing regulatory edges to other signature genes highlighting major regulators by large circles. The color coding of the circle represents the average expression level of the corresponding gene in AS II, AS III and GBM IV diagnosed in adults relative to PA I diagnosed in children and young adults: underexpressed (*blue*) and overexpressed (*red*) in adult astrocytomas. Inferred regulatory dependencies between TFs and signature genes are displayed by directed edges: activator (*orange*) and repressor (*purple*)

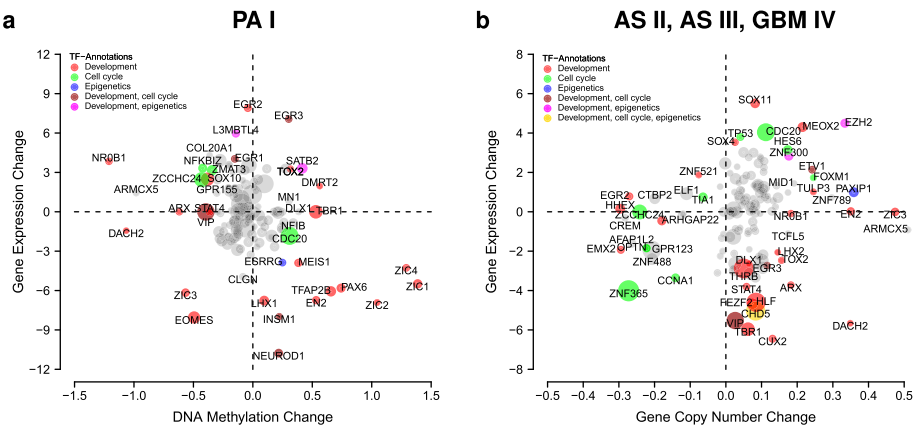

Fig. 8 Epigenomic and genomic mutations associated with TF expression changes distinguishing PA I from AS II, AS III and GBM IV. TFs are displayed by circles. The circle size increases with the number of outgoing regulatory edges to other signature genes highlighting major regulators by large circles. Names of TFs strongly deviating from the center are shown in black. TFs belonging to selected annotation categories are highlighted by colored circles. Gene-specific changes in expression, DNA methylation or copy number are quantified by log-ratios comparing tumor to normal brain tissue. A log-ratio close to zero indicates no change in tumor, whereas a strong deviation from zero indicates a change. **a**, Average DNA methylation changes associated with TFs plotted against their average expression profiles in PA I. **b**, Average copy changes of TFs plotted against their average expression profiles in AS II, AS III and GBM IV

PA I without strong directly underlying DNA methylation changes (e.g. EGR2, INSM1, LHX1, NEUROD1). None of the central hub TFs in Fig. 7 showed strong expression changes in PA I in response to directly underlying DNA methylation changes, except for CDC20 and ZCCHC24. Other TFs with fewer outgoing links to signature genes showed greatly altered expression levels in PA I in response to strong DNA methylation changes (e.g. EGR3, EN2, EOMES, NR0B1, PAX6, SATB2, ZIC1, ZIC2, ZIC3, ZIC4).

This situation was quite different for AS II, AS III and GBM IV (Fig. 8b). Four central hub TFs showed strongly altered expression levels in response to directly underlying gene copy number mutations (CDC20, GPR123, ZNF365, ZNF488), whereas other hub TFs showed strong underexpression without underlying deletions (e.g. CHD5, HLF, TBR1, THRB, VIP). Again, TFs with altered expression and/or copy number mutations were part of three major functional categories as observed for PA I before. The majority of TFs was involved in development and differentiation (e.g. EGR2, EMX2, DACH2, MEOX2, SOX11). Other TFs were involved in cell cycle control, proliferation, apoptosis and DNA repair (e.g. CCNA1, CDC20, CHD5, TP53, ZNF365). Some TFs were involved in the regulation of chromatin remodeling and DNA methylation (CHD5, EZH2, PAXIP1, ZNF300).

Conclusions

Our computational study revealed similarities and differences in gene expression levels between astrocytomas of all four WHO grades under consideration of astrocytoma type-specific normal brain references. We compared all four considered astrocytoma grades (PA I, AS II, AS III, GBM IV) at the level of individual genes and cancer-relevant signaling pathways. Thereby, we identified many genes that were exclusively under- or overexpressed in a specific astrocytoma grade. In addition, we also revealed many genes that showed the same pattern of under- or overexpression in specific subsets of astrocytoma grades. We discussed many of these genes in the background of the currently existing literature and we summarized selected astrocytoma type-specific differentially expressed genes that might be of interest for future studies that aim at the development of novel markers. We further observed at the level of individual genes and cancer-relevant signaling pathways that the number of differentially expressed genes typically increased with the astrocytoma grade. This trend suggests an association of transcriptional alterations with the increased tumor aggressiveness of the different astrocytoma grades. Interestingly, the cytokine receptor interaction pathway escaped this general trend. Nearly the same number of overexpressed genes were observed for PA I and GBM IV in this pathway. Detailed studies further identified

commonly and exclusively overexpressed genes in the cytokine receptor interaction pathway for PA I and GBM IV and further revealed that only genes that were overexpressed in GBM IV were significantly enriched for known cancer genes involved in aggressiveness, invasion and poor outcome. Moreover, this in-depth analysis also revealed a characteristic expression patterns of CX3CL1 (fractalkine) and its receptor CX3CR1 that distinguished PA I from AS II, AS III and GBM IV. These genes are involved in glioma invasion and progression of malignant astrocytomas [117]. Strong overexpression of both genes in PA I in comparison to higher grade astrocytomas suggests a potential contribution to the non-invasive growth behavior of PA I. Thus, it might be worth to validate this potential link by gene knockdowns in a future study.

Surprisingly, PA I was strongly associated with the mesenchymal subtype, which is typically observed for very aggressive GBM IV. Additional analyses indicated that the tumor micro-environment may have a greater contribution to the manifestation of the mesenchymal subtype than the tumor biology itself, which might explain the seemingly contradiction between the similarity in terms of subtype classification and the very different clinical course of mostly benign PA I and highly malignant GBM IV. In accordance with this, we found that the endothelial cell marker ANGPT2 (alias ANG2) was highly overexpressed in PA I and GBM IV but not in AS II or AS III. Using immunohistochemistry, we confirmed that PA I and GBM IV showed Ang2-positive endothelial cells in regions with activated blood vessels. This feature was largely absent in AS II and AS III. Thus, our study suggests that microvascular proliferation and necrosis, which both have been described as common histological features of PA I and GBM IV [8], contribute at least to some extent to the observation of the mesenchymal subtype.

We also revealed major transcriptional regulators that distinguished PA I from AS II, AS III and GBM IV based on a computationally inferred signature-specific transcriptional regulatory network. We found that many of the differentially expressed central transcriptional regulators play important roles in cell cycle regulation, chromatin remodeling, or brain development and differentiation. Further analyses indicated that the differential expression of transcriptional regulators was mainly driven by directly underlying DNA methylation changes in PA I or gene copy number alterations in AS II, AS III and GBM IV. We note that the impacts of DNA methylation changes on transcriptional regulators in AS II, AS III and GBM IV could not be compared to those in PA I, because DNA methylation profiles were not available for AS II, AS III and GBM IV tumors from Rembrandt. This could be addressed in a future study using DNA methylation profiles measured for AS II, AS III and GBM IV from TCGA brain tumor cohorts.

We are aware that our network approach can also be utilized for the analysis of a molecular signature that distinguishes all four astrocytoma types. However, this should be done based on a larger data set including additional astrocytoma samples from other resources to ensure robustness and transferability. A future study could for example utilize additional publicly available astrocytoma data sets (e.g. TCGA and ICGC data sets and other smaller studies) and further try to directly integrate additional omics layers (e.g. gene copy numbers, DNA methylation profiles, single nucleotide polymorphisms).

Altogether, our study confirmed many known findings and revealed novel interesting insights into astrocytoma biology and therefore represents a valuable resource for future studies.

Additional files

Additional file 1: Contains supporting Tables S1–S7. Table S1: Summary of considered astrocytoma samples. **Table S2:** Summary of t-test results for PA I, AS II, AS III and GBM IV. **Table S3:** Signature genes distinguishing PA I from AS II, AS III and GBM IV. **Table S4:** Regulatory network associated with expression differences of signature genes. **Table S5:** Summary of integrated gene annotations. **Table S6:** Verhaak classification results for PA I, AS II, AS III and GBM IV. **Table S7:** Correlation statistics for macrophage marker genes included in the Verhaak-classifier. (ZIP 3287 kb)

Additional file 2: Contains supporting Texts S1–S5 and supporting Figures S1–S9. Text S1: Processing of Rembrandt gene copy number data. **Text S2:** Lasso-based regulatory network inference. **Text S3:** Endothelial cells express Ang2 in PA I and GBM IV. **Text S4:** Network validation based on independent glioma cohorts. **Text S5:** Comparison of motif search-based and gene expression-based TF-target links. **Figure S1:** Age distribution of PA I, AS II, AS III and GBM IV patients. **Figure S2:** Functional categorization of differentially expressed genes. **Figure S3:** Verhaak classification results of two independent PA I cohorts. **Figure S4:** Ang2 immunohistochemistry. **Figure S5:** Associations of PA I, AS II, AS III and GBM IV with hypermethylator subtype. **Figure S6:** Differential expression in individual signaling pathways. **Figure S7:** Gene expression-based regulatory network validation. **Figure S8:** Overlap of network-based TF-target interactions and motif-based TF-binding sites. **Figure S9:** TF-TF interaction network. (PDF 1249 kb)

Abbreviations

WHO: World Health Organization; PA I: pilocytic astrocytoma WHO grade I; AS II: diffuse astrocytoma WHO grade II; AS III: anaplastic astrocytoma WHO grade III; GBM IV: glioblastoma WHO grade IV; TF: transcription factor/cofactor.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Wrote manuscript: MS; Designed studies: MS, Performed studies: MS, MG; Curated gene annotations: MG, BF, MS; Immunohistochemistry: MM; Interpretation of results: MS, MM, BK; Read and approved the final manuscript: all.

Acknowledgments

We thank Thomas Buder (ZIH TU Dresden) and the reviewers for valuable comments. This work was done in the frame of GlioMath-Dresden funded by the European Social Fund and the Free State of Saxony. MS and BK are members and were supported by the University CancerCentrum Dresden Network Brain Tumors.

Author details
[1]Innovative Methods of Computing, Center for Information Services and High Performance Computing, Dresden University of Technology, Dresden, Germany. [2]Cellular Networks and Systems Biology, University of Cologne, CECAD, Cologne, Germany. [3]Institute of Molecular Systems Biology, Zurich, Switzerland. [4]Institute of Neurology (Edinger Institute), Goethe University, Frankfurt, Germany. [5]Institute for Clinical Genetics, Faculty of Medicine Carl Gustav Carus, Dresden University of Technology, Dresden, Germany. [6]German Cancer Consortium (DKTK), Dresden, Germany. [7]German Cancer Research Center (DKFZ), Heidelberg, Germany.

Received: 17 July 2015 Accepted: 1 November 2015
Published online: 16 December 2015

References

1. Ohgaki H, Kleihues P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. J Neuropathol Exp Neurol. 2005;64:479–89.
2. Canoll P, Goldman JE. The interface between glial progenitors and gliomas. Acta Neuropathol. 2008;116:465–77.
3. Chen J, McKay RM, Parada LF. Malignant Glioma: Lessons from Genomics, Mouse Models, and Stem Cells. Cell. 2012;149:36–47.
4. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. WHO classification of tumours of the central nervous system. Acta Neuropathol. 2007;11:97–109.
5. Jones DT, Gronych J, Lichter P, Witt O, Pfister SM. MAPK pathway activation in pilocytic astrocytoma. Cell Mol Life Sci. 2012;69:1799–1811.
6. Armstrong GT, Conklin HM, Huang S, Srivastava D, Sanford R, Ellison DW, et al. Survival and long-term health and cognitive outcomes after low-grade glioma. Neuro Oncol. 2011;13:223–34.
7. Jones DT, Hutter B, Jäger N, Korshunov A, Kool M, Warnatz HJ, et al. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. Nat Genet. 2013;45:927–932.
8. Kurwale NS, Suri V, Suri A, Sarkar C, Gupta DK, Sharma BS, et al. Predictive factors for early symptomatic recurrence in pilocytic astrocytoma: does angiogenesis have a role to play? J Clin Neurosci. 2011;18:472–7.
9. Rodriguez EF, Scheithauer BW, Giannini C, Rynearson A, Cen L, Hoesley B, et al. PI3K/AKT pathway alterations are associated with clinically aggressive and histologically anaplastic subsets of pilocytic astrocytoma. Acta Neuropathol. 2011;121:407–20.
10. Tonn JC, Westphal M, Rutka JT, Grossman SA. Neuro-oncology of CNS tumors. ISBN: 978-3540258339. Berlin Heidelberg: Springer; 2005.
11. Kelly PJ. Gliomas: Survival, origin and early detection. Surg Neurol Int. 2010;1:96.
12. Ohgaki H, Kleihues P. Genetic alterations and signaling pathways in the evolution of gliomas. Cancer Sci. 2009;100:2235–245.
13. Ohgaki H, Kleihues P. The definition of primary and secondary glioblastoma. Clin Cancer Res. 2013;19:764–72.
14. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science. 2014;343:189–193.
15. Tove LL, Hansson HA, Stein S, Sverre HT. Prognostic value of histological features in diffuse astrocytomas WHO grade II. Int J Clin Exp Pathol. 2012;5:152–8.
16. Smoll NR, Hamilton B. Incidence and relative survival of anaplastic astrocytomas. Neuro Oncol. 2014;16:1400–07.
17. Nuno M, Birch K, Mukherjee D, Sarmiento JM, Black KL, Patil CG. Survival and prognostic factors of anaplastic gliomas. Neurosurgery. 2013;73:458–65.
18. Sturm D, Bender S, Jones DT, Lichter P, Grill J, Becher O, et al. Paediatric and adult glioblastoma: multiforme (epi)genomic culprits emerge. Nat Rev Cancer. 2014;14:92–107.
19. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321:1807–1812.
20. Gorovets D, Kannan K, Shen R, Kastenhuber ER, Islamdoust N, Campos C, et al. IDH mutation and neuroglial developmental features define clinically distinct subclasses of lower grade diffuse astrocytic glioma. Clin Cancer Res. 2012;18:2490–501.
21. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–996.
22. Taveras JM, Thompson HG, Pool JL. Should we treat glioblastoma multiforme? A study of survival in 425 cases. Am J Roentgenol Radium Ther Nucl Med. 1962;87:473–9.
23. The Cancer Genome Atlas Nework. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–1068.
24. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155:462–477.
25. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110.
26. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17:510–522.
27. Cooper LAD, Gutman DA, Long Q, Johnson BA, Cholleti SR, Kurc T, et al. The Proneural Molecular Signature Is Enriched in Oligodendrogliomas and Predicts Improved Survival among Diffuse Gliomas. PLoS One. 2010;5:12548.
28. Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. Autoregressive Higher-Order Hidden Markov Models: Exploiting Local Chromosomal Dependencies in the Analysis of Tumor Expression Profiles. PLoS One. 2014;9:100295.
29. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463:318–25.
30. Jörnsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling TE, et al. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. Mol Syst Biol. 2011;7:486. doi:10.1038/msb.2011.17.
31. Deshmukh H, Yu J, Shaik J, MacDonald TJ, Perry A, Payton JE, et al. Identification of transcriptional regulatory networks specific to pilocytic astrocytoma. BMC Med Genomics. 2011;4:57. doi:10.1186/1755-8794-4-57.
32. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. Mol Syst Biol. 2012;8:605. doi:10.1038/msb.2012.37.
33. Wang C, Funk CC, Eddy JA, Price ND. Transcriptional analysis of aggressiveness and heterogeneity across grades of astrocytomas. PLoS One. 2013;8:76694.
34. Huang H, Hara A, Homma T, Yonekawa Y, Ohgaki H. Altered expression of immune defense genes in pilocytic astrocytomas. J Neuropathol Exp Neurol. 2005;64:891–901.
35. Rorive S, Maris C, Debeir O, Sandras F, Vidaud M, Bièche I, et al. Exploring the distinctive biological characteristics of pilocytic and low-grade diffuse astrocytomas using microarray gene expression profiles. J Neuropathol Exp Neurol. 2006;65:794–807.
36. Rickman DS, Bobek MP, Misek DE, Kuick R, Blaivas M, Kurnit DM, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res. 2001;61:6885–895.
37. Hunter S, Young A, Olson J, Brat DJ, Bowers G, Wilcox JN, et al. Differential expression between pilocytic and anaplastic astrocytomas: identification of apolipoprotein D as a marker for low-grade, non-infiltrating primary CNS neoplasms. J Neuropathol Exp Neurol. 2002;61:275–81.
38. Lambert SR, Witt H, Hovestadt V, Zucknick M, Kool M, Pearson DM, et al. Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. Acta Neuropathol. 2013;126:291–301.
39. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Statist Assoc. 2004;99:909–17.
40. Madhavan S, Zenklusen JC, Kotliarov Y, Sahmi H, Fine HA, Buetow K. Rembrandt: Helping personalized medicine become a reality through integrative translational research. Mol. Cancer Res. 2009;7:157–67.

41. Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell. 2006;9:287–300.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57: 289–300.
43. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B. 1996;58:267–88.
44. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. 2014;42:413–68.
45. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nar Rev Cancer. 2004;4:177–83.
46. Sharma MK, Mansur DB, Reifenberger G, Perry A, Leonard JR, Aldape KD, et al. Distinct genetic signatures among pilocytic astrocytomas relate to their brain region origin. Cancer Res. 2007;67:890–900.
47. Klein R, Roggendorf W. Increased microglia proliferation separates pilocytic astrocytomas from diffuse astrocytomas: a double labeling study. Acta Neuropathol. 2001;101:245–8.
48. Herder V, Iskandar CD, Kegler K, Hansmann F, Elmarabet SA, Khan MA, et al. Dynamic changes of microglia/macrophage M1 and M2 polarization in Theiler's murine encephalomyelitis. Brain Pathol. 20151750–3639. doi:10.1111/bpa.12238.
49. Stratmann A, Risau W, Plate KH. Cell type-specific expression of angiopoietin-1 and angiopoietin-2 suggests a role in glioblastoma angiogenesis. Am J Pathol. 1998;153:1459–66.
50. Turcan S, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483:479–83.
51. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48. doi: 10.1186/1471-2105-10-48.
52. Baala L, Briault S, Etchevers HC, Laumonnier F, Natiq A, Amiel J, et al. Homozygous silencing of T-box transcription factor EOMES leads to microcephaly with polymicrogyria and corpus callosum agenesis. Nat Genet. 2007;39:454–6.
53. Reuss DE, Sahm F, Schrimpf D, Wiestler B, Capper D, Koelsche C, et al. ATRX and IDH1-R132H immunohistochemistry with subsequent copy number analysis and IDH sequencing as a basis for an integrated? diagnostic approach for adult astrocytoma, oligodendroglioma and glioblastoma. Acta Neuropathol. 2015;129:133–46.
54. Fuxe J, Akusjärvi G, Goike HM, Roos G, Collins VP, Pettersson RF. Adenovirus-mediated overexpression of p15INK4B inhibits human glioma cell growth, induces replicative senescence, and inhibits telomerase activity similarly to p16INK4A. Cell Growth Differ. 2000;11: 373–84.
55. Otero JJ, Rowitch D, Vandenberg S. OLIG2 is differentially expressed in pediatric astrocytic and in ependymal neoplasms. J Neurooncol. 2011;104:423–38.
56. Wadhwa S, Nag TC, Jindal A, Kushwaha R, Mahapatra AK, Sarkar C. Expression of the neurotrophin receptors Trk A and Trk B in adult human astrocytoma and glioblastoma. J Biosci. 2003;28:181–8.
57. Murat A, Migliavacca E, Gorlia T, Lambiv WL, Shay T, Hamou MF, et al. Stem cell-related self-renewal signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. J Clin Oncol. 2008;26:315–24.
58. Li F, Jang H, Puttabyatappa M, Jo EJJM Curry. Ovarian FAM110C (Family with Sequence Similarity 110C): Induction during the periovulatory period and regulation of granulosa cell cycle kinetics in rats. Biol Reprod. 2012;86:185.
59. Miotto E, Sabbioni S, Veronese A, Calin GA, Gullini S, Liboni A, et al. Frequent aberrant methylation of the CDH4 gene promoter in human colorectal and gastric cancer. Cancer Res. 2004;64:8156–159.
60. Xie Q, Flavahan WA, Bao S, Rich J. The tailless root of glioma: Cancer stem cells. Cell Stem Cell. 2014;15:114–6.
61. Phillips HS, Karbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006;9:157–73.
62. Auvergne RM, Sim FJ, Wang S, Chandler-Militello D, Burch J, Al Fanek Y, et al. Transcriptional differences between normal and glioma-derived glial progenitor cells identify a core set of dysregulated genes. Cel Rep. 2013;3:2127–141.
63. Zhang L, Chen LH, Wan H, Yang R, Wang Z, Feng J, et al. Exome sequencing identifies somatic gain-of-function PPM1D mutations in brainstem gliomas. Nat Genet. 2014;46:726–30.
64. Wang P, Rao J, Yang H, Zhao H, Yang L. PPM1D silencing by lentiviral-mediated RNA interference inhibits proliferation and invasion of human glioma cells. J Huazhong Univ Sci Technolog Med Sci. 2011;31:94–9.
65. Park KH, Choi SE, Eom M, Kang Y. Downregulation of the anaphase-promoting complex (APC)7 in invasive ductal carcinomas of the breast and its clinicopathologic relationships. Breast Cancer Res. 2005;7:238–47.
66. Chen Y, Cai J, Murphy TJ, Jones DP. Overexpressed human mitochondrial thioredoxin confers resistance to oxidant-induced apoptosis in human osteosarcoma cells. J Biol Chem. 2002;277: 33242–3248.
67. Harper J, Yan L, Loureiro RM, Wu I, Fang J, D'Amore PA, et al. Repression of vascular endothelial growth factor expression by the zinc finger transcription factor ZNF24. Cancer Res. 2007;67:8736–741.
68. Hatanpaa KJ, Burma S, Zhao D, Habib AA. Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. Neoplasia. 2010;12:675–84.
69. Guo P, Hu B, Gu W, Xu L, Wang D, Huang HJ, et al. Platelet-derived growth factor-B enhances glioma angiogenesis by stimulating vascular endothelial growth factor expression in tumor endothelia and by promoting pericyte recruitment. Am J Pathol. 2003;162:1083–1093.
70. Nazarenko I, Hede SM, He X, Hedrén A, Thompson J, Lindström MS, et al. PDGF and PDGF receptors in glioma. Ups J Med Sci. 2012;117:99–112.
71. Riemenschneider MJ, Büschges R, Wolter M, Reifenberger J, Boström J, Kraus JA, et al. Amplification and overexpression of the MDM4 (MDMX) gene from 1q32 in a subset of malignant gliomas without TP53 mutation or MDM2 amplification. Cancer Res. 1999;59:6091–096.
72. Riemenschneider MJ, Knobbe CB, Reifenberger G. Refined mapping of 1q32 amplicons in malignant gliomas confirms MDM4 as the main amplification target. Int J Cancer. 2003;104:752–7.
73. Pu P, Kang C, Li J, Wang G. Suppression of glioma-cell survival by antisense and dominant-negative AKT2 RNA. Cancer Biol Med. 2005;2: 609–14.
74. Zhang J, Han L, Zhang A, Wang Y, Yue X, You Y. AKT2 expression is associated with glioma malignant progression and required for cell survival and invasion. Oncol Rep. 2010;24:65–72.
75. Morrision RS, Yamaguchi F, Bruner JM, Tang M, McKeehan W, Berger MS. Fibroblast growth factor receptor gene expression and immunoreactivity are elevated in human glioblastoma multiforme. Cancer Res. 1994;54:2794–799.
76. Bai J, Mei PJ, Liu H, Li C, Li W, Wu YP, et al. BRG1 expression is increased in human glioma and controls glioma cell proliferation, migration and invasion in vitro. J Cancer Res Clin Oncol. 2012;138:991–8.
77. Yan H, Yang K, Xiao H, Zou YJ, Zhang WB, Liu HY. Over-expression of cofilin-1 and phosphoglycerate kinase 1 in astrocytomas involved in pathogenesis of radioresistance. CNS Neurosci Ther. 2012;18:729–36.
78. Ding H, Cheng YJ, Yan H, Zhang R, Zhao JB, Qian CF, et al. Phosphoglycerate kinase 1 promotes radioresistance in U251 human glioma cells. Oncol Rep. 2014;31:894–900.
79. Phung TL, Du W, Xue Q, Ayyaswamy S, Gerald D, Antonello Z, et al. Akt1 and Akt3 exert opposing roles in the regulation of vascular tumor growth. Cancer Res. 2015;75:40–50.
80. Toedt G, Barbus S, Wolter M, Felsberg J, Tews B, Blond F, et al. Molecular signatures classify astrocytic gliomas by IDH1 mutation status. Int J Cancer. 2011;128:1095–103.
81. Palani M, Arunkumar R, Vanisree AJ. Methylation and expression patterns of tropomyosin-related kinase genes in different grades of glioma. Neuromolecular Med. 2014;16:529–39.
82. Xu Y, Stamenkovic I, Yu Q. CD44 attenuates activation of the hippo signaling pathway and is a prime therapeutic target for glioblastoma. Cancer Res. 2010;70:2455–464.
83. Li Z, Bao S, Wu Q, Wang H, Eyler C, Sathornsumetee S, et al. Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells. Cancer Cell. 2009;15:501–13.
84. Ludwig S, Engel K, Hoffmeyer A, Sithanandam G, Neufeld B, Palm D, et al. 3pK, a novel mitogen-activated protein (MAP) kinase-activated protein kinase, is targeted by three MAP kinase pathways. Mol Cell Biol. 1996;16:6687–697.
85. Nakada M, Kita D, Watanabe T, Hayashi Y, Teng L, Pyko IV, et al. Aberrant signaling pathways in glioma. Cancer. 2011;3:3242–278.

86. Woitach JT, Zhang M, Niu CH, Thorgeirsson SS. A retinoblastoma-binding protein that affects cell-cycle control and confers transforming ability. Nat Genet. 1998;19:371–4.
87. Zou J, Wang K, Han L, Zhang A, Shi Z, Pu P, et al. AKT1 and AKT2 promote malignant transformation in human brain glioma LN229 cells. Clin Oncol Cancer Res. 2011;8:144–8.
88. Orian JM, Vasilopoulos K, Yoshida S, Kaye AH, Chow CW, Gonzales MF. Overexpression of multiple oncogenes related to histological grade of astrocytic glioma. Br J Cancer. 1992;66:106–12.
89. Nagpal J, Jamoona A, Gulati ND, Mohan A, Braun A, Murdi R, et al. Revisiting the role of p53 in primary and secondary glioblastomas. Anticancer Res. 2006;26:4633–640.
90. Pollack IF, Hamilton RL, Finkelstein SD, Campbell JW, Martinez AJ, Sherwin RN. The relationship between TP53 mutations and overexpression of p53 and prognosis in malignant gliomas of childhood. Cancer Res. 1997;57:304–9.
91. Holland EC, Hively WP, Gallo V, Varmus HE. Modeling mutations in the G1 arrest pathway in human gliomas: overexpression of CDK4 but not loss of INK4aARF induces hyperploidy in cultured mouse astrocytes. Genes Dev. 1998;12:3644–649.
92. Huang ZY, Baldwin RL, Hedrick NM, Gutmann DH. Astrocyte-specific expression of CDK4 is not sufficient for tumor formation, but cooperates with p53 heterozygosity to provide a growth advantage for astrocytes in vivo. Oncogene. 2002;21:1325–34.
93. Lyustikman Y, Momota H, Pao W, Holland EC. Constitutive activation of Raf-1 induces glioma formation in mice. Neoplasia. 2008;10:501–10.
94. Wang JB, Dong DF, Wang MD, Gao K. IDH1 overexpression induced chemotherapy resistance and IDH1 mutation enhanced chemotherapy sensitivity in glioma cells in vitro and in vivo. Asian Pac J Cancer Prev. 2014;15:427–32.
95. Ahn YH, Yang Y, Gibbons DL, Creighton CJ, Yang F, Wistuba II, et al. Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor γ2 expression. Mol Biol Cell. 2011;31: 4270–285.
96. Lee EW, Lee MS, Camus S, Ghim J, Yang MR, Oh W. Differential regulation of p53 and p21 by MKRN1 E3 ligase controls cell cycle arrest and apoptosis. EMBO Journal. 2009;28:2100–113.
97. Wang E, Zhang C, Polavaram N, Liu F, Wu G, Schroeder MA, et al. The role of factor inhibiting HIF (FIH-1) in inhibiting HIF-1 transcriptional activity in glioblastoma multiforme. PLoS One. 2014;9(e86102). doi:10.1371/journal.pone.0086102.
98. Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature. 2006;439:871–4.
99. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 2007;39: 232–6.
100. Toda M. Glioma stem cells and immunotherapy for the treatment of malignant gliomas. ISRN Oncology. 2013;2013:673793.
101. Xie Q, Wu Q, Mack S, Yang K, Kim L, Hubert C, et al. CDC20 maintains tumor initiating cells. Oncotarget. 2015;6:13241–13254.
102. Annovazzi L, Mellai M, Caldera V, Valente G, Schiffer D. SOX2 expression and amplification in gliomas and glioma cell lines. Cancer Genomics Proteomics. 2011;8:139–47.
103. Berezovsky AD, Poisson LM, Cherba D, Webb CP, Transou AD, Lemke NW, et al. Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. Neoplasia. 2014;16:193–206.
104. Ducan CG, Killela PJ, Payne CA, Lampson B, Chen WC, Liu J, et al. Integrated genomic analyses identify ERRFI1 and TACC3 as glioblastoma-targeted genes. Oncotarget. 2010;1:265–77.
105. Parker BC, Annala MJ, Cogdell DE, Granberg KJ, Sun Y, Ji P, et al. The tumorigenic FGFR3-TACC3 gene fusion escapes mir-99a regulation in glioblastoma. J Clin Invest. 2013;123:855–65.
106. Li S, Chou AP, Chen W, Chen R, Deng Y, Phillips HS, et al. Overexpression of isocitrate dehydrogenase mutant proteins renders glioma cells more sensitive to radiation. Neuro Oncol. 2013;15:57–68.
107. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. Nature. 2008;454:436–44.
108. Allavena P, Germano G, Marchesi F, Mantovani A. Chemokines in cancer related inflammation. Exp Cell Res. 2011;317:664–73.
109. Guven-Maiorov E, Acuner-Ozbabacan SE, Keskin O, Gursoy A, Nussinov R. Structural pathways of cytokines may illuminate their roles in regulation of cancer development and immunotherapy. Cancers (Basel). 2014;6:663–83.
110. Ilyin SE, González-Gómez I, Gilles FH, Plata-Salamán CR. Interleukin-1 alpha (IL-1 alpha), IL-1 beta, IL-1 receptor type I, IL-1 receptor antagonist, and TGF-beta 1 mRNAs in pediatric astrocytomas, ependymomas, and primitive neuroectodermal tumors. Mol Chem Neuropathol. 1998;33: 125–37.
111. Sasaki A, Ishiuchi S, Kanda T, Hasegawa M, Nakazato Y. Analysis of interleukin-6 gene expression in primary human gliomas, glioblastoma xenografts, and glioblastoma cell lines. Brain Tumor Pathol. 2001;18: 13–21.
112. Plata-Salamán CR. Brain cytokines and disease. Acta Neuropsychiatrica. 2002;14:262–78.
113. Zhou Y, Larsen PH, Hao C, Yong VW. CXCR4 is a major chemokine receptor on glioma cells and mediates their survival. J Biol Chem. 2002;277:49481–87.
114. Kouno J, Nagai H, Nagahata T, Onda M, Yamaguchi H, Adachi K, et al. Up-regulation of CC chemokine, CCL3L1, and receptors, CCR3, CCR5 in human glioblastoma that promotes cell growth. J Neurooncol. 2004;70: 301–7.
115. Ludwig A, Schulte A, Schnack C, Hundhausen C, Reiss K, Brodway N, et al. Enhanced expression and shedding of the transmembrane chemokine CXCL16 by reactive astrocytes and glioma cells. J Neurochem. 2005;93:1293–303.
116. Sciumé G, Soriani A, Piccoli M, Frati L, Santoni A, Bernardini G. CX3CR1/CX3CL1 axis negatively controls glioma cell invasion and is modulated by transforming growth factor–1. Neuro Oncol. 2010;12: 701–10.
117. Yao X, Liu Y, Huang J, Zhou Y, Chen K, Gong W, et al. The role of chemoattractant receptors in the progression of glioma, Glioma - Exploring its biology and practical relevance. InTech, Anirban Ghosh (Ed.) 2011. doi: 10.5772/22154.
118. Zhou J, Xiang Y, Yoshimura T, Chen K, Gong W, Huang J, et al. The role of chemoattractant receptors in shaping the tumor microenvironment. Biomed Res Int. 2014;2014:751392. doi: 10.1155/2014/751392.
119. Talasila KM, Soentgerath A, Euskirchen P, Rosland GV, Wang J, Huszthy PC, et al. EGFR wild-type amplification and activation promote invasion and development of glioblastoma independent of angiogenesis. Acta Neuropathol. 2013;125:683–98.
120. Gong J, Zhu S, Zhang Y, Wang J. Interplay of VEGFa and MMP2 regulates invasion of glioblastoma. Tumour Biol. 2014;35:11879–85.
121. Bazan JF, Bacon KB, Hardiman G, Wang W, Soo K, Rossi D, et al. A new class of membrane-bound chemokine with a CX3C motif. Nature. 1997;385:640–4.
122. Imai T, Hieshima K, Haskell C, Baba M, Nagira M, Nishimura M, et al. Identification and molecular characterization of fractalkine receptor CX3CR1, which mediates both leukocyte migration and adhesion. Cell. 1997;91:521–30.
123. Marchesi F, Locatelli M, Solinas G, Erreni M, Allavena P, Mantovani A. Role of CX3CR1/CX3CL1 axis in primary and secondary involvement of the nervous system by cancer. J Neuroimmunol. 2010;224:39–44.
124. Lauro C, Catalano M, Trettel F, Mainiero F, Ciotti MT, Eusebi F, et al. The chemokine CX3CL1 reduces migration and increases adhesion of neurons with mechanisms dependent on the beta1 integrin subunit. J Immunol. 2006;177:7599–606.
125. Claes A, Idema AJ, Wesseling P. Diffuse glioma growth: a guerilla war. Acta Neuropathol. 2007;114:443–58.
126. Cheng Y, Pang JC, Ng HK, Ding M, Zhang SF, Zheng J, et al. Pilocytic astrocytomas do not show most of the genetic changes commonly seen in diffuse astrocytomas. Histopathology. 2000;37:437–44.
127. Yang L, Li N, Wang C, Yu Y, Yuan L, Zhang M, et al. Cyclin L2, a novel RNA polymerase II-associated cyclin, is involved in pre-mRNA splicing and induces apoptosis of human hepatocellular carcinoma cells. J Biol Chem. 2004;279:11639–48.
128. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456:66–72.

129. Bulfone A, Smiga SM, Shimamura K, Peterson A, Puelles L, Rubenstein JL. T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. Neuron. 1995;15:63–78.
130. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucl Acids Res. 2000;28:27–30.

4.2 Publication:

Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups

Journal: BMC Cancer

Received: 21 February 2017; **Accepted**: 20 March 2018; **Published**: 10 April 2018

Citation: Chris Lauber, Barbara Klink and Michael Seifert (2018): Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups. BMC Cancer, 18:399.

Placement and summary of the publication

Oligodendrogliomas represent a specific class of human brain tumors that were diagnosed based on the presence of specific histological features over many years (Louis et al. (2007)). With the availability of molecular data and diagnostic tests, characteristic molecular markers (1p/19q co-deletion, IDH1/2 mutation) are now considered to improve the diagnosis (Louis et al. (2016)), but our knowledge about molecular tumor subtypes and their differences at the level of signaling and metabolic pathways, underlying regulatory networks and involved major regulators is still limited. We therefore considered publicly available gene copy number, single nucleotide variation and transcriptome data of histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network (2015); Ceccarelli et al. (2016)) to reveal molecular subtypes and to characterize differences between them by utilizing well-established standard bioinformatics methods for statistical data analysis in combination with a network-based approach to predict altered major regulators.

Our study revealed three molecular subgroups for histologically classified oligodendrogliomas purely based on hierarchical clustering of their gene copy number profiles. Further

analysis of the IDH mutation status, associations with known Verhaak and G-CIMP subtypes and overall survival behavior of patients confirmed that these subgroups largely resembled known molecular glioma subtypes that were previously determined by a combination of different omics layers (The Cancer Genome Atlas Research Network (2015); Kamoun et al. (2016)). After the exclusion of the classical glioblastoma-like subgroup, we derived a signature of 5113 differentially expressed genes that distinguished histologically classified oligodendrogliomas with a concurrent 1p/19q co-deletion and an IDH mutation from those that predominantly showed an IDH mutation. These signature showed strong differences at the level of signaling and metabolic pathways including known pathways involved in glioma development (e.g. cell proliferation, differentiation, migration, cell-cell contacts). To further explore differences between both molecular subtypes, we also learned a gene regulatory network revealing putative major regulators with functions in cytoskeleton remodeling, apoptosis, and neural development potentially contributing to the manifestation of differences between both subgroups. We also identified characteristic differences in the expression of several HOX and SOX transcription factors between both subgroups indicating that different glioma stemness programs are active in both subgroups. This is also supported by single cell transcriptome analyses by Tirosh et al. (2016) and by Venteicher et al. (2017), which were both published during our work on this study.

Moreover, the considered oligodendroglioma data set represents an important resource, but one has to be aware that all oligodendrogliomas were classified by TCGA according to the WHO 2007 brain tumor classification system (Louis et al. (2007)). This histology-based system is relatively error-prone (Coons et al. (1997); van den Bent (2010)) and has therefore been replaced by the WHO 2016 brain tumor classification system that uses a combination of histological and molecular markers (Louis et al. (2016)). To address this, we discussed our findings in the context of this new classification system. This is important for the interpretation of our results and supports others that want to work with the TCGA lower-grade glioma data set.

Our study demonstrated that gene copy number profiles alone can be used to derive known molecular subgroups of histologically classified oligodendrogliomas. In addition, the revealed potential major regulators and the characteristic differences in the activity of potential stemness programs provide a basis for future experimental validation studies.

Author contribution

I designed the concept of the study. I supervised the implementation of the methods and the realization of the computational analysis by Chris Lauber, who was a postdoc in my group. I developed the structure of the manuscript and wrote the manuscript together with Chris Lauber. I discussed all findings with Barbara Klink, who supported the biological interpretation of our results. I performed the revisions of the manuscript.

Lauber *et al. BMC Cancer* (2018) 18:399
https://doi.org/10.1186/s12885-018-4251-7

BMC Cancer

RESEARCH ARTICLE Open Access

Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups

Chris Lauber[1], Barbara Klink[2,3] and Michael Seifert[1,3*]

Abstract

Background: Molecular data of histologically classified oligodendrogliomas are available offering the possibility to stratify these human brain tumors into clinically relevant molecular subtypes.

Methods: Gene copy number, mutation, and expression data of 193 histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) were analyzed by well-established computational approaches (unsupervised clustering, statistical testing, network inference).

Results: We applied hierarchical clustering to tumor gene copy number profiles and revealed three molecular subgroups within histologically classified oligodendrogliomas. We further screened these subgroups for molecular glioma markers (1p/19q co-deletion, *IDH* mutation, gain of chromosome 7 and loss of chromosome 10) and found that our subgroups largely resemble known molecular glioma subtypes. We excluded glioblastoma-like tumors (7a10d subgroup) and derived a gene expression signature distinguishing histologically classified oligodendrogliomas with concurrent 1p/19q co-deletion and *IDH* mutation (1p/19q subgroup) from those with predominant *IDH* mutation alone (IDHme subgroup). Interestingly, many signature genes were part of signaling pathways involved in the regulation of cell proliferation, differentiation, migration, and cell-cell contacts. We further learned a gene regulatory network associated with the gene expression signature revealing novel putative major regulators with functions in cytoskeleton remodeling (e.g. *APBB1IP*, *VAV1*, *ARPC1B*), apoptosis (*CCNL2*, *CREB3L1*), and neural development (e.g. *MYTIL*, *SCRT1*, *MEF2C*) potentially contributing to the manifestation of differences between both subgroups. Moreover, we revealed characteristic expression differences of several *HOX* and *SOX* transcription factors suggesting the activity of different glioma stemness programs in both subgroups.

Conclusions: We show that gene copy number profiles alone are sufficient to derive molecular subgroups of histologically classified oligodendrogliomas that are well-embedded into general glioma classification schemes. Moreover, our revealed novel putative major regulators and characteristic stemness signatures indicate that different developmental programs might be active in these subgroups, providing a basis for future studies.

Keywords: Histologically classified oligodendrogliomas, Molecular subgroup, Gene expression signature, Gene regulatory network, Bioinformatics, Computational systems biology

*Correspondence: michael.seifert@tu-dresden.de
[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
[3]National Center for Tumor Diseases, Dresden, Germany
Full list of author information is available at the end of the article

Background

Oligodendrogliomas belong to the class of diffuse gliomas that represent the most frequent primary brain tumors in adults [1]. About 4 to 8% of all diagnosed tumors of the central nervous system are oligodendrogliomas [2]. Diffuse gliomas are generally characterized by infiltration of the surrounding brain tissue, and fast progression and relapse are common [3]. Traditionally, histological similarities to normal glial cells (astrocytes and oligodendrocytes) were used to distinguish between different types of diffuse gliomas according to the World Health Organization (WHO) 2007 grading system [4]. Known downsides of this histological classification include a considerable variability of diagnoses between neuropathologists and difficulties in discriminating oligodendrogliomas from other types of diffuse gliomas like astrocytomas and "mixed-type" oligoastrocytomas, which complicates diagnostics and treatment decisions for individual patients [5, 6]. These challenges led to the exploration of molecular markers for glioma diagnostics [7]. The majority of oligodendrogliomas shows a characteristic allelic loss of chromosomal arms 1p and 19q (1p/19q) that contributes to better chemotherapy sensitivity and longer recurrence-free survival [8, 9]. Three different gene expression subtypes of 1p/19q co-deleted oligodendrogliomas have recently been revealed, but the analysis of the clinical relevance of these subtypes requires additional studies [10]. Further, specific heterozygous somatic point mutations of the isocitrate dehydrogenase gene (IDH1/2) were found in more than three-fourths of all oligodendrogliomas and nearly three-fourths of all astrocytomas of WHO grades II and III [11–13] and in all 1p/19q codeleted gliomas [14]. These mutations are associated with the glioma-CpG island methylator phenotype (G-CIMP) [15, 16] and with a better prognosis compared to IDH wild-type tumors [11, 17].

These molecular markers were integrated into a recent update of the classification of tumors of the central nervous system by the WHO [18]. As a consequence, some diffuse glioma classes became obsolete, like the "mixed-type" oligoastrocytomas that should now be classified as either oligodendrogliomas or astrocytomas. According to this new classification, oligodendrogliomas are characterized by the co-occurrence of the mutation of *IDH1/2* and the 1p/19q co-deletion. Notably, this class does not accommodate *IDH*-mutated tumors with 1p/19q wild-type that were classified as oligodendrogliomas based on histology before. Such discrepancies between histological and molecular tumor classification still remain a great challenge for further improvements of glioma diagnostics, but in terms of prognosis molecular markers can outweigh histological characteristics. Recently, it has been shown that glioma subgroups can be defined based on *IDH* mutation and 1p/19q co-deletion status deriving genetic subgroups that are more reflective of disease subtypes than glioma classes defined by histology [19]. These results were further refined through the analysis of DNA methylation profiles revealing clinically relevant molecular subtypes [20]. In addition, single cell transcriptome data has allowed to gain novel insights into the molecular architecture of oligodendrogliomas showing that the majority of tumor cells express either a specialized astrocyte-like or oligodendrocyte-like program, whereas a subpopulation of cells remains undifferentiated and is associated with a neural stem cell expression program that most likely drives tumor development [21]. This has been further extended by analyzing single cell transcriptomes of oligodendrogliomas and astrocytomas suggesting a common stemness program for both tumor types that drives tumor growth, whereas differences between both types are mainly driven by the tumor microenvironemt and specific genetic signatures [22]. This has important consequences for the clinical management of oligodendrogliomas and may also explain in part differences between molecular and histological classifications. All these and many other studies have greatly contributed to a better understanding of molecular characteristics of oligodendrogliomas. Still, also in the light of differences between histological and molecular classifications, our knowledge about specific molecular characteristics of oligodendrogliomas is incomplete.

Here, we present an in-depth computational analysis of histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) revealing novel differences between molecular subgroups at the level of individual genes, pathways, and gene regulatory networks. We first stratified these tumors based on their gene copy number profiles into three subgroups utilizing unsupervised clustering. Additional screening for the presence of known glioma markers showed that these subgroups largely resembled already known molecular glioma subtypes. To further characterize molecular differences, we derived a signature of differentially expressed genes distinguishing tumors with 1p/19 co-deletion and *IDH* mutation from tumors that predominantly showed an *IDH* mutation. We further learned a gene regulatory network that is capable to explain this observed expression signature. This enabled us to identify novel putative major regulators that are potentially involved in the manifestation of differences between both subgroups. Interestingly, this network also contained a characteristic expression signature of *HOX* and *SOX* genes that distinguishes both subgroups indicating the activity of different glioma stemness programs.

Methods

Molecular data of oligodendrogliomas and normal brains

DNA copy number, RNA-seq gene expression, and somatic mutation data was obtained from the TCGA

data portal (https://gdc.cancer.gov/) for 193 histologically classified oligodendrogliomas of the TCGA lower grade glioma (LGG) cohort (Additional file 1). The vast majority of tumor samples represented primary tumors, except five recurrent tumors. We determined gene-specific copy number log-ratios for each oligodendroglioma based on its corresponding DNA copy number profile (see [23] for details). Three commercially available normal brain samples were obtained from StrataGen, BioChain, and Clonetech for which RNA-seq gene expression has been measured previously. All considered gene copy (17,677 genes) and gene expression (15,988 genes) profiles are provided in Additional file 2.

Clustering based on CNV data

Hierarchical clustering (euclidean distance, complete linkage) of tumors was done in R using the processed gene copy number variation (CNV) log-ratio data of tumor compared to normal. One obvious outlier (TCGA-P5-A5F6-01A) was removed from subsequent analyses. Three tumor subgroups were derived by cutting the clustering dendrogram into three sub-trees. These subgroups were named taking into account the following molecular properties: (i) 1p/19q - co-deletion of chromosomal arms 1p and 19q and presence of characteristic *IDH1/2* mutation, (ii) IDHme - predominance of *IDH1/2* mutation but no co-deletion of 1p and 19q, and (iii) 7a10d - no co-deletion of 1p and 19q, lack of *IDH1/2* mutations, amplification of chromosome 7, and deletion of chromosome 10.

Data normalization and identification of differentially expressed genes

Raw RNA-seq gene expression counts were loaded into R. Combined normalization of tumor and normal brain RNA-seq data was done using the voom function of the limma package [24] with normalization method cyclic loess. Differential gene expression analysis between CNV-derived tumor subgroups was done following limma's standard workflow. Differentially expressed (signature) genes were selected using an FDR-adjusted p-value (q-value) [25] cut-off of 0.01.

Verhaak and G-CIMP classification

Gene expression $\log_2$-ratios of genes in tumor compared to the average expression in normal brain tissue were computed for each oligodendroglioma sample. 756 of 840 genes that were used to derive the four Verhaak classes [26] were part of our data set. We calculated pearson correlation and associated p-values between the gene expression log-ratios in the glioma reference set and our tumor subgroups. Similarly, 42 of 50 genes of the glioma-CpG island methylator phenotype (G-CIMP) set [15] were part of our data set, for which we calculated Pearson correlations and p-values. Note that genes missing from the Verhaak and G-CIMP signature do not strongly affect the classification, because there are other genes in these signatures that show expression levels that are strongly correlated with those of the missing genes [27].

Survival analysis

Information about days to death or days to last follow-up was taken from Table S1 of [20]. This table represented the most recent survival information in months at the time of our study. We transformed the survival information from months into days using the factor 30.4167 followed by a rounding to the nearest integer (Additional file 1). We generated survival curves and performed log-rank tests using the R package survival [28].

Gene and pathway annotation enrichment analysis

Gene, signaling, and metabolome pathway annotations were obtained from [23]. The number of signature genes per annotation category was counted separately for up- and downregulated genes, and the significance of gene enrichment was calculated using Fisher's exact test.

Signature-specific regulatory network inference

We inferred transcriptional regulatory networks associated with the normalized expression of the signature genes that discriminate between the 1p/19q and IDHme subgroups following the approach detailed in [27] with few modifications. We constructed two types of networks that differed in the set of predictor variables: (i) only the gene copy number of a signature gene was used to predict its own expression and (ii) in addition to the copy numbers, the gene expression of all signature genes that were annotated as transcription factors (TFs) were used to predict the expression of a signature gene. The expression value of a particular TF was excluded from its own prediction in the latter analysis. For each signature gene, lasso (least absolute shrinkage and selection operator) regression [29] and a significance test for lasso [30] were used to estimate the coefficients and their corresponding significance for each predictor of the underlying signature gene-specific linear model as implemented in [31]. We only considered the most significant predictors with p-values less than 5×10^{-5} specified by the standard detection limit of the covariance test implementation [30]. We further validated each network through cross-validation by repeated random subsampling. To this end, the data was randomly partitioned into a training set constituting two-third of the tumors on which the network was constructed and a test set constituting the remaining one-third of tumors for which the expression of the signature genes was predicted and compared to the experimentally measured expression. This was repeated 100 times. To assess prediction accuracy we calculated pearson correlation of predicted and measured gene expression averaged over the 100

networks. For network visualization we only kept links that occurred in at least 75% of the 100 networks.

Results

Gene copy number variations and *IDH* mutations characterize three molecular subgroups of histologically classified oligodendrogliomas

It has been shown previously that the majority of histologically classified oligodendrogliomas has a co-deletion of chromosomal arms 1p and 19q and a characteristic mutation of *IDH1/2* [19, 32]. We thus analyzed genome-wide gene copy number data that were available for 193 histologically classified oligodendrogliomas from TCGA (Additional files 1 and 2). Unsupervised clustering of the tumors based on their CNV profiles alone revealed three subgroups (Fig. 1). We further analyzed the mutation status of *IDH1/2* of tumors in these subgroups (Table 1). The largest subgroup comprised 133 tumors (68.9%) and showed the characteristic 1p/19q co-deletion as well as *IDH1* or *IDH2* mutations in each tumor. We refer to this subgroup as 1p/19q. The second largest subgroup included 45 tumors (23.3%) that showed no obvious pattern of gene deletions or amplifications. Since the majority of tumors in this subgroup had an *IDH1/2* mutation (82%), we named this subgroup IDH mutation-enriched (IDHme). The third subgroup comprised 15 tumors (7.8%) that were characterized by an amplification of chromosome 7 and a deletion of chromosome 10 as typically observed in classical glioblastomas [3]. Only three tumors in this subgroup had an *IDH1* or *IDH2* mutation (20%).

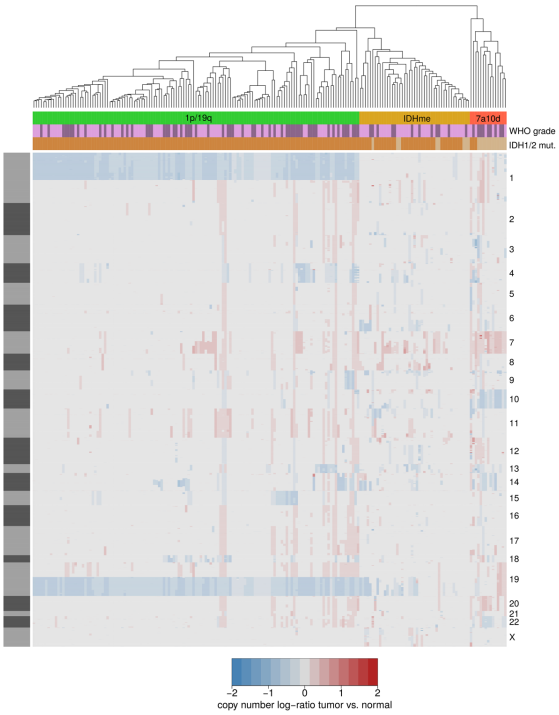

Fig. 1 CNV-derived molecular subgroups of histologically classified oligodendrogliomas. Heatmap of genome-wide gene copy number log-ratios of 193 histologically classified oligodendrogliomas (columns) compared to normal brain for 17677 genes (rows); blue: deletions, gray: unchanged, red: amplifications. Chromosomes are highlighted by alternating gray and black bars to the left; chromosome midpoints are indicated by labels to the right. Oligodendroglioma subgroups were revealed by unsupervised clustering and are shown using green (1p/19q), yellow (IDHme), and red (7a10d) column coloring. Tumors of WHO grade II are labeled in light purple. WHO grade III is labeled in dark purple. Tumors with an IDH1/2 mutation are labeled in dark brown. The absence of an IDH mutation is labeled in light brown

Table 1 Frequency of mutations of known cancer-relevant genes per oligodendroglioma subgroup

Gene	Mutated	1p/19q	IDHme	7a10d
IDH1/2	Yes	133	37	3
	No	0	8	12
TP53	Yes	7	35	6
	No	126	10	9
ATRX	Yes	3	30	3
	No	130	15	12
CIC	Yes	85	2	0
	No	48	43	15
FUBP1	Yes	38	0	0
	No	95	45	15
NOTCH1	Yes	31	3	0
	No	102	42	15

We refer to this subgroup as 7a10d. It is important to note that the 7a10d subgroup formed an own subcluster that is relatively distant to the 1p/19q and IDHme subgroups, which were both part of one larger subcluster (Fig. 1).

Tumors of the three subgroups differ in mutational status of other cancer-relevant genes

We further observed differences in mutational profiles of known glioma-relevant genes (*TP53*, *ATRX*, *CIC*, *FUBP1*, *NOTCH1* [3, 19]) between tumors of the three subgroups (Table 1). Only 5% and 2% of the 1p/19q tumors showed a mutation of, respectively, *TP53* and *ATRX*, while about two-third of the IDHme tumors had at least one of these two genes mutated. For 7a10d tumors, these numbers were 40% and 25%, respectively. In contrast, *CIC* and *FUBP1* were relatively frequently mutated in the 1p/19q subgroup (64% and 29%, respectively), but only two *CIC* and no *FUBP1* mutations were observed in the IDHme tumors and none of the 7a10d tumors showed *CIC* and *FUBP1* mutations. Also for *NOTCH1* the IDHme and 7a10d subgroups resemble each other in terms of mutation frequency (7% and 0%, respectively), while about one-fourth of the 1p/19q tumors showed a *NOTCH1* mutation.

Subgroup 7a10d differs in Verhaak and G-CIMP subtype classification and patient survival from 1p/19q and IDHme

In order to explore whether tumors of the three oligodendroglioma subgroups differ in their gene expression profiles compared to known molecular glioma subtypes we first considered the Verhaak subtypes [26]. We computed the correlation between the given signature-specific expression levels of the Verhaak subtypes and the corresponding gene expression levels of each individual oligodendroglioma. We observed moderate but still significant correlation values with the Verhaak subtypes for the vast majority of tumors (P <0.05 for 130 of 133 1p/19q tumors, for 43 of 45 IDHme tumors, and for all 7a10d tumors considering the Verhaak subtype with the strongest correlation). The 1p/19q and IDHme subgroups showed a similar association pattern (Fig. 2a top and middle). Tumors in both subgroups have highest similarity to the proneural and classical subtypes followed by the mesenchymal subtype, while there is generally a negative correlation with the neural subtype. In contrast, the vast majority of tumors in the 7a10d subgroup had a negative correlation with the proneural subtype (Fig. 2a bottom). This is expected for tumors without an *IDH* mutation [15].

In a similar analysis, we compared the associations of the three oligodendroglioma subgroups with the expression signature of the G-CIMP subtype driven by the mutation of *IDH* [15]. Like for the Verhaak classification, the 1p/19q and IDHme subgroups resembled each other and the tumors in these subgroups had generally positive correlation values to G-CIMP (P <0.1 for 73 of 133 1p/19q tumors and 27 of 45 IDHme tumors), as opposed to 7a10d tumors that showed no or very weak positive and negative correlation (P <0.1 for 2 of 15 tumors, Fig. 2b).

We also analyzed whether there are differences in patient survival between the three subgroups by using the clinical data available for 125 1p/19q, 34 IDHme, and 15 7a10d tumors. Patients from the 1p/19q and IDHme subgroups showed no differences in survival (Fig. 2c top and middle, log-rank test, $P = 0.7843$). In sharp contrast, patients from 7a10d showed significantly shorter survival than patients from the 1p/19q and IDHme subgroups (Fig. 2c bottom, log-rank tests, $P = 4.9 \times 10^{-6}$ and $P = 1.1 \times 10^{-4}$, respectively) consistent with previous findings [19].

All three subgroups are part of known glioma subtypes

Recent studies have defined molecular subtypes for gliomas [19, 20]. We thus analyzed how our three subgroups 1p/19q, IDHme, and 7a10d observed for histologically classified oligodendrogliomas are embedded in these general classification schemes. Diffuse gliomas were grouped into three major subtypes based on the IDH mutation status and the presence of the 1p/19q co-deletion in [19]. Our 1p/19q subgroup corresponds to the 1p/19q subtype in [19]. The IDHme subgroup is included in the subtype that has no 1p/19q co-deletion but an IDH mutation in [19]. The 7a10d subgroup is included in the subtype that has no IDH mutation and no 1p/19q co-deletion, which contains gliomas of which about 50% showed a gain of chromosome 7 and a loss of chromosome 10 [19]. Further, our purely CNV-based derivation

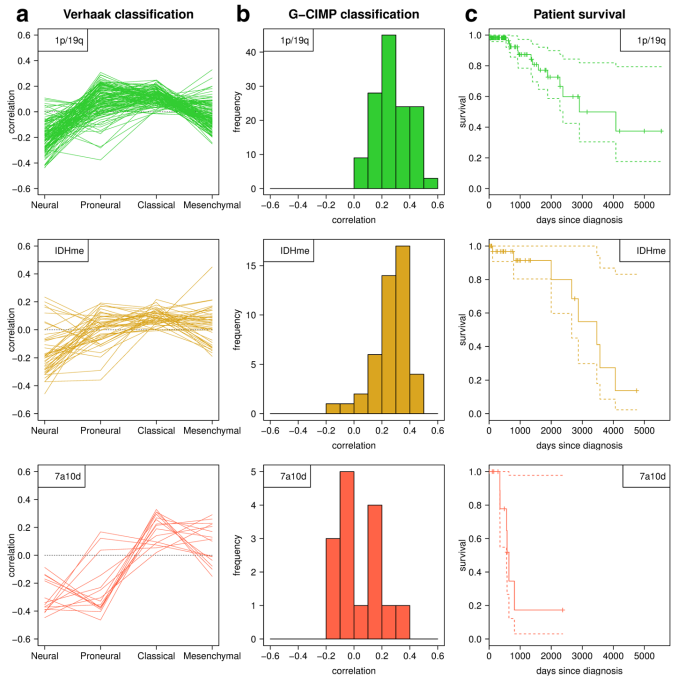

Fig. 2 Oligodendroglioma subgroup comparison. **a** Pearson correlation between the expression log-ratios of 756 marker genes in the Verhaak glioma reference set and our tumor subgroups is shown for the four Verhaak classes. **b** Pearson correlation between the expression log-ratios of 42 genes in the G-CIMP reference set and our tumor subgroups. **c** Kaplan-Meier curves (solid lines) and 95%-confidence intervals (dashed lines) of patient survival according to the corresponding clinical data

of the three subgroups (Fig. 1) shows that tumors with an IDH mutation are more similar to each other than tumors without an IDH mutation. This is in accordance with [19]. Also highly similar gene mutation patterns and survival times are observed for our subgroups and those by [19].

The classification scheme in [19] has been refined in [20] subdividing the IDH mutant group into a G-CIMP-low, G-CIMP-high, and a 1p/19q co-deletion subtype. Our 1p/19q subgroup is included in the 1p/19q co-deletion group in [20]. Further, the vast majority of tumors in our IDHme subgroup belong to the G-CIMP-high group in [20] indicated by the observation of positive correlations with the G-CIMP subtype in our analysis (Fig. 2b middle). Only four IDHme tumors may belong to the G-CIMP-low subtype (correlation with G-CIMP less than 0.1, Fig. 2b middle). This is in good accordance with the molecular classification of histologically classified oligodendrogliomas by [20]. In addition, the non-IDH mutant group was further subdivided in [20] into a classic-like, mesenchymal-like, and two other subtypes. Tumors of our 7a10d subgroup are represented by these subtypes. About half of the 7a10d tumors belong to the classic-like group (Fig. 2a bottom). The majority of the remaining tumors belong to the mesenchymal-like group, but they also show a relatively strong correlation with the classical group (Fig. 2a bottom). This is similar to [20] where also a large proportion of the tumors in the mesenchymal-like group were classified to belong to the classical group of Verhaak [26].

We further tested if the three subgroups were well-embedded in molecular data of closely related histologically classified oligoastrocytomas and astrocytomas of the TCGA lower grade glioma cohort. Therefore, we performed unsupervised clustering of the gene copy number profiles and found that all three subgroups were present among the oligoastrocytomas and that the astrocytomas were split up into the IDHme and 7a10d subgroup. In addition, Verhaak and G-CIMP subtype classifications,

patient survival, and gene expression behavior were highly similar between the oligodendroglioma subgroups and corresponding subgroups of oligoastrocytomas and astrocytomas (Additional file 3). This clearly indicates that each of our derived subgroups was adequately covered based on molecular data of histologically classified oligodendrogliomas.

Generally, strong differences in chromosomal mutations, subtype characteristics, and patient survival between the 7a10d subgroup and the other two subgroups 1p/19q and IDHme (Figs. 1 and 2) indicate that 7a10d tumors rather resemble glioblastoma-like tumors [3, 19, 20]. We therefore focused our further analysis on the comparison of tumors from the 1p/19q and IDHme subgroups.

A signature of differential gene expression discriminates 1p/19q from IDHme

To compare genome-wide gene expression profiles of the 1p/19q and IDHme subgroups we conducted a differential gene expression analysis contrasting these two subgroups. Using a q-value cut-off of 0.01 we identified 5113 genes to be differentially expressed between 1p/19q and IDHme (Fig. 3, Additional file 4). The expression of half of the signature genes was downregulated in 1p/19q compared to IDHme, while the other half was upregulated. When comparing tumors of grade II and grade III within each subgroup we found no large-scale differences. Only 104 signature genes where differentially expressed between tumor grades II and III for the 1p/19q subgroup (73 grade II vs. 60 grade III tumors, Additional file 5), while there were no significant expression differences of signature genes between tumor grades II and III for the IDHme subgroup (33 grade II vs. 12 grade III tumors).

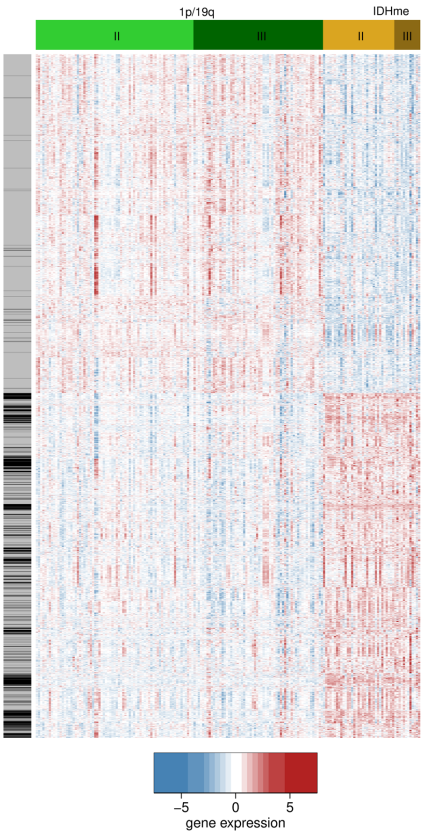

Fig. 3 Gene signature distinguishing 1p/19q from IDHme. The heatmap shows the expression values of 5113 differentially expressed (q-value <0.01) signature genes (rows) for 178 histologically classified oligodendrogliomas (columns). Rows are Z-score-scaled and ordered based on a hierarchical clustering of the data (dendrogram not shown). Subgroups are shown in green (1p/19q) and yellow (IDHme) with tumor grades highlighted using light (grade II) and dark (grade III) shadings. Genes that are located on chromosomal arms 1p or 19q are indicated by black lines to the left of the heatmap; other genes are in gray

The signature is enriched for signaling and metabolic pathway genes and transcription factors

Looking at the annotations of the 5113 signature genes we found that the categories transcription factor/cofactor, kinase, phosphatase, signaling pathway gene, and tumor suppressor gene were significantly enriched for downregulated genes in tumors of the 1p/19q subgroup compared to that of the IDHme subgroup (P <0.05, Fig. 4). For signature genes upregulated in 1p/19q compared to IDHme only the transcription factor/cofactor category was found to be significantly enriched (P <0.1). Among the affected signaling pathways several pathways known to be involved in cancer were significantly enriched with genes downregulated in 1p/19q tumors compared to IDHme (Fig. 4c). These were the MAPK signaling, ErbB signaling, mTOR signaling, PI3k-Akt signaling, Apoptosis, Wnt signaling, TGF-Beta signaling, VEGF signaling, Focal adhesion, Adherence junction, Jak-STAT signaling, and Hedgehog signaling pathway, which are known to affect proliferation, differentiation, migration, adhesion, cell growth and survival, cell cycle arrest and progression, and metabolism (see Table S4 in [33]). For genes upregulated in 1p/19q no enrichment of signaling pathways was observed.

Regarding metabolic pathways (Fig. 4b), the pentose phosphate pathway (generating NADPH, pentoses, and

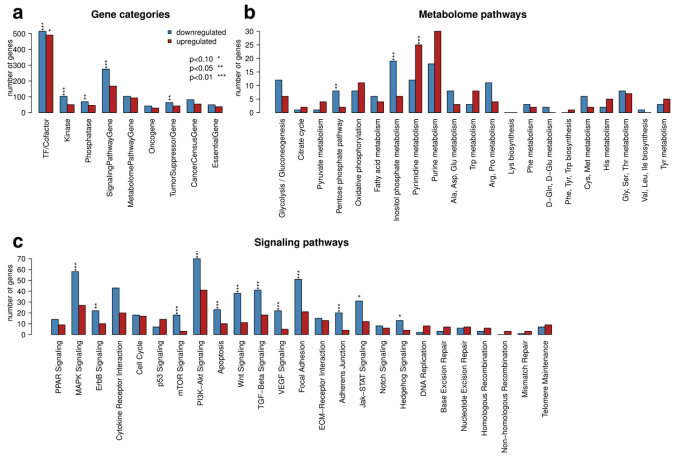

Fig. 4 Functional analysis of signature genes. Enrichment of gene categories (**a**), metabolic pathways (**b**), and signaling pathways (**c**) with signature genes. Bars are shown separately for genes down- and upregulated in 1p/19q compared to IDHme using blue and red color, respectively. Significance of enrichment was calculated using Fisher's exact test and is highlighted by asterisks symbols

Ribose 5-phosphate, a precursor for nucleotide synthesis) and inositol phosphate pathway (generating inositol phosphates that play a role in various cellular processes including cell growth and differentiation, cell migration and apoptosis) were significantly enriched with genes downregulated in 1p/19q (P <0.05 and P <0.01, respectively). The pyrimidine pathway (generating cytosine, thymine, and uridine nucleotides) was enriched with genes showing an increased expression in 1p/19q tumors (P <0.01).

Moreover, there were in total 1006 transcription factors/cofactors present in the signature (Additional file 6), forming the basis for the subsequent reconstruction of a gene regulatory network that is associated with the observed expression differences between the 1p/19q and IDHme subgroups.

A gene regulatory network is associated with expression differences between 1p/19q and IDHme

We sought to construct a gene regulatory network which can predict the expression of the 5113 signature genes distinguishing 1p/19q from IDHme. In this analysis, 100 cross-validated networks were computed and used to calculate an average predicted expression value for each signature gene (see "Methods" for details). We applied the approach to two different predictor sets. First, we started to learn a network for which only the copy number of a gene was used to predict its expression. For 1442 signature genes (28.2%) no prediction of gene expression based on the underlying gene copy number was obtained. For the vast majority of the remaining signature genes the average predicted expression value correlated positively with the measured expression for the test data (Fig. 5a), and the median correlation coefficient over all signature genes was 0.292 (P <0.05 for 53.7% of the genes).

In the second analysis, we learned a regulatory network by utilizing both the gene-specific copy numbers and the expression values of transcription factors that were part of the signature as predictors. This network yielded significantly better predictions than the CNV-only network (Fig. 5a, Mann-Whitney U test, $P \approx 0$). Predictions were obtained for all signature genes, and the median correlation coefficient was 0.676 on the test data (P <0.05 for 95.8% of the genes). We chose this second network (Additional file 7) for further analysis because of its superior prediction accuracy and the possibility to identify potential regulators of other signature genes.

Hubs in the network, e.g. nodes with high degree that have many connections to other nodes, may help to identify potential key regulators involved in the manifestation of differences between the 1p/19q and IDHme subgroups. We thus looked at the out-degree of nodes representing transcription factors and found that few of them (49 of 1006, 4.9%) had an out-degree of at least 10, while the vast majority were connected to few signature genes (Fig. 5b). A sub-network containing only these hub transcription factors and the signature genes connected to them by ingoing or outgoing links is shown in Fig. 6. The vast majority of network connections represent activating

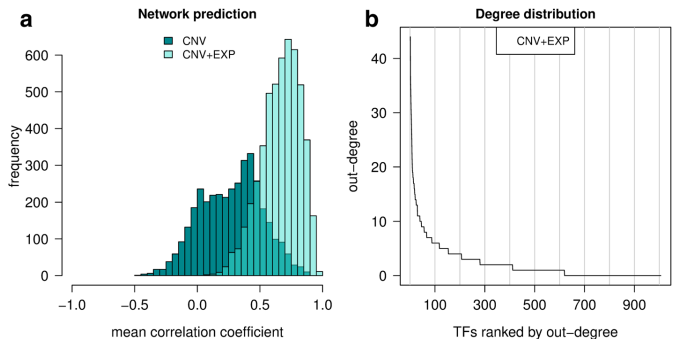

Fig. 5 Regulatory network predictions and hub transcription factors. **a** Pearson correlations between predicted and measured expression of the 5113 signature genes on the test data are shown for the network utilizing only gene copy number data (CNV: dark-turquoise) and the network utilizing both gene copy number and gene expression data (CNV + EXP: light-turquoise) as predictors. Correlation values are averaged over 100 cross-validation iterations (see Methods for details). **b** Ranking of transcription factors by out-degree for the CNV + EXP network. Only very few transcription factors have high out-degree values (hubs), whereas the large majority shows few connections to other signature genes

links. Moreover, this sub-network can be further partitioned into potential gene regulatory modules that (i) show many internal connections, (ii) have few or no external links to other gene clusters, and (iii) comprise signature genes with comparable patterns of expression differences between 1p/19q and IDHme (see node coloring in Fig. 6).

Regulatory hubs and gene network modules affect cancer-relevant functions

One of the gene modules in our regulatory network (Fig. 6) contains *APBB1IP*, the gene with the highest out-degree in the network, as well as other hub transcription factors including *VAV1*, *ARPC1B*, *SPI1*, *TFEC*, *FERMT3*, and *IKZF1*, among others. The expression of genes in this cluster is downregulated in 1p/19q compared to IDHme. According to GeneCards [34] and UniProtKB/Swiss-Prot [35] annotations, *APBB1IP* functions in signal transduction from Ras activation to actin cytoskeletal remodeling [36, 37], *VAV1* is a guanine nucleotide exchange factor for Rho family GTPases also known to be involved in the regulation of cytoskeletal rearrangements and a known proto-oncogene [38], *ARPC1B* regulates actin polymerization and mediates the formation of branched actin networks [39], *SPI1* is a proto-oncogene potentially involved in the regulation of pre-mRNA splicing [40], *TFEC* has been associated with breast cancer and is part of the cancer-related C-MYB transcription factor network [41], *FERMT3* has been associated with cell adhesion deficiencies [42], and *IKZF1* is known to be involved in different types of cancer [43].

A second gene module includes the hub transcription factors *CDK5R2*, *MYT1L*, *CELF3*, *RGS7*, and *SCRT1* (Fig. 6). In contrast to the first gene cluster described above, the expression of genes in this second cluster is upregulated in 1p/19q compared to IDHme. *CDK5R2* is a regulator of the cell division protein Cyclin-dependent kinase 5 and has been associated with neuronal migration and development [44], *MYT1L* is a pan-neural transcription factor involved in neuronal differentiation and is thought to play a role in the development of neurons and oligodendroglia [35], *CELF3* is involved in the regulation of pre-mRNA alternative splicing [45], *RGS7* is associated with benign neoplasms in different organs and regulates G-protein-coupled receptor signaling [46], and *SCRT1* is a Zinc finger DNA-binding protein critical for neuronal differentiation [47].

There are other individual hub transcription factors in the network with potentially relevant functions in cancer development. One of them is *PHB* (upregulated in 1p/19q compared to IDHme) that codes for prohibitin, which inhibits DNA synthesis, has been associated with breast cancer, and plays a role in regulating proliferation [48, 49]. *CREB3L1* (upregulated in 1p/19q) is thought to be involved in the protection of astrocytes from ER stress-induced cell death [50]. *CENPT* (upregulated in 1p/19q) encodes one of the inner kinetochore proteins and is required for normal chromosome organization and progress through mitosis [51]. *MEF2C* (dowregulated in 1p/19q) is crucial for normal neuronal development and has been suggested to be involved in neurogenesis and in the development of cortical architecture [52, 53]. *EIF3K* (downregulated in 1p/19q) is a component of the eukaryotic translation initiation complex regulating protein synthesis [54]. *CCNL2* (downregulated in 1p/19q) regulates a critical factor involved in cell apoptosis [55].

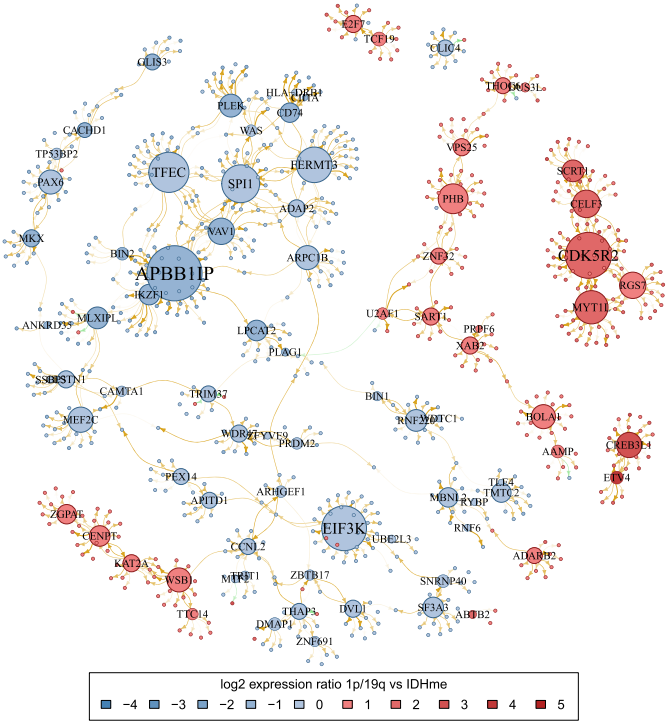

Fig. 6 Gene regulatory network distinguishing 1p/19q from IDHme Sub-network showing transcription factors with an out-degree of at least 10 and the signature genes connected to them. Genes (nodes) are colored according to gene expression log-ratios between the average expression in 1p/19q and IDHme. Nodes with labels represent transcription factors and the size of a node is proportional to its out-degree. Activating and repressing links are shown in yellow and green color, respectively, and transparency of a link is inversely proportional to the number of times the link is present across the 100 cross-validation iterations

Further, *ETV4* involved in developmental processes and oncogenesis [34] was upregulated in 1p/19q compared to IDHme.

Comparison of 1p/19q and IDHme to closely related oligodendrogliomas and astrocytomas

Recently, bulk and single cell transcriptomes of *IDH*-mutant oligodendrogliomas and astrocytomas have been compared [22]. This study suggested shared glial lineages and developmental hierarchies where most differences resulted from characteristic mutations and microenvironmental compositions. In more detail, they observed that differences in bulk gene expression profiles between oligodendrogliomas and astrocytomas can be primarily explained by the impact of characteristic tumor class-specific mutations (oligodendrogliomas: 1p/19q co-deletion, *CIC* mutations; astrocytomas: *TP53* mutations) and differences in the composition of the tumor microenvironment, but not by distinct expression programs of glial lineages of malignant cells. They compared oligodendrogliomas defined based on their histology and the presence of the 1p/19q co-deletion to astrocytomas defined based on their histology and the presence of mutations in *TP53* or *ATRX*. This is similar to our analysis. Our 1p/19q subgroup has the same histological and genetic features as their oligodendrogliomas. Our IDHme subgroup is closely related to their astrocytomas, except for differences in histology. In accordance with [22], we observed downregulations of genes on 1p and 19q (Fig. 3) and upregulations of genes of the p53 signaling pathway (Fig. 4c) in 1p/19q compared to IDHme. We found similar evidences that genes involved in cytoskeleton remodeling (e.g. *APBB1IP*,

VAV1, *ARPC1B*, Fig. 6) were downregulated in 1p/19q compared to IDHme, which might indicate potentially existing morphological differences. Further, we found significant expression differences between 1p/19q and IDHme analyzing oligodendrocyte-like and astrocyte-like expression programs from [22] (Additional file 8: Figure S1A-B, t-test, $P = 4.8 \times 10^{-11}$). The 1p/19q subgroup showed higher expression of genes of the oligodendrocyte-like program than the IDHme subgroup, whereas IDHme showed higher expression of genes of the astrocyte-like program. Similarly, both groups also differed in their expression of microglia/macrophage marker genes (Additional file 8: Figure S1C, t-test, $P < 0.03$). Interestingly, we found a weak trend that the 1p/19q and IDHme subgroups differ in the expression of the stemness program from [22]. Still, the majority of genes of the stemness program showed similar expression levels in both groups, but there were several genes with stronger expression differences (Additional file 8: Figure S1D). This included genes involved in cytoskeleton remodeling (absolute average log-ratio for 1p/19q compared to IDHme >1; *DCX*,*TMSB15A*: upregulated in 1p/19q; *FNBP1L*: downregulated in 1p/19q) and *MYT1L*, a known key factor of neural differentiation, upregulated in 1p/19q compared to IDHme.

1p/19q and IDHme tumors differ in stemness programs

Glioma stemness programs have been characterized over the last years suggesting important regulatory roles for different members of the *HOX* [20, 56] and *SOX* [20–22, 57] gene families. Roles of *SOX* genes in development and pathology have been reviewed in [58]. We thus analyzed our regulatory network (Additional file 7) for characteristic expression differences of both gene families between 1p/19q and IDHme. Our network includes seven *HOX* genes (*HOXA4*, *HOXA5*, *HOXA6*, *HOXA7*, *HOXA11*, *HOXA13*, *HOXC4*) and four *SOX* genes (*SOX6*, *SOX8*, *SOX12*, *SOX13*). Interestingly, all *HOX* genes were downregulated in 1p/19q compared to IDHme, whereas all *SOX* genes were upregulated in 1p/19q compared to IDHme (Fig. 7). This indicates the activity of different stemness programs between 1p/19q (potentially *SOX*-driven) and IDHme (potentially *HOX*-driven) tumors.

Moreover, this is also supported by already known cancer-relevant functions of different genes. *HOXA4* overexpression suppressed cell motility and spreading in ovarian cancer [59]. *HOXA5* downregulation increased stemness, cell plasticity and aggressiveness of breast cancer [60], and upregulation induced stemness loss in colon cancer [61]. *HOXA7* overexpression enhanced proliferation, migration, invasion and metastasis of liver cancer [62]. *HOXA11* was reported to represent a potential tumor suppressor in different cancers [63, 64]. *HOXC4* overexpression of was observed in lymph node metastases of prostate cancer [65]. Interestingly, different *SOX* genes have already been reported to be involved in oligodendrocyte development. Alterations of corresponding gene expression patterns can therefore be important for tumor development. *SOX6* regulates different stages of oligodendrocyte development by repressing cell specification and terminal differentiation and by influencing cell migration patterns [66]. *SOX8* is expressed in immature glia of the developing cerebellum and in cerebellar tumors [67] and has important functions in oligodendrocyte development and differentiation [68, 69]. *SOX13* regulates the differentiation of specific neurons [70].

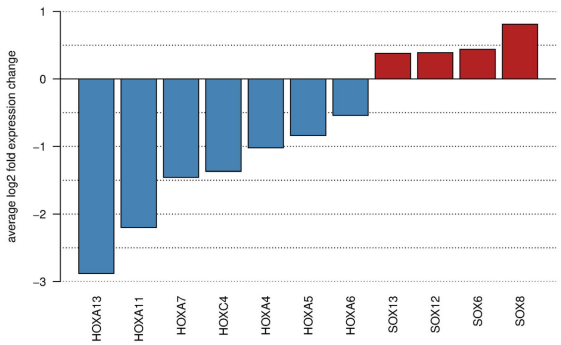

Fig. 7 HOX and SOX signature distinguishing 1p/19q from IDHme. Average log-fold expression differences of *HOX* and *SOX* genes between the 1p/19q and the IDHme subgroup. Genes downregulated in 1p/19q compared to IDHme are shown in blue and upregulated genes are shown in red. Gene expression differences between tumors of both groups were highly significant with q-values clearly less than 0.01 (Additional file 6)

Discussion

First, we analyzed gene copy number data of histologically classified oligodendrogliomas from TCGA and revealed three molecular subgroups by hierarchical clustering of gene copy number data alone (Fig. 1). We used additional information about the presence of a 1p/19q co-deletion [8] and an *IDH* mutation [11] to further characterize these subgroups. In accordance with previous findings for histologically classified oligodendrogliomas [10, 71] and gliomas in general [19], we observed a large 1p/19q subgroup characterized by concurrent 1p/19q co-deletion and *IDH* mutation, an intermediate IDHme subgroup of tumors that mainly show an *IDH* mutation but no commonly overrepresented gene copy number alterations, and a small 7a10d subgroup showing a concurrent duplication of chromosome 7 and a deletion of chromosome 10 where most tumors lacked *IDH* mutations. In addition, considering Verhaak [26] and G-CIMP [15] classes, the 1p/19q and the IDHme subgroup resembled each other, whereas the 7a10d subgroup strongly deviated from these two subgroups also in terms of significantly lower overall patient survival (Fig. 2). This, in combination with the molecular characteristics of the 7a10d subgroup, suggests that these tumors might rather represent glioblastoma-like tumors [3]. This is also supported by a refined molecular classification of gliomas in [20]. Thus, tumors of our small 7a10d subgroup may have been falsely classified as oligodendrogliomas based on histology alone, which is not unlikely considering difficulties of pure histological classifications [6]. We therefore decided to focus our further analyses on the comparison of the 1p/19q and the IDHme subgroups.

Second, we performed an in-depth analysis of the 1p/19q and IDHme subgroups deriving a characteristic gene expression signature that distinguished tumors of both groups (Fig. 3). Interestingly, many of these signature genes were part of signaling pathways involved in the regulation of cell proliferation, differentiation, migration, and cell-cell contacts (Fig. 4). Several of these pathways have already been reported to be involved in glioma development (e.g. PI3K-AKT, MAPK, VEGF signaling) [27, 33, 72, 73]. The strong downregulation of these pathways in the 1p/19q subgroup compared to the IDHme subgroup might be associated with a better sensitivity to treatment and prognosis of (1p/19q) oligodendrogliomas compared to other low-grade gliomas [74, 75].

Third, to better understand differences between the 1p/19q and the IDHme subgroup, we reconstructed a gene regulatory network capable to explain gene expression differences between both subgroups (Figs. 5 and 6). Interestingly, we revealed that several potential hub transcription factors involved in remodeling of the cytoskeleton (e.g. *APBB1IP*, *VAV1*, *ARPC1B*), apoptosis (*CCNL2*, *CREB3L1*), and neural development (e.g. *MYTIL*, *SCRT1*, *MEF2C*) were differentially expressed between both subgroups. Since all or the vast majority of tumors of these two subgroups show *IDH* mutations, the globally observed expression differences are likely to be strongly influenced by the 1p/19q co-deletion. Moreover, we observed characteristic expression differences between *HOX* and *SOX* transcription factors (Fig. 7). All *HOX* genes included in our network were downregulated and all *SOX* genes were upregulated in 1p/19q compared to IDHme. This indicates that the 1p/19q subgroup and the IDHme subgroup express different stemness programs. Recent findings of specific *HOX* and *SOX* gene expression patterns for different types of gliomas indicate an important role of both gene families in brain tumors [20–22]. This is also supported by the recent finding that *SOX2* repression is an early driver of gliomagenesis that blocks the differentiation of neural stem cells in an *in-vitro* model of low-grade astrocytomas [76]. Further experimental studies are required to analyze our revealed stemness signatures.

Finally, it is important to discuss the revealed molecular subtypes in the light of the new WHO 2016 brain tumor classification scheme [18]. All oligodendrogliomas that we analyzed have been classified by the TCGA according to the WHO 2007 brain tumor classification scheme [4], which was state-of-the-art when the tumors were obtained. This older classification is purely based on histology, whereas the new WHO 2016 classification additionally considers the 1p/19q-co-deletion and the *IDH* mutation status. There would be differences in the grouping of tumors, but a reclassification of the analyzed tumors is not straightforward and would require expert knowledge of neuropathologists that have to consider histological and molecular data. Therefore, we cannot realize this reclassification for the considered TCGA data set, but we can interpret our subgroups with respect to the new WHO 2016 classification. Considering our 7a10d subgroup, information about the gain of chromosome 7 and the deletion of chromosome 10 are not considered at all in the new WHO 2016 classification system [18]. Thus, tumors of these subgroup would still not be classified as glioblastomas if no clear signs of high malignancy (necrosis, pathological vascular proliferation) are observed in histology. It is likely that such signs were not present in nearly half of the 7a10d tumors (6 of 15), otherwise these tumors would have been assigned the WHO grade IV instead of grade II according to the WHO 2007 brain tumor classification system. Therefore, these tumors of our 7a10d subgroup might rather be classified as astrocytoma *IDH*-wildtype or *IDH*-mutant (if histological and molecular data are conclusive) or even as oligodendroglioma, NOS (if histological and molecular data are inconclusive) according to the WHO 2016 brain tumor classification system. This may change in future [77]. Such low-grade gliomas without any signs of

high malignancy and without *IDH* mutation still represent an area of ongoing research [78]. Further, like for the WHO 2007 brain tumor classification, all tumors of our 1p/19q subgroup would also be classified as oligodendrogliomas (*IDH*-mutant and 1p/19q-codeleted) according to the WHO 2016 brain tumor classification system. This is also supported by the characteristic overexpression of *SOX* genes. In contrast, tumors of our IDHme subgroup would now be classified as astrocytoma *IDH*-mutant or *IDH*-wildtype also when oligodendroglia-like features are present in histology. This is further supported by the presence of characteristic *ATRX* (30 of 45 tumors) or *TP53* (35 of 45 tumors) mutations in *IDH*-mutated tumors [18]. It is important to note that the new WHO 2016 brain tumor classification system does not change the results of our study. The observed molecular differences between subgroups exist independent of the underlying classification system. Still, one should always be aware of the underlying classification system. In the light of the new WHO 2016 brain tumor classification system, we performed an in-depth comparison of oligodendrogliomas (*IDH*-mutant and 1p/19q co-deleted) represented by our 1p/19q subgroup to astrocytomas (vast majority *IDH*-mutant) represented by our IDHme subgroup. This is supported by our finding that the 1p/19q subgroup expressed an oligodendrocyte-like program and that the IDHme subgroup expressed an astrocyte-like program [22].

Conclusions

Our study confirms prior findings about the molecular subtyping of histologically classified oligodendrogliomas and further provides novel insights into gene expression differences between subtypes. It is important to note that we were able to derive these subtypes purely based on gene copy number data alone. Additional information about the presence of a 1p/19q co-deletion and an *IDH* mutation were only considered subsequently to further characterize these subgroups. The in-depth comparison of the 1p/19q and IDHme subgroups provides novel insights into differences at the level of single genes, pathways, and regulatory networks that have not been reported so far. We identified a characteristic gene expression signature that distinguishes both subgroups including several known signaling pathways that impact on cell proliferation, migration, and angiogenesis. We derived a gene regulatory network that can explain expression differences between both subgroups. Our network-based analysis enabled us to predict novel putative major regulators that contribute to the manifestation of differences between both subgroups. Several of these major regulators are known to be involved in the regulation of cytoskeleton remodeling, apoptosis, and neural development. Moreover, we also revealed a characteristic *HOX* and *SOX* gene expression signature that distinguishes both subgroups suggesting the activity of different glioma stemness programs.

Further, the analyzed oligodendroglioma data set represents an important resource for future research, but researchers have to be aware that these tumors were classified by TCGA according to the WHO 2007 brain tumor classification system. We hope that the discussion of our findings in the context of the new WHO 2016 classification will raise awareness for the fact that brain tumor classification systems can vary considerably. This is important for the interpretation of the results of our retrospective study and for future studies based on the considered TCGA data set.

In summary, our in-depth study focused on the analysis of molecular data of histologically classified oligodendrogliomas. Especially with respect to an oligodendroglial phenotype, characteristic expression differences associated with histological classification may also exist for other types of gliomas. Future studies with already existing molecular data of histologically classified oligodendrogliomas, oligoastrocytomas, and astrocytomas could search for such patterns and evaluate their value for molecular tumor classification.

Additional files

Additional file 1: Oligodendroglioma TCGA identifiers and survival information. (XLSX 14.1 kb)

Additional file 2: Gene copy number and gene expression data of oligodendrogliomas and gene expression data of normal brain references. (XLSX 35942 kb)

Additional file 3: Comparison of revealed subtypes to subtypes revealed for histologically classified oligoastrocytomas and astrocytomas. (PDF 2365 kb)

Additional file 4: Normalized expression values of 5113 signature genes. (TXT 6266 kb)

Additional file 5: 104 signature genes differentially expressed between tumor grades II and III. (XLSX 13.7 kb)

Additional file 6: 1006 transcription factors of the gene signature. (XLSX 92.4 kb)

Additional file 7: Consensus gene regulatory network. Rows of the matrix represent 5113 response variables (signature genes); columns represent 1007 predictors (transcription factors + gene-specific copy number). A cell's value shows how many times the link from the respective predictor gene to the respective response gene was present across the 100 cross-validation iterations. Negative values indicate repressing links and positive values indicate activating links. (TXT 10209 kb)

Additional file 8: **Figure S1.** Comparison to Venteicher et al. (PDF 49.5 kb)

Abbreviations

CNV: Gene copy number variation; G-CIMP: Glioma-CpG island methylator phenotype; IDH: Isocitrate dehydrogenase; OD: Oligodendroglioma; TCGA: The Cancer Genome Atlas; TF: Transcription factor; WHO: World Health Organization; 1p/19q subgroup: Tumors with 1p/19q co-deletion and IDH mutation; IDHme subgroup: Tumors that predominantly show IDH mutations but no 1p/19q co-deletion; 7a10d subgroup: Tumors that show an amplification of chromosome 10 and a deletion of chromosome 7 but mostly no IDH mutation

Acknowledgements
This study would have been impossible without the comprehensive data sets made publicly available by the TCGA Research Network. We thank the reviewers for their valuable comments.

Funding
We did not receive third party funding to perform this study. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the SLUB/TU Dresden to pay the article processing charge.

Availability of data and materials
Data of all considered TCGA tumors are publicly available from the The Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Additional file 1 lists identifiers of all utilized samples and corresponding survival information. Additional file 2 contains all considered gene copy number and gene expression profiles of TCGA tumors and gene expression profiles of the considered normal brain references.

Authors' contributions
MS designed the study. CL and MS performed the analysis and wrote the manuscript. BK contributed to the interpretation of the results. All authors read the manuscript, revised it critically, and approved the final version.

Ethics approval and consent to participate
No ethical approval was required for this study. All utilized public omics data sets were generated by others who obtained ethical approval.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [2]Institute for Clinical Genetics, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [3]National Center for Tumor Diseases, Dresden, Germany.

Received: 21 February 2017 Accepted: 20 March 2018
Published online: 10 April 2018

References

1. Ohgaki H, Kleihues P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. J Neuropathol Exp Neurol. 2005;64(6):479–89.
2. Wesseling P, van den Bent M, Perry A. Oligodendroglioma: pathology, molecular mechanisms and markers. Acta Neuropathologica. 2015;129(6):809–27. https://doi.org/10.1007/s00401-015-1424-1.
3. Ohgaki H, Kleihues P. The definition of primary and secondary glioblastoma. Clin Cancer Res. 2012;19(4):764–72. https://doi.org/10.1158/1078-0432.CCR-12-3002.
4. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathologica. 2007;114(2):97–109. https://doi.org/10.1007/s00401-007-0243-4.
5. Coons S, Johnson P, Scheithauer B, Yates A, Pearl D. Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. Cancer. 1997;79:1381–93.
6. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathologica. 2010;120(3):297–304. https://doi.org/10.1007/s00401-010-0725-7.
7. Riemenschneider MJ, Jeuken JWM, Wesseling P, Reifenberger G. Molecular diagnostics of gliomas: state of the art. Acta Neuropathologica. 2010;120(5):567–84. https://doi.org/10.1007/s00401-010-0736-4.
8. Cairncross J, Ueki K, Zlatescu M, Lisle D, Finkelstein D, Hammond R, et al. Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. J Natal Cancer Institute. 1998;90(19):1473–9.
9. Jansen M, Yip S, Louis DN. Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. Lancet Neurol. 2010;9(7):717–26. https://doi.org/10.1016/S1474-4422(10)70105-8.
10. Kamoun A, Idbaih A, Dehais C, Elarouci N, Carpentier C, Letouzé E, et al. Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. Nat Commun. 2016;7:11263. https://doi.org/10.1038/ncomms11263.
11. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, et al. IDH1 and IDH2 mutations in gliomas. N Eng J Med. 2009;360(8):765–73. https://doi.org/10.1056/NEJMoa0808710.
12. Hartmann C, Meyer J, Balss J, Capper D, Mueller W, Christians A, et al. Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: a study of 1010 diffuse gliomas. Acta Neuropathologica. 2009;118(4):469–74. https://doi.org/10.1007/s00401-009-0561-9.
13. Ichimura K, Pearson DM, Kocialkowski S, Backlund LM, Chan R, Jones DTW, et al. IDH1 mutations are present in the majority of common adult gliomas but rare in primary glioblastomas. Neuro Oncol. 2009;11(4):341–7. https://doi.org/10.1215/15228517-2009-025.
14. Labussiere M, Idbaih A, Wang X-W, Marie Y, Boisselier B, Falet C, et al. All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. Neurology. 2010;74(23):1886–90. https://doi.org/10.1212/WNL.0b013e3181e1cf3a.
15. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17(5):510–22. https://doi.org/10.1016/j.ccr.2010.03.017.
16. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483(7390):479–83. https://doi.org/10.1038/nature10866.
17. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321(5897):1807–12. https://doi.org/10.1126/science.1164382.
18. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World health organization classification of tumors of the central nervous system: a summary. Acta Neuropathologica. 2016;131(6):803–20. https://doi.org/10.1007/s00401-016-1545-1.
19. The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Eng J Med. 2015;372(26):2481–98. https://doi.org/10.1056/NEJMoa1402121.
20. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell. 2016;164(3):550–63. https://doi.org/10.1016/j.cell.2015.12.028.
21. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539:309–13. https://doi.org/10.1038/nature20123.
22. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017;355:8478.
23. Seifert M, Friedrich B, Beyer A. Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. Genome Biol. 2016;17:204. https://doi.org/10.1186/s13059-016-1058-1.
24. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):47. https://doi.org/10.1093/nar/gkv007.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
26. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17(1):98–110. https://doi.org/10.1016/j.ccr.2009.12.020.

27. Seifert M, Garbe M, Friedrich B, Mittelbronn M, Klink B. Comparative transcriptomics reveals similarities and differences between astrocytoma grades. BMC Cancer. 2015;15:952. https://doi.org/10.1186/s12885-015-1939-9.
28. Therneau TM. A Package for Survival Analysis in S. 2015. https://CRAN.R-project.org/package=survival. Version 2.38.
29. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1994;58:267–88.
30. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. 2014;42(2):413–68. https://doi.org/10.1214/13-AOS1175.
31. Seifert M, Beyer A. regNet: An R package for network-based propagation of gene expression alterations. Bioinformatics. 2018;34(2):308–11. https://doi.org/10.1093/bioinformatics/btx544.
32. Aldape K, Burger PC, Perry A. Clinicopathologic aspects of 1p/19q loss and the diagnosis of oligodendroglioma. Arch Pathol Lab Med. 2007;131(2):242–51.
33. Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. Autoregressive higher-order hidden Markov models: Exploiting local chromosomal dependencies in the analysis of tumor expression profiles. PLoS ONE. 2014;9(6):e100295. https://doi.org/10.1371/journal.pone.0100295.
34. Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, et al. GeneCards version 3: the human gene integrator. Database. 2010;2010:020. https://doi.org/10.1093/database/baq020.
35. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods Mol Bio. 2016;1374:23–54.
36. Inagaki T, Suzuki S, Miyamoto T, Takeda T, Yamashita K, Komatsu A, et al. The retinoic acid-responsive proline-rich protein is identified in promyeloleukemic HL-60 Cells. J Biol Chem. 2003;278(51):51685–92. https://doi.org/10.1074/jbc.M308016200.
37. Lafuente EM, van Puijenbroek AAFL, Krause M, Carman CV, Freeman GJ, Berezovskaya A, et al. RIAM, an Ena/VASP and profilin ligand, interacts with Rap1-GTP and mediates Rap1-induced adhesion. Dev Cell. 2004;7(4):585–95. https://doi.org/10.1016/j.devcel.2004.07.021.
38. Razanadrakoto L, Cormier F, Laurienté V, Dondi E, Gardano L, Katzav S, et al. Mutation of Vav1 adaptor region reveals a new oncogenic activation. Oncotarget. 2015;6(4):2524–38. https://doi.org/10.18632/oncotarget.2629.
39. Welch MD, Iwamatsu A, Mitchison TJ. Actin polymerization is induced by Arp 2/3 protein complex at the surface of Listeria monocytogenes. Nature. 1997;385(6613):265–9. https://doi.org/10.1038/385265a0.
40. Hallier M. The transcription factor Spi-1/PU,1 interacts with the potential splicing factor TLS. J Biol Chem. 1998;273(9):4838–42. https://doi.org/10.1074/jbc.273.9.4838.
41. Chung M-C, Kim H-K, Kawamoto S. TFEC can function as a transcriptional activator of the nonmuscle myosin II heavy chain-a gene in transfected cells. Biochemistry. 2001;40(30):8887–97. https://doi.org/10.1021/bi002847d.
42. Suratannon N, Yeetong P, Srichomthong C, Amarinthnukrowh P, Chatchatee P, Sosothikul D, et al. Adaptive immune defects in a patient with leukocyte adhesion deficiency type III with a novel mutation in FERMT3. Pediatr Allergy Immunol. 2015;27(2):214–7. https://doi.org/10.1111/pai.12485.
43. Yang L, Luo Y, Wei J. Integrative genomic analyses on Ikaros and its expression related to solid cancer prognosis. Oncol. Rep. 2010;24(2):571–7.
44. Tang D, Yeung J, Lee K.-Y, Matsushita M, Matsui H, Tomizawa K, et al. An isoform of the neuronal cyclin-dependent kinase 5 (Cdk5) activator. J Biol Chem. 1995;270(45):26897–903. https://doi.org/10.1074/jbc.270.45.26897.
45. Ladd AN, Charlet-B N, Cooper TA. The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cellular Biol. 2001;21(4):1285–96. https://doi.org/10.1128/MCB.21.4.1285-1296.2001.
46. Koelle MR, Horvitz HR. EGL-10 regulates G protein signaling in the C, elegans nervous system and shares a conserved domain with many mammalian proteins. Cell. 1996;84(1):115–25. https://doi.org/10.1016/S0092-8674(00)80998-8.
47. Nakakura EK, Watkins DN, Schuebel KE, Sriuranpong V, Borges MW, Nelkin BD, et al. Mammalian scratch: a neural-specific snail family transcriptional repressor. Proc Natl Acad Sci. 2001;98(7):4010–5. https://doi.org/10.1073/pnas.051014098.
48. Mishra S, Murphy LC, Murphy LJ. The prohibitins: emerging roles in diverse functions. J Cellular Mol Med. 2006;10(2):353–63. https://doi.org/10.1111/j.1582-4934.2006.tb00404.x.
49. Sato T, Saito H, Swensen J, Olifant A, Wood C, Danner D, et al. The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer. Cancer Res. 1992;52(6):1643–6.
50. Kondo S, Hino S.-I, Saito A, Kanemoto S, Kawasaki N, Asada R, et al. Activation of OASIS family, ER stress transducers, is dependent on its stabilization. Cell Death Differ. 2012;19(12):1939–49. https://doi.org/10.1038/cdd.2012.77.
51. Prendergast L, van Vuuren C, Kaczmarczyk A, Doering V, Hellwig D, Quinn N, et al. Premitotic assembly of human CENPs -t and -w switches centromeric chromatin to a mitotic state. PLoS Biol. 2011;9(6):1001082. https://doi.org/10.1371/journal.pbio.1001082.
52. Leifer D, Krainc D, Yu YT, McDermott J, Breitbart RE, Heng J, et al. MEF2C, a MADS/MEF2-family transcription factor expressed in a laminar distribution in cerebral cortex. Proc Natl Acad Sci. 1993;90(4):1546–50. https://doi.org/10.1073/pnas.90.4.1546.
53. McDermott JC, Cardoso MC, Yu YT, Andres V, Leifer D, Krainc D, et al. hMEF2C gene encodes skeletal muscle- and brain-specific transcription factors. Mol Cellular Biol. 1993;13(4):2564–77. https://doi.org/10.1128/mcb.13.4.2564.
54. Shen X, Yang Y, Liu W, Sun M, Jiang J, et al. Identification of the p28 subunit of eukaryotic initiation factor 3(eIF3k) as a new interaction partner of cyclin D3. FEBS Lett. 2004;573(1-3):139–46. https://doi.org/10.1016/j.febslet.2004.07.071.
55. Yang L. Cyclin I2, a novel RNA polymerase II-associated cyclin, is involved in pre-mRNA splicing and induces apoptosis of human hepatocellular carcinoma cells. J Biol Chem. 2004;279(12):11639–48. https://doi.org/10.1074/jbc.M312895200.
56. Kurscheid S, Baddy P, Sciuscio D, Samarzija I, Shay T, Vassallo I, et al. Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. Genome Biol. 2015;16:16. https://doi.org/10.1186/s13059-015-0583-7.
57. Ferletta M. The Role of Sox Transcription Factors in Brain Tumourigenesis, Molecular Targets of CNS Tumors. London: InTech; 2011. https://www.intechopen.com/contact.html.
58. Chew LJ, Gallo V. The Yin and Yang of Sox proteins: Activation and repression in development and disease. J Neurosci Res. 2009;87(15):3277–87. https://doi.org/10.1002/jnr.22128.
59. Klausen C, Leung PC, Auersperg N. Cell motility and spreading are suppressed by HOXA4 in ovarian cancer cells: possible involvement of beta1 integrin. Mol Cancer Res. 2009;7:1425–37.
60. Teo WW, Merino VF, Cho S, Korangath P, Liang X, Wu R-C, et al. HOXA5 determines cell fate transition and impedes tumor initiation and progression in breast cancer through regulation of E-cadherin and CD24. Oncogene. 2016;35:5539–51.
61. Ordóñez-Morán P, Dafflon C, Imajo M, Nishida E, Huelsken J. HOXA5 counteracts stem cell traits by inhibiting Wnt signaling in colorectal cancer. Cancer Cell. 2015;28:815–29.
62. Tang B, Qi G, Sun X, Tang F, Yuan S, Wang Z, et al. HOXA7 plays a critical role in metastasis of liver cancer associated with activation of Snail. Mol Cancer. 2016;15:57. https://doi.org/10.1186/s12943-016-0540-4.
63. Bai Y, Fang N, Gu T, Kang Y, Wu J, Yang D, et al. HOXA11 gene is hypermethylation and aberrant expression in gastric cancer. Cancer Cell Int. 2014;14:79. https://doi.org/10.1186/s12935-014-0079-7.
64. Wang L, Cui Y, Sheng J, Yang Y, Kuang G, Fan Y, et al. Epigenetic inactivation of HOXA11, a novel functional tumor suppressor for renal cell carcinoma, is associated with RCC TNM classification. Oncotarget. 2017;8:21861–70.
65. Miller GJ, Miller HL, van Bokhoven A, Lambert JR, Werahera PN, Schirripa O, et al. Aberrant HOXC expression accompanies the malignant phenotype in human prostate. Cancer Res. 2003;63:5879–88.
66. Stolt CC, Schlierf A, Lommes P, Hillgärtner S, Werner T, Kosian T, et al. SoxD proteins influence multiple stages of oligodendrocyte development and modulate SoxE protein function. Dev Cell. 2006;11:697–709.
67. Cheng Y-C, Lee C-J, Badge RM, Orme AT, Scotting PJ. Sox8 gene expression identifies immature glial cells in developing cerebellum and cerebellar tumours. Mol Brain Res. 2001;92:193–200.

68. Stolt CC, Lommes P, Friedrich RP, Wegner M. Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy. Development. 2004;131:2349–58.
69. Stolt C. C, Schmitt S, Lommes P, Sock E, Wegner M. Impact of transcription factor Sox8 on oligodendrocyte specification in the mouse embryonic spinal cord. Dev Biol. 2005;281(2):309–17. https://doi.org/10.1016/j.ydbio.2005.03.010.
70. Wang Y, Bagheri-Fam S, Harley VR. SOX13 is up-regulated in the developing mouse neuroepithelium and identifies a sub-population of differentiating neurons. Dev Brain Res. 2005;157:201–8.
71. Eisenreich S, Abou-El-Ardat K, Szafranski K, Campos Valenzuela JA, Rump A, Nigro JM, et al. Novel ClC point mutations and an exon-spanning, homozygous deletion identified in oligodendroglial tumors by a comprehensive genomic approach including transcriptome sequencing. PLoS ONE. 2013;8(9):e76623. https://doi.org/10.1371/journal.pone.0076623.
72. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.
73. Pearson JRD, Regad T. Targeting cellular pathways in glioblastoma multiforme. Signal Transduct Targeted Ther. 2017;2:17040. https://doi.org/10.1038/sigtrans.2017.40.
74. Sakata K, Hareyama M, Komae T, Shirato H, Watanabe O, Watarai J, et al. Supratentorial astrocytomas and oligodendrogliomas treated in the MRI era. Jpn J Clin Oncol. 2001;31:240–5.
75. Koeller KK, Rushing EJ. Oligodendroglioma and its variants: Radiologic-pathologic correlation. RadioGraphics. 2005;25(6):1669–88. https://doi.org/10.1148/rg.256055137.
76. Modrek AS, Golub D, Khan T, Bready D, Prado J, Bowman C, et al. Low-grade astrocytoma mutations in IDH1, P53, and ATRX cooperate to block differentiation of human neural stem cells via repression of SOX2. Cell Rep. 2017;21(5):1267–80.
77. Reifenberger G, Wirsching HG, Knobbe-Thomsen CB, Weller M. Advances in the molecular genetics of gliomas - implications for classification and therapy. Nat Rev Clin Oncol. 2017;14(7):434–52. https://doi.org/10.1038/nrclinonc.2016.204.
78. van der Vlis TAMB, Hoeben A, Beckervordersandforth JC, Ackermans L, Eekers DBP, Wennekes RMJ, et al. Impact of the revised WHO classification of diffuse low-grade glioma on clinical decision making: A case report. Surg Neurol Int. 2017;8:223.

4.3 Publication:

Survival differences and associated molecular signatures of DNMT3A-mutant acute myeloid leukemia patients

Journal: Scientific Reports

Received: 27 February 2019; **Accepted**: 13 July 2020; **Published**: 29 July 2020

Citation: Chris Lauber , Nádia Correia, Andreas Trumpp, Michael A. Rieger, Anna Dolnik, Lars Bullinger, Ingo Roeder and Michael Seifert (2020): Survival differences and associated molecular signatures of DNMT3A-mutant acute myeloid leukemia patients. Scientific Reports, 10:12761.

Placement and summary of the publication

The work on this study was motivated by my involvement in the SyTASC (Systems-based Therapy of AML Stem Cells) project, which was funded by the German Cancer Aid with the goal to identify molecular factors that can explain survival differences of acute myeloid leukemia (AML) patients with a mutation of DNMT3A.

AML is a highly malignant and very heterogeneous cancer affecting myeloid blood cells. AML is characterized by a rapid growth of abnormal immature myeloblasts that lost their ability to differentiate leading to the replacement of normal cells in bone marrow and blood (Döhner et al. (2015)). One of the most frequently mutated genes in AML is the DNA methyltransferase DNMT3A (The Cancer Genome Atlas Research Network (2013a)), which is known to be important for normal hematopoiesis (Challen et al. (2011); Yang et al. (2015)). DNMT3A-mutations have been associated with shorter survival of AML patients in different studies (e.g. Ribeiro et al. (2012); Renneville et al. (2012)). Still, some DNMT3A-mutant patients have also been

reported to show long survival or even a long-term remission (Ploen et al. (2014); Sun et al. (2016)). Molecular factors that explain these survival differences were not known so far and molecular differences distinguishing short- and long-lived DNMT3A-mutant patients had not been intensively studied, but such knowledge would be very important to improve patient stratification. We therefore decided to search for such factors in publicly available omics profiles of DNMT3A-mutant AML patients.

We initially analyzed genome-wide somatic mutation profiles of DNMT3A-mutant patients from TCGA (The Cancer Genome Atlas Research Network (2013a)) by hierarchical clustering revealing two patient subgroups with strong differences in survival. We further determined molecular mutation and expression signatures that distinguished both subgroups. We found that the presence of FLT3 and/or NPM1 mutations, two known genes frequently mutated in AML and associated with poor prognosis, contribute to the observed survival differences of DNMT3A-mutant patients. We also observed an upregulation of genes of the p53, VEGF and DNA replication pathway and a downregulation of genes of the PI3K-Akt pathway in short- compared to long-lived patients. We further identified that the majority of microRNAs was downregulated in the short-lived group compared to the long-lived group and that some of these microRNAs have not been linked to AML so far (miR-153-2, miR-3065, miR-95, miR-6718). We learned gene regulatory networks to identify potential major regulators that distinguished both subgroups revealing several genes and microRNAs with known roles in AML pathogenesis, but also novel candidates involved in the regulation of hematopoiesis, cell cycle, cell differentiation and immunity. Moreover, the characteristic gene mutation and expression signatures that distinguished short- from long-lived patients were also predictive for independent DNMT3A-mutant AML patients from other cohorts and could also contribute to further improve existing prognostic scoring systems.

Our study represents the first in-depth computational approach that characterizes molecular factors associated with survival differences of DNMT3A-mutant AML patients. This could contribute to the development of robust markers for an improved patient stratification.

Author contribution

I designed the study. I supervised the computational analyses by Chris Lauber, who was a postdoc in my group. We both wrote the initial version of the manuscript together. I performed all computational studies for the four revisions of the manuscript and revised the manuscript. Anna Dolnik and Lars Bullinger provided the data of the independent Ulm validation cohort. Nádia Correia, Andreas Trumpp, Michael A. Rieger and Ingo Roeder supported the discussion of the results.

www.nature.com/scientificreports

SCIENTIFIC REPORTS
nature research

OPEN

Survival differences and associated molecular signatures of *DNMT3A*-mutant acute myeloid leukemia patients

Chris Lauber[1], Nádia Correia[2], Andreas Trumpp[2], Michael A. Rieger[3], Anna Dolnik[4], Lars Bullinger[4], Ingo Roeder[1,5] & Michael Seifert[1,5]✉

Acute myeloid leukemia (AML) is a very heterogeneous and highly malignant blood cancer. Mutations of the DNA methyltransferase *DNMT3A* are among the most frequent recurrent genetic lesions in AML. The majority of *DNMT3A*-mutant AML patients shows fast relapse and poor survival, but also patients with long survival or long-term remission have been reported. Underlying molecular signatures and mechanisms that contribute to these survival differences are only poorly understood and have not been studied in detail so far. We applied hierarchical clustering to somatic gene mutation profiles of 51 *DNMT3A*-mutant patients from The Cancer Genome Atlas (TCGA) AML cohort revealing two robust patient subgroups with profound differences in survival. We further determined molecular signatures that distinguish both subgroups. Our results suggest that *FLT3* and/or *NPM1* mutations contribute to survival differences of *DNMT3A*-mutant patients. We observed an upregulation of genes of the p53, VEGF and DNA replication pathway and a downregulation of genes of the PI3K-Akt pathway in short- compared to long-lived patients. We identified that the majority of measured miRNAs was downregulated in the short-lived group and we found differentially expressed microRNAs between both subgroups that have not been reported for AML so far (*miR-153-2*, *miR-3065*, *miR-95*, *miR-6718*) suggesting that miRNAs could be important for prognosis. In addition, we learned gene regulatory networks to predict potential major regulators and found several genes and miRNAs with known roles in AML pathogenesis, but also interesting novel candidates involved in the regulation of hematopoiesis, cell cycle, cell differentiation, and immunity that may contribute to the observed survival differences of both subgroups and could therefore be important for prognosis. Moreover, the characteristic gene mutation and expression signatures that distinguished short- from long-lived patients were also predictive for independent *DNMT3A*-mutant AML patients from other cohorts and could also contribute to further improve the European LeukemiaNet (ELN) prognostic scoring system. Our study represents the first in-depth computational approach to identify molecular factors associated with survival differences of *DNMT3A*-mutant AML patients and could trigger additional studies to develop robust molecular markers for a better stratification of AML patients with *DNMT3A* mutations.

Acute myeloid leukemia (AML) is a highly malignant cancer of myeloid blood cells affecting about one million people globally in 2015[1,2]. It most frequently occurs in older adults and shows a relatively poor five-year survival rate of about 25%, which is worsening with increasing age of a patient at diagnosis[3]. AML is characterized by a rapid growth of abnormal, immature myeloblasts that lost their ability to differentiate, which replace normal

[1]Institute for Medical Informatics and Biometry (IMB), Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [2]Division of Stem Cells and Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany. [3]Department of Medicine, Hematology/Oncology, Goethe University Hospital Frankfurt, Frankfurt, Germany. [4]Department of Hematology, Oncology and Tumorimmunology, Charité University Medicine Berlin, Campus Virchow Klinikum Berlin, Germany. [5]National Center for Tumor Diseases (NCT), Dresden, Germany. ✉email: michael.seifert@tu-dresden.de

cells in the bone marrow and blood. At the level of underlying genetic aberrations, AML is very heterogeneous. Mutations in several genes are required for leukemic transformation affecting multiple steps of the differentiation pathway[4,5]. In addition, different cytogenetic abnormalities of significant prognostic relevance, ranging from translocations (t(8;21), t(15;17)) and inversions (inv(16)) with relatively good prognosis to deletions of whole chromosomes (5, 7) or chromosomal arms (5q) and abnormalities on the q-arm of chromosome 3 (3q) associated with high risk, have been observed in AML patients[6–8].

The first genome of a cytogenetically normal AML patient was sequenced in 2008[9]. The Cancer Genome Atlas (TCGA) Research Network made enormous efforts to perform whole-genome or exome sequencing, transcriptome and microRNA (miRNA) sequencing, and DNA methylome analysis of a large cohort of adult AML cases in 2013[10]. These and other sequencing-based studies (e.g.[11–14]) enabled the identification of several genetic and genomic alterations acquired during AML pathogenesis. Subtypes of AML are associated with distinctive patterns of altered gene expression (e.g.[15–17]). Likewise, a prognostic and functional role of widespread dysregulation of miRNAs has emerged[18,19]. Regarding somatic mutations, it was found that only about a dozen genes are affected on average in an AML patient, which is considerably less than in most other human cancers[10]. The by far top-ranking recurrently mutated genes in AML are *FLT3*, *NPM1* and *DNMT3A*[10].

The DNA methlytransferase 3A (*DNMT3A*) forms a gene family of DNA methyltransferases together with *DNMT3B* and *DNMT1*, where the encoded proteins DNMT3A and DNMT3B add methyl groups to unmodified DNA by conversion of cytosine to 5-methylcytosine, while DNMT1 maintains existing DNA methylation after cell division[20]. *DNMT3A* is highly expressed in embryonic stem cells[21,22]. A *DNMT3A* deletion in mouse hematopoietic stem cells has been shown to inhibit differentiation[23] and a deletion of *DNMT3A* in human hematopoietic stem cells resulted in increased self-renewal and blockage of differentiation[24]. This importance of *DNMT3A* for normal hematopoiesis is in line with its high frequency of somatic mutations in AML, which are found in about 20% of patients[9,25]. It is assumed that *DNMT3A* mutations are acquired months or years before a potential onset of AML from hematopoietic stem cells or multipotent precursor cells leading to a pre-leukemic state that potentially leads to the development of AML[26,27]. In addition, significant associations of *DNMT3A* mutations with IDH1/2 mutations, *FLT3* internal tandem duplications (ITD) and tyrosine kinase domain mutations (TKD), and *NPM1* mutations have been reported[9,28].

Notably, around two-thirds of the *DNMT3A* mutations affect the R882 codon in the methyltransferase domain of *DNMT3A*[9,25]. Moreover, *DNMT3A* mutations in general or those affecting the R882 residue have been linked to shorter survival rates of patients[9,14,25,29–31], but there is also an ongoing debate about the prognostic values of R882 and non-R882 *DNMT3A* mutations. This debate is fueled by the fact that, in contrast to generally poor prognosis, some *DNMT3A*-mutant patients show relatively long survival or even go into long-term remission with *DNMT3A* mutations remaining stable[32,33]. Molecular characteristics associated with such prognosis differences of *DNMT3A*-mutant patients have not been intensively studied so far.

Here, we initially analyzed genome-wide somatic mutation profiles of *DNMT3A*-mutant patients from the TCGA AML cohort by hierarchical clustering. Our analysis revealed two patient subgroups with profound differences in overall survival rates. Additional analyses of gene and miRNA expression data in combination with inference of gene regulatory networks enabled us to identify molecular patterns of expression dysregulation as well as gene modules that distinguish both subgroups. The characteristic gene mutation and expression signatures also enabled to separate *DNMT3A*-mutant AML patients of two independent cohorts into a short- and long-lived group. The results of our computational analysis point toward several genetic regulators and cellular processes that are potentially involved in a manifestation of apparent survival differences of AML patients with *DNMT3A* mutations.

Results

Two subgroups of *DNMT3A*-mutated AML patients differ in overall survival. Considering the gene mutation data from TCGA for all 197 AML patients, we found that 51 of them had a *DNMT3A* mutation. We observed in total 5 frame-shift, 43 missense, 6 nonsense and 3 splice site *DNMT3A* mutations including 6 patients that had two of these mutations. For 29 (57%) of the patients, the mutation affected the R882 codon at second (n=22) or first (n=7) codon position (Supplementary Table 1). The 51 *DNMT3A*-mutated patients had on average 13.3 mutated genes (min=2, max=24) from 1,890 genes analyzed in total. Hierarchical clustering of the 51 *DNMT3A*-mutated patients based on binary mutational profiles of the 1,890 genes revealed two well-separated subgroups of nearly equal size (24 vs. 27 patients; Fig. 1A). Importantly, the two subgroups of *DNMT3A*-mutant patients showed a significant difference in overall survival ($P < 0.013$; Fig. 1B, Supplementary Table 2). Compared to 138 AML patients without a *DNMT3A* mutation, only the subgroup with shorter overall survival (short-lived subgroup from here on) showed a statistically significant difference in survival ($P < 0.0001$), while the other (long-lived subgroup) did not ($P < 0.345$), although a considerable deviation of its survival curve from that of the non-mutated patients was observed (Fig. 1B). Generally, *DNMT3A*-mutant patients showed significantly shorter survival than patients without a *DNMT3A* mutation (Fig. 1B, $P = 0.004$). Further, the short-lived subgroup was enriched with patients harboring a R882 *DNMT3A* mutation (n=17, 71%) compared to patients with non-R882 mutations (n=7, 29%), while the long-lived subgroup was composed of 12 patients with R882 (44%) and 15 patients with non-R882 mutations (56%). However, this difference in the proportion of R882 mutations of both subgroups was not significant (Chi-squared test, $P = 0.106$). We further compared the number of mutated genes and cytogenetic abnormalities between the short- and long-lived subgroup. The median number of mutated genes of short-lived patients was significantly smaller than for long-lived patients (Supplementary Fig. 5; U-Test: $P < 0.004$; short-lived: 10.5; long-lived: 17). The majority of short- (71%) and long-lived patients (59%) had normal cytogenetic profiles. Interestingly, the long-lived group contained 7 patients (26%) with duplications or rearrangements of chromosome 8 that have not been observed in the short-lived group.

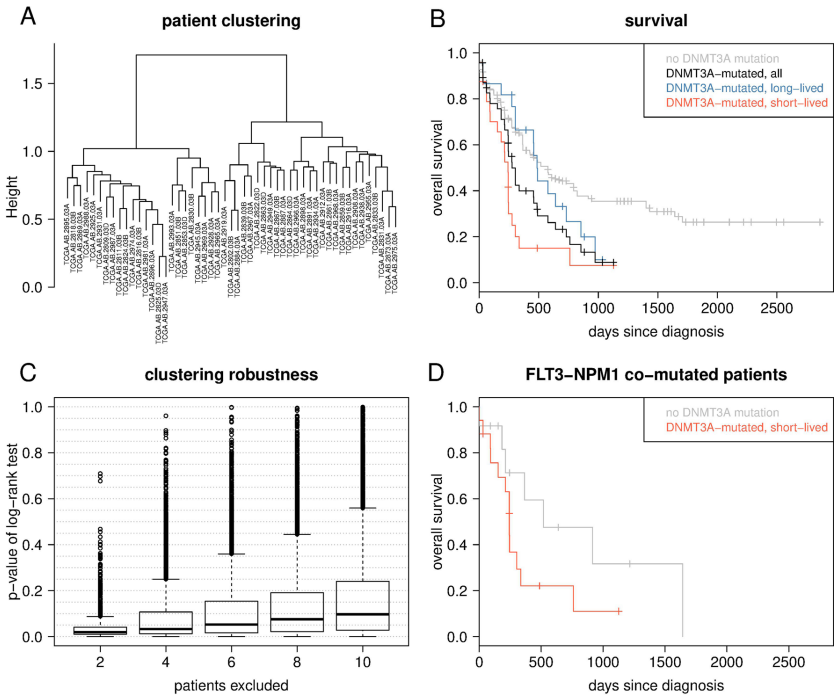

Figure 1. Clustering of *DNMT3A*-mutated AML patients into two subgroups that differ in survival. (**A**) Hierarchical clustering of 51 *DNMT3A*-mutated AML patients; tip labels indicate TCGA identifiers (left subtree: short-lived, right subtree: long-lived). (**B**) Kaplan-Meier survival curves for the patients from (**A**) (black) and the two subgroups (short-lived: red, left subtree in A, survival data available for all 24 patients; long-lived: blue, right subtree in A, survival data available for 23 of 27 patients) as well as for 138 AML patients without a *DNMT3A* mutation (gray). Log-rank tests: $P < 0.013$ for red vs. blue, $P < 0.0001$ for red vs. grey, $P = 0.345$ for blue vs. grey, $P = 0.004$ for black vs. grey. (**C**) Robustness of clustering the *DNMT3A*-mutated patients into two subgroups that differ in survival, as assessed by randomly excluding patients and performing a hierarchical clustering and subsequent log-rank test on the data subset. Each boxplot shows the distributions of p-values of the log-rank tests for 10,000 data subsets. (**D**) Kaplan-Meier survival curves analyzing the impact of *FLT3* and *NPM1* co-mutations for all 17 affected patients of the short-lived subgroup (red) and all 12 affected patients of the 138 patients without a *DNMT3A*-mutation (grey). Log-rank test: $P < 0.094$.

To evaluate the robustness of the grouping of the 51 *DNMT3A*-mutated patients into two subgroups that differ in survival, we repeated the clustering for data subsets obtained by excluding different randomly selected fractions of patients considering 10,000 repetitions of this procedure (see Methods for details). For the vast number of subsets, the difference in patient survival between the subgroups remained significant or stayed close to the level of significance obtained for the full data set, although p-values of the log-rank tests increased with increasing number of excluded patients (Fig. 1C). The latter is not unexpected considering the limited number of *DNMT3A*-mutated AML patients.

For the analysis in which two patients were excluded at random, we observed that few subsets showed exceptionally high p-values of the corresponding log-rank tests. The patients excluded from these subsets exclusively belonged to a set of in total 4 members of the short-lived subgroup (TCGA case identifiers: TCGA-AB-2931-03, TCGA-AB-2824-03, TCGA-AB-2896-03, TCGA-AB-2945-03). Each of these four patients died, and their survival times were, respectively, 0, 30, 214, and 243 days after diagnosis. The four patients showed mutations in 13, 5, 3 and 13 genes, respectively; all four had an *NPM1* and two of them an *FLT3* mutation, while the remaining mutations were found only once among the four patients.

Frequent *FLT3* and *NPM1* mutations distinguish short- and long-lived *DNMT3A*-mutated patients. In order to understand whether and how patients from the two identified subgroups differ at the molecular level, we first searched for somatic mutations of specific genes that were enriched in one subgroup

compared to the other. We found that each patient of the short-lived subgroup had at least one of either *FLT3* (20 of 24 patients) or *NPM1* (21 of 24 patients) mutated, with 17 (71%) of them showing mutations in both of these genes. In sharp contrast, *FLT3* and *NPM1* were mutated in only one and seven patients of the 27 patients from the long-lived subgroup, respectively. We did not find any gene with strong enrichment of mutations in patients from the long-lived subgroup. Instead, we only observed slightly increased numbers of five *IDH2* and four *MT-CYB* mutations in this subgroup. These two genes were not mutated in any of the patients from the short-lived subgroup.

To test whether or not the short survival of patients from the short-lived subgroup is mainly driven by *FLT3-NPM1* co-mutations, we separately analyzed a subset of in total 29 AML patients from the TCGA AML cohort, which had these two genes mutated. Seventeen of them also had a *DNMT3A* mutation and showed a considerably shorter survival compared to the remaining 12 patients without a *DNMT3A* mutation (Fig. 1D; Log-rank test, $P < 0.094$). Although not statistically significant but considering the small sample size, this points towards an effect of *DNMT3A* mutations on survival that is independent of *FLT3* and *NPM1* co-mutations.

Also patients either having a *FLT3* mutation or a *NPM1* mutation in combination with a *DNMT3A* mutation showed shorter overall survival than patients without a *DNMT3A* mutation (Supplementary Fig. 2). Further, the overall survival of patients with *NPM1-DNMT3A* co-mutations was very similar to those of patients with *FLT3-NPM1-DNMT3A* co-mutations. Co-mutations of *DNMT3A* with *FLT3*, *NPM1* or both genes were generally associated with poor survival.

In addition, we determined the specific type of *FLT3* mutation for each patient and analyzed if *FLT3*-ITD and *FLT3*-TKD differ in their impact on survival of *DNMT3A*-mutant AML patients from TCGA (Supplementary Table 1, Supplementary Fig. 6). The 20 *FLT3* mutations in the short-lived subgroup were split into 11 *FLT3*-ITD and 9 *FLT3*-TKD mutations. The one *FLT3* mutation in the long-lived group was a *FLT3*-ITD mutation. There was no significant difference in survival of *DNMT3A*-mutant AML patients distinguished by their type of *FLT3* mutation. Both groups did also not significantly differ in survival in comparison to *DNMT3A*-mutant AML patients without *FLT3* mutations.

We further analyzed the gene mutation profiles within the short- and long-lived group by additionally dividing each corresponding subtree in Fig. 1A into its two major patient subgroups (Supplementary Fig. 3). Both derived short-lived subgroups strongly differed in the number of co-mutations of *DNMT3A* with *FLT3* or *NPM1*. The two derived long-lived subgroups strongly differed in the number of co-mutations of *DNMT3A* with *IDH1* or *IDH2* and also in the number of *NPM1* mutations.

Since *DNMT3A*-R882 mutations were increased in the short-lived group, we analyzed if mutations of *FLT3* or/and *NPM1* are found more frequently in AML patients with *DNMT3A*-R882 mutations compared to patients with other *DNMT3A* mutations. We therefore considered a large independent cohort of AML patients[34] and found a significant enrichment of *DNMT3A*-R882 and *NPM1* co-mutations and a significant enrichment of concurrent *DNMT3A*-R882, *NPM1*, *FLT3* mutations compared to the corresponding groups of patients with other *DNMT3A* mutations (Supplementary Fig. 4, Fisher's exact test: $P < 0.01$), whereas no significant difference in the proportion of *FLT3* mutations was found. We also observed systematic differences considering the percentage of peripheral blood blasts, white blood cell counts, platelet counts, and the hemoglobin level indicating that differentiation capabilities of AML cells with R882 and non-R882 *DNMT3A* mutations may differ at least to some extent (Supplementary Fig. 4).

A gene expression signature discriminates short- and long-lived *DNMT3A*-mutated patients. Next, we used RNA-Seq gene expression data from TCGA for the 51 *DNMT3A*-mutated patients and conducted a differential gene expression analysis to search for genes that differ in their expression levels between the short- and long-lived subgroup. We identified 260 differentially expressed genes (DEGs) using an FDR-corrected p-value (q-value) cut-off of 0.1 (Fig. 2A, Supplementary Table 3).

When grouping the 260 DEGs into different functional categories (transcription factors, oncogenes, tumor suppressor genes, kinases, phosphatases, signaling and metabolic pathway genes, etc.[35]), we only found a significant enrichment for known cancer-relevant signaling pathway genes. This included four genes involved in cytokine receptor interactions (*CCL23, FAS, KITLG, TSLP*) that were upregulated in the short-lived relative to the long-lived subgroup, two genes of the p53 signaling pathway (*FAS, TP53I3*) that were also upregulated, six genes involved in PI3K-Akt signaling (*EFNA1, FGF9, GNG11, GNG2, GNG7, ITGA6*) that were downregulated, two genes of the VEGF signaling pathway (*PIK3CB, PLA2G4A*) that were upregulated, and two genes involved in DNA replication (*POLE4, RNASEH2C*) that were also found to be upregulated in the short-lived subgroup (Fig. 2B).

Since we compared expression profiles of two relatively large groups, individual genes can also vary in their expression within a group while still being differentially expressed between both groups. This can result in additional subgroups that are masked by the global differential expression analysis. We therefore performed a hierarchical clustering of the patients based on expression profiles of the 260 DEGs, which resulted in four expression groups (EGs, Supplementary Table 2) of patients with characteristic large-scale expression differences for the 260 genes (Fig. 2C). EG1 exclusively contained 15 patients from the long-lived subgroup, while EG3 included 16 short-lived and a single long-lived patient (Fig. 2C, Table 1). These two groups with evident differences in gene expression thus strongly resemble the long- and short-lived subgroup clustering based on the somatic mutation data. The other two groups of patients (EG2a and EG2b) represented a mixture of in total five patients from the short-lived and ten patients from the long-lived subgroup with intermediate expression levels for most of the 260 DEGs (Fig. 2C, Table 1).

When inspecting additional meta-information from TCGA for the patients of the different expression groups, we observed no systematic differences regarding cytogenetic abnormality types. Instead, there was a notable

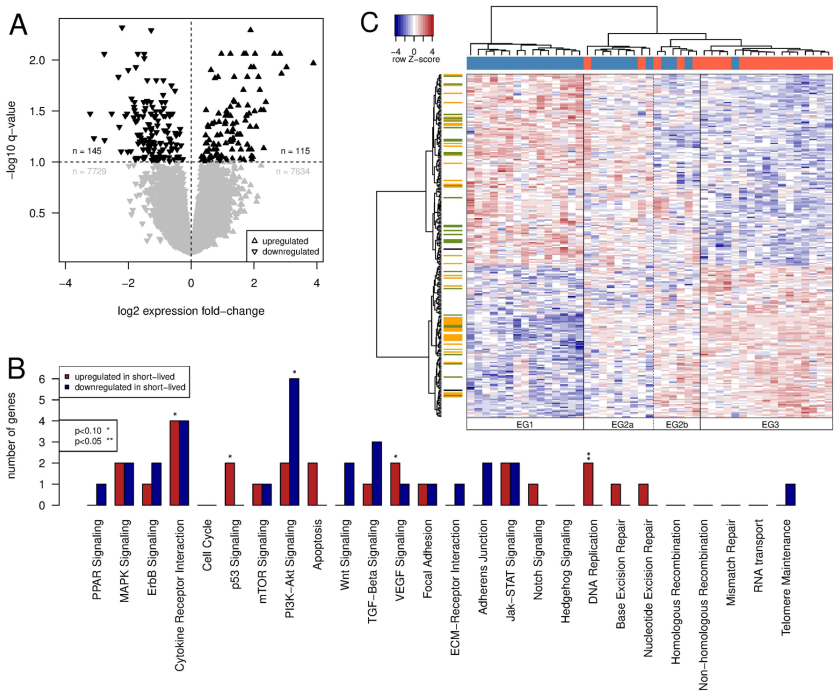

Figure 2. Genes differentially expressed between patient subgroups and enrichment analysis. (**A**) Volcano plot showing the relative expression change of the 15,623 genes between patients from the short-lived and long-lived subgroup. Genes with a significant change in expression ($q < 0.1$) are in black, others in gray. (**B**) Signaling pathways enriched with genes that are differentially expressed between the short- and the long-lived subgroup; separately shown for genes upregulated (red) and downregulated (blue), respectively, in the short-lived relative to the long-lived subgroup. (**C**) Gene expression heatmap of 260 differentially expressed genes. Rows are Z score-scaled. Column coloring indicates patients from the short-lived (red) and the long-lived (blue) subgroup. Row coloring highlights known transcription factors (yellow), genes involved in signaling pathways (green) and genes showing both of these annotations (black).

tendency that patients of EG2b and EG3, the two groups with a high or very high fraction of short-lived patients, were more frequently classified to have FAB type M4 (acute myelomocytic leukemia) or M5 (acute monoblastic leukemia or acute monocytic leukemia) (Table 1). The FAB types M4 and M5 have previously been associated with a high mutational burden at diagnosis[36]. This was not confirmed for our cohort, where the median number of mutated genes for patients within EG2b and EG3 was significantly smaller than for patients within EG1 and EG2a (U-Test: $P < 0.002$; EG2b and EG3: 11; EG1 and EG2a: 17; Supplementary Fig. 1).

A miRNA expression signature discriminates short- and long-lived *DNMT3A*-mutated patients. To further analyze differences between the short- and the long-lived subgroup with respect to gene regulation, we considered miRNA expression data from TCGA available for 42 of the 51 *DNMT3A*-mutated AML patients. As for the gene expression data, we conducted a differential expression analysis and identified 25 differentially expressed miRNAs discriminating patients from the two subgroups using a q-value cut-off of 0.1 (Fig. 3A, Supplementary Table 3).

Interestingly, the relative fractions of up- and downregulated miRNAs in the short-lived compared to the long-lived subgroup were highly uneven. The large majority of miRNAs (21 out of 25) were downregulated in the short-lived subgroup, while only four miRNAs were upregulated. An altered miRNA expression can have different reasons: (i) it could be caused by the altered expression of a host gene that contains the affected miRNA, or (ii) the expression of a miRNA can be altered directly and independent of its host gene or in the absence of a host gene (e.g. a miRNA encoded in an intergenic chromosomal region). Therefore, we tested whether or not the expression of a miRNA is significantly correlated with the expression of its host gene across all *DNMT3A*-mutant patients.

	Expression group			
	EG1	EG2a	EG2b	EG3
Group composition				
Short-lived patients	0	2	3	16
Long-lived patients	15	7	3	1
Cytogenetic abnormality types				
n.a.	1	1	0	2
8+	3	1	0	0
7q-	1	1	0	0
Complex	1	0	0	1
Complex 5p-	1	0	0	0
Normal	7	6	4	14
Normal 8+	1	0	1	0
Normal 7q-	0	0	1	0
FAB types				
n.a.	0	1	0	0
M0	2	0	0	0
M1	5	3	0	3
M2	4	3	0	3
M3	0	0	0	1
M4	3	2	1	6
M5	0	0	5	4
M7	1	0	0	0

Table 1. Assignment of short- and long-lived patients to our revealed gene expression groups in combination with meta-information about cytogenetic abnormality types and FAB types of the *DNMT3A*-mutant AML patients from TCGA. See also Supplementary Fig. 1 for an overview of the number of mutated genes per subgroup.

Based on this correlation analysis, we found that the first category, e.g. miRNAs with significant host gene expression correlation, contained 6 of the 25 differentially expressed miRNAs (Fig. 3B). The expression of *miR-199a-2*, whose gene co-localizes with the dynamin gene *DNM3*, was positively correlated with *DNM3* expression ($r = 0.432$, $P = 0.004$). Also the expression of *miR-3154* and *miR-199a-1*, which co-localize with the other two dynamin genes, were positively correlated with the expression of their host genes (*miR-3154* vs. *DNM1*: $r = 0.390$, $P = 0.011$; *miR-199a-1* vs. *DNM2*: $r = 0.266$, $P = 0.089$), although not statistically significant after correction for multiple testing (i.e., $q > 0.1$). Comparing short- to long-lived patients, the expression of *DNM1* and *DNM3* was moderately decreased, whereas the expression of *DNM2* did not differ between both subgroups (Supplementary Table 3). The other five miRNAs that had significantly positive expression correlations with their host genes were *miR-10a* (host gene *HOXB3*), *miR-126* (*EGFL7*), *miR-362* (*CLCN5*), *miR-26a-1* (*CTDSPL*) and *miR-551b* (*EGFEM1P*).

The second category contained 19 of 25 differentially expressed miRNAs that did not show coexpression with their host genes or are encoded in inter-genic regions and do not have a host gene (Fig. 3B). An association with AML has been reported previously for 13 of them (Supplementary Table 4). For instance, *miR-181a-2*, *miR-181b-2* and *miR-30a* are known to be associated with a favorable prognosis upon up-regulation of their expression[19,37], which is in line with a strong down-regulation of these three miRNAs in the short-lived relative to the long-lived subgroup. Similarly, we could reconfirm an up-regulation of *let-7b* in the context of *NPM1* mutations and a down-regulation of *miR-130a* in the context of *FLT3* mutations[37] in the short-lived subgroup. Further, we found a down-regulation of *miR-331* in the short-lived subgroup, which differs from[19] reporting that the up-regulation of *miR-331* was associated with poor prognosis. We also observed decreased expression of *miR-98* in the short-lived subgroup, which differs from previous findings that *miR-98* is up-regulated in the background of *NPM1* mutations (Supplementary Table 4). In addition, no direct associations with AML have been reported so far for the 4 miRNAs (*miR-153-2, miR-3065, miR-6718, miR-95*) (Supplementary Table 4), but associations with other types of cancer suggest that differences in their expression between short- and long-lived *DNMT3A*-mutant AML patients could also be important for prognosis (see "Discussion").

Regulatory networks reveal potential molecular major regulators distinguishing short- from long-lived *DNMT3A*-mutated patients. In order to investigate the combined effect of gene and miRNA expression on gene regulation we integrated these two types of data using a regulatory network-based approach. We started by reconstructing a signature gene-specific network to reveal potential regulators that distinguish short- from long-lived patients. Considering the 260 differentially expressed genes observed between both groups (Fig. 2A, Supplementary Table 3), we modeled the expression of a signature gene as a linear combination of the expression levels of the other 259 signature genes distinguishing short- from long-lived patients

Regulator gene	GeneCards annotation summary
Network module 1	
SLC4A1	Anion exchanger; role in O2/CO3 exchange in erythrocytes
HBM	Hemoglobin subunit Mu; iron ion and oxygen binding
RHD	Rh blood group D antigen; ammonium transmembrane transporter activity
GYPA	erythrocyte membrane protein; MN blood group receptor; hematopoietic stem cell differentiation; associated with Anemia, Autoimmune Hemolytic
CA1	Carbonate dehydratase and hydro-lyase activity; highest concentration in erythrocytes; nitrogen metabolism
Network module 2	
ZAP70	T cell receptor associated kinase; T cell development; lymphocyte activation
CD3D	Part of T cell receptor/CD3 complex; associated with immunodeficiencies
EVL	Enhances actin nucleation and polymerization; actin and profilin binding
IFITM1	Interferone-induced transmembrane protein; antiviral activity; cell adhesion and control of cell growth and migration; regulates osteoblast differentiation
Network module 3	
TNFRSF17	TNF receptor of major B lymphocytes; autoimmune response; transduces signals for cell survival and proliferation
IGKV4-1	V segment of variable domain of immunoglobulin light chain
IGKC	Constant region of immunoglobulin heavy chains
Network module 4	
GYPC	Erythrocyte membrane protein; Gerbich blood group; response to elevated platelet cytosolic Ca2+; regulation of mechanical cell stability
MREG	Melanoregulin; incorporation of pigments into hair; membrane fusing
Network module 5	
SORL1	Transmembrane signaling receptor activity; low-density lipoprotein binding
C1QTNF4	Pro-inflammatory cytokine; activation of NF-kappa-B; IL6 up-regulation
Network module 6	
RPS3	Ribosomal protein; mRNA activation
RPS19	Ribosomal protein; mRNA activation; associated with anemia

Table 2. Non-HOX network modules and potential major regulators.

(see "Methods" for details). The prediction of robust links between genes during reconstruction of the network was complicated due to the small number of *DNMT3A*-mutated patients. Therefore, we repeated the network inference 100 times with different, randomly selected training sets of patients to identify network links that robustly occurred in at least two-thirds of the networks with a q-value of 0.1 or smaller. This enabled to predict the expression values of on average 18.3% of the 260 signature genes. In a second step, we further added the expression values of all 514 miRNAs as additional predictors to the network model and repeated the analysis. This slightly improved the fraction of signature genes with predictable expression levels to 21.9%. The prediction accuracy of those genes, quantified by computing correlation coefficients between measured and predicted expression levels on the network-specific test sets, was high and significantly shifted into the positive range (mean correlation: 0.805, Wilcoxon signed rank test: $P < 0.0001$, Supplementary Fig. 5).

The resulting consensus network included 76 genes and 9 miRNAs (Fig. 4, Supplementary Table 5). This network consisted of several modules that were composed of two to eight genes. Two of the larger network modules were up-regulated in the short-lived subgroup and contained, respectively, four *HOXA* and three *HOXB* genes which are well-known major regulators of cell development and that have frequently been reported to be dysregulated in cancers including AML[38–40]. Six additional network modules with at least three genes and their components are summarized in Table 2. Each of the potential regulators in these network modules (labeled nodes in Fig. 4) was down-regulated in the short-lived compared to the long-lived subgroup. Interestingly, two of the six modules (network modules 1 and 4) contained genes that code for proteins expressed in erythrocytes or other blood components (*HBM, RHD, GYPA, GYPC, CA1*), or have been implicated in blood-associated diseases like anemia (*GYPA*) (Table 2). In addition, *RPS19* of network module 6, encoding a ribosomal protein, has been linked to anemia, too (Table 2). Genes of the remaining three modules (network modules 2, 3, and 5) are involved in innate or adaptive immunity (*ZAP70, CD3D, IFITM1, TNFRSF17, IGKV4-1, IGKC, C1QTNF4*) (Table 2).

We further analyzed the protein-coding genes that were directly connected to one of the nine miRNAs in the network representing potential targets for miRNA-based post-transcriptional regulation (Fig. 4). Among them were five genes with known transcription factor activity (*PBX3, HOXB3, LEF1, HOXA7, LBH*) and three genes with oncogenic potential (*PAPD7, PBX3, LEF1*) for which a role in other cancers has been suggested previously (Table 3). Interestingly, a role during leukemogenesis and/or implications for clinical prognosis in AML has been reported for eight of the nine miRNAs (Supplementary Table 4). This included the differential regulation of *let-7b* and *miR-130a* already mentioned above as well as of *miR-10a* and *miR-486* in the context of *NPM1* or *FLT3* mutations, effects on prognosis upon differential regulation of *miR-128-1* and *miR-150*, an increased cell survival and proliferation prompted by expression changes of *miR-196b* targeting *HOXB8*, and regulation of *miR-628* by cytokines[18,19,37,41,42].

miRNA	logFC	Connected gene	GeneCards annotation summary
hsa-let-7b	1.01	*PAPD7*	Poly(A) RNA polymerase; oncogenic MAPK signaling; DNA repair; sister chromatin adhesion
hsa-mir-10a	2.97	*PBX3*	Astrocytoma association; misregulation in cancer; transcription factor activity
hsa-mir-10a	2.97	*HOXB3*	Transcription factor in development, host gene of *hsa-mir-10a*
hsa-mir-128-1	− 0.54	*ARPP21*	cAMP-regulated phosphoprotein; nucleic acid and calmodulin binding; enriched expression in CNS
hsa-mir-130a	− 1.88	*FAM69B*	Cysteine-rich type II transmembrane protein of unknown function
hsa-mir-150	− 0.81	*LEF1*	T cell receptor binding; Wnt signaling, cancer association; transcription factor activity
hsa-mir-196b	1.16	*HOXA7*	Transcription factor in development
hsa-mir-486	− 0.93	*LBH*	transcriptional activator in mitogen-activated protein kinase signaling pathway
hsa-mir-628	− 0.60	*BEND2*	participation in protein and DNA interactions during chromatin restructuring or transcription
hsa-mir-6718	2.61	*LRMDA*	Leucin-rich; melanocyte differentiation

Table 3. Network miRNAs and potentially directly or indirectly regulated protein-coding genes. The logFC-column quantifies the expression level of the miRNA within the short-lived subgroup relative to the long-lived subgroup.

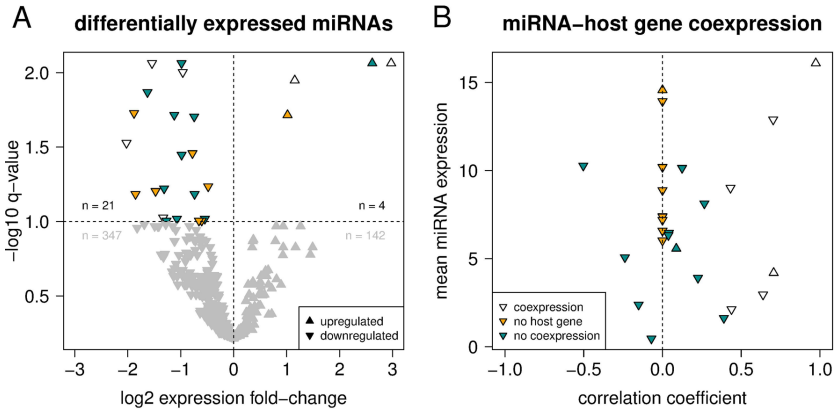

Figure 3. Differentially expressed miRNAs and co-expression of miRNAs and their host genes. (**A**) Volcano plot showing the relative expression change of 514 miRNAs between patients from the short-lived and long-lived subgroup. miRNAs with a significant change in expression ($q < 0.1$) are colored, others in gray. (**B**) Correlation of miRNA and corresponding host gene expression values across 42 *DNMT3A*-mutated patients. Pearson correlation coefficients were set to zero (dashed vertical line, yellow coloring) for miRNAs without a protein-coding host gene. For both figure panels, triangles indicate miRNAs that show (white) or do not show (turquoise) a significant coexpression (positive correlation) with their respective host gene ($q < 0.1$).

Validation based on independent *DNMT3A*-mutant AML patients. We considered gene mutation and gene expression data of independent *DNMT3A*-mutated AML patients from the German-Austrian AML Study Group[34,43–45] to analyze whether the characteristic gene mutation and expression profiles that distinguished short- and long-lived *DNMT3A*-mutated TCGA AML patients are also of potential prognostic relevance for other patients.

To analyze the transferability of the prognostic relevance of our initial grouping of gene mutation profiles of *DNMT3A*-mutated TCGA AML patients into a short- and long-lived subgroup (Fig. 1A, Supplementary Table 1), we considered gene mutation data of 208 *DNMT3A*-mutant AML patients from the German-Austrian AML Study Group that were initially treated in a similar manner followed by a bone marrow transplantation. We determined for each of these new patients the most similar *DNMT3A*-mutated TCGA AML patient and assigned its corresponding label (short- or long-lived) to the new patient. We found that the gene mutation profiles of short- and long-lived *DNMT3A*-mutated TCGA AML patients enabled to separate the 208 *DNMT3A*-mutant AML patients from the German-Austrian AML Study Group into a short- and long-lived subgroup that differed significantly in survival (Fig. 5A, log-rank test: $P < 0.003$).

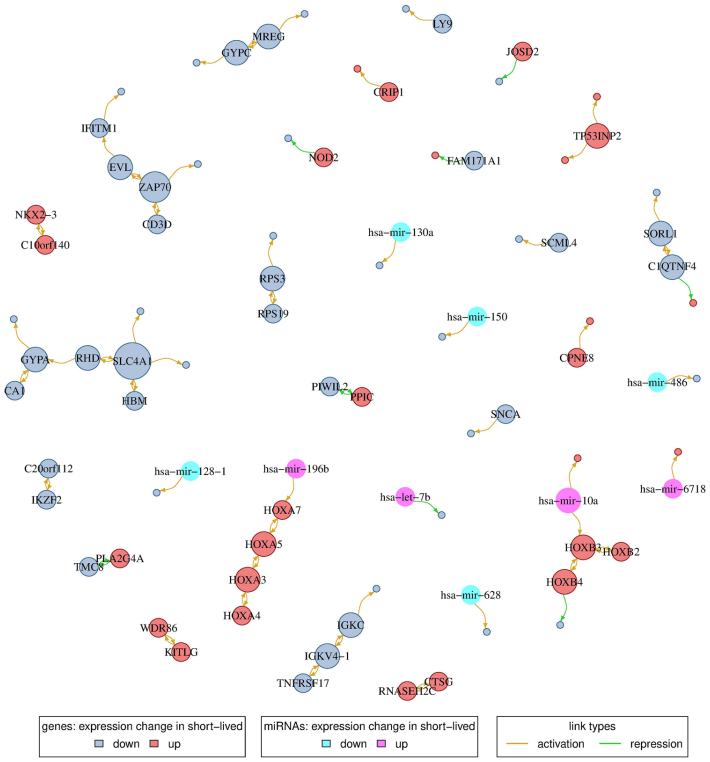

Figure 4. Gene and miRNA regulatory network. Nodes represent either genes that are differentially expressed between the two patient subgroups or miRNAs selected as predictors during network inference. Nodes are colored according to whether a gene/miRNA shows an increase or decrease in expression in the short-lived relative to the long-lived patient subgroup. Gene/miRNA names are shown for putative regulator nodes (out-degree > 0) with node sizes being proportional to their out-degree. Potential activating and repressing links are shown in yellow and green color, respectively; only links present in at least two-thirds of the networks were considered. Note that links can represent direct or indirect regulatory dependencies or may only represent correlations.

In addition, we also analyzed the transferability of the prognostic relevance of the characteristic gene expression signature that distinguished short- and long-lived *DNMT3A*-mutated TCGA AML patients (Fig. 2A, Supplementary Table 2, $q < 0.1$). We therefore considered gene expression data of 63 *DNMT3A*-mutant AML patients from the University Hospital of Ulm that were also part of two clinical trials of the German-Austrian AML Study Group[44,45]. The majority of these patients received a bone marrow transplantation (47 of 63). We determined for each of these new patients the similarity to the TCGA-based short- and long-lived signature and assigned to each patient the label of the most similar class (short- or long-lived). We found that the characteristic gene expression signature that distinguished short- and long-lived *DNMT3A*-mutated TCGA AML patients also enabled to separate the 63 *DNMT3A*-mutant AML patients from the University Hospital of Ulm into a short- and long-lived subgroup that differed significantly in survival (Fig. 5B, log-rank test: $P < 0.03$). This separation significance was further improved when we only considered the 47 patients that received a bone marrow transplantation (log-rank test: $P < 0.016$).

Further, we also analyzed if our short- and long-lived classification of *DNMT3A*-mutant AML patients can help to improve the widely considered European LeukemiaNet (ELN) prognostic scoring systems[46,47]. Risk classifications according to the ELN 2010 system[46] were publicly available for 192 of 208 patients of the German-Austrian AML Study Group considered for the gene mutation-based validation[34]. Our additional stratification into short- and long-lived patients significantly improved the risk stratification of patients of the ELN 2010

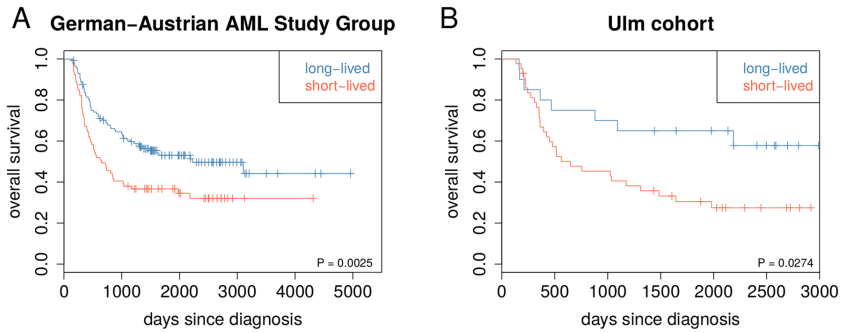

Figure 5. Validation based on independent *DNMT3A*-mutant AML patients. (**A**) Gene mutation based validation. Kaplan-Meier curves for an independent cohort of 208 *DNMT3A*-mutated AML with bone marrow transplantation from the German-Austrian AML Study Group. For each of these patients, the most similar *DNMT3A*-mutated AML patient of the TCGA cohort was determined by counting mismatches between the corresponding gene mutation profiles. Each patient was assigned to the short-lived or to the long-lived group based on the class label of the most similar TCGA patient (short-lived: red, 79 patients; long-lived: blue, 129 patients). Log-rank test for short- vs. long-lived: $P < 0.003$. (**B**) Gene expression based validation. Kaplan-Meier curves for an independent cohort of 63 *DNMT3A*-mutated AML patients from the University Hospital of Ulm that were also part of the German-Austrian AML Study Group. For each of these patients, correlations between its signature gene expression profile with the average short-lived and long-lived signature gene expression profiles of the *DNMT3A*-mutated AML patients from TCGA were computed. Each patient was assigned to the short-lived or to the long-lived group based on the maximum of both correlations (short-lived: red, 43 patients; long-lived: blue, 20 patients). Log-rank test for short- vs. long-lived: $P < 0.03$.

intermediate-1 risk category (Fig. 6A, log-rank test: $P = 0.0008$). Also patients of the ELN 2010 adverse risk group could potentially benefit from our additional stratification (Supplementary Fig. 7). Further, our additional stratification had no impact on the stratification of patients of the ELN 2010 intermediate-2 or favorable risk categories (Supplementary Fig. 7). We also analyzed the impact of our short- and long-lived stratification on the revised ELN 2017 risk classification[47]. This was possible for 134 of 208 patients, but excluded patients with a *FLT3*-ITD mutation, because the *FLT3*-ITD-to-wild-type allelic ratios required for a reclassification were not publicly available[48]. In this limited analysis, we found that our additional stratification had no impact on the ELN 2017 favorable risk category, but there were too few patients to interpret the additional stratification of the other risk categories (Supplementary Fig. 9A-B).

In addition, the ELN 2010 risk classification was also available for 62 of 63 patients of the Ulm cohort considered for the gene expression-based validation[34]. We found that our additional stratification into short- and long-lived patients again significantly improved the risk stratification for patients of the ELN 2010 intermediate-1 risk category (Fig. 6B, log-rank test: $P = 0.0011$). Interestingly, there was also a clear tendency that patients of the ELN 2010 favorable risk category could potentially benefit from our additional stratification (Supplementary Fig. 8). Further, our additional stratification had no impact on patients of the ELN 2010 intermediate-2 or adverse risk categories (Supplementary Fig. 8). We also analyzed the impact of our additional stratification on the revised ELN 2017 risk classification that was available for 37 of 63 patients[48]. We observed that patients of the ELN 2017 favorable risk group can potentially benefit from our additional stratification (Supplementary Fig. 9C,D). Similar trends were also present for the ELN 2017 intermediate and adverse risk categories. However, there were too few patients within the different ELN 2017 risk categories to analyze the significance of these trends.

Nevertheless, all these results round off our different computational studies for the TCGA cohort and indicate that the characteristic discriminative gene mutation and expression signatures that distinguished short- from long-lived *DNMT3A*-mutated TCGA AML patients are also predictive for other independent patient cohorts and potentially useful to improve patient stratification.

Discussion

A somatic mutation of *DNMT3A* occurs in about one fourth of adult AML cases. Mutations of this gene have frequently been associated with poor survival[9,14,25,30,31], but also substantially longer survival or long-term remissions have been reported for some *DNMT3A*-mutant AML patients[32,33]. Detailed molecular differences that may contribute to these survival differences have not been characterized so far. This motivated us to analyze all *DNMT3A*-mutant patients of the TCGA AML cohort with the help of well-established computational tools. We identified two robust subgroups of *DNMT3A*-mutant patients purely based on clustering of somatic gene mutation profiles and further found that both subgroups showed significant survival differences.

Further comparisons showed that the short-lived subgroup had a strong enrichment of mutations of the R882 codon of the catalytic methyltransferase domain of *DNMT3A*, whereas the number of R882 and non-R882 mutations was nearly equal within the long-lived subgroup. This mutation type-specific effect on prognosis has been

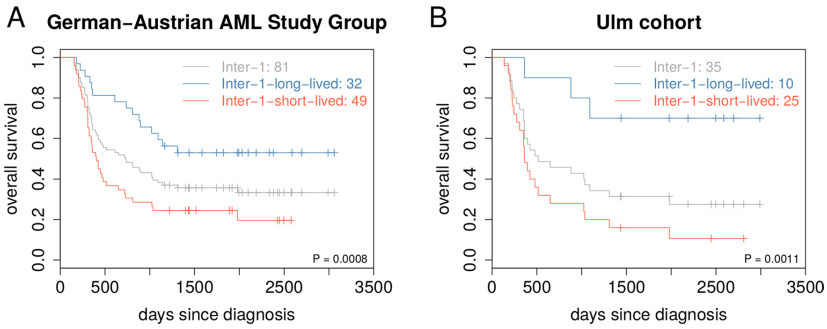

Figure 6. Improvement of ELN 2010 risk classification by additional short- and long-lived stratification. (**A**) Gene mutation-based validation of the 81 independent validation patients from the German-Austrian AML Study Group of the ELN 2010 risk category intermediate-1 (Inter-1). For each of these patients, the most similar *DNMT3A*-mutated AML patient of the TCGA cohort was determined by counting mismatches between the corresponding gene mutation profiles. Each patient was assigned to the short-lived or to the long-lived subgroup based on the class label of the most similar TCGA patient. Kaplan-Meier curves of this additional stratification are shown in red for the 49 Inter-1-short-lived patients and in blue for the 32 Inter-1-long-lived patients. The basic Kaplan-Meier curve without additional stratification of these patients is shown in grey. Log-rank test for Inter-1-short- vs. Inter-1-long-lived: P = 0.0008. A global overview of the additional stratification of all ELN 2010 risk categories is shown in Supplementary Fig. 7. (**B**) Gene expression based validation of the 35 independent validation patients from the Ulm cohort of the ELN 2010 risk category intermediate-1 (Inter-1). For each of these patients, correlations between its signature gene expression profile with the average short-lived and long-lived signature gene expression profiles of the *DNMT3A*-mutated AML patients from TCGA were computed. Each patient was assigned to the short-lived or to the long-lived subgroup based on the maximum of both correlations. Kaplan-Meier curves of this additional stratification are shown in red for the 25 Inter-1-short-lived patients and in blue for the 10 Inter-1-long-lived patients. The basic Kaplan-Meier curve without additional stratification of these patients is shown in grey. Log-rank test for Inter-1-short- vs. Inter-1-long-lived: P = 0.0011. A global overview of the additional stratification of all ELN 2010 risk categories is shown in Supplementary Fig. 8.

noted before[25], but was not sufficient for a full discrimination of our two subgroups. Thus, additional molecular factors are likely to contribute to the observed survival differences.

Mutated *DNMT3A* has been shown to induce genomic instability in a human leukemic cell line model[49]. We therefore compared the short- and long-lived subgroup in terms of mutated genes and cytogenetic rearrangements. Interestingly, the number of mutated genes was significantly smaller in the short-lived subgroup. In addition, the majority of patients of both subgroups had normal cytogenetic profiles, but especially some patients of the long-lived subgroup showed duplications or rearrangements of chromosome 8 that have not been observed within the short-lived subgroup. Thus, the overall shorter survival of patients in the short-lived group cannot be explained by a greater mutational burden or increased rates of abnormal cytogenetic profiles.

We found *NPM1* and/or *FLT3* mutations in every short-lived patient but only in few long-lived patients. This overrepresentation of *NPM1* and/or *FLT3* mutations in the short-lived subgroup is not unexpected, because *DNMT3A*, *NPM1*, and *FLT3* are the most frequently mutated genes found in AML[10]. The co-occurrence of mutations of all three genes has previously been suggested to define a specific subtype of AML with unique epigenetic features[10] and frequent mutations of *NPM1* and *FLT3* in *DNMT3A*-mutant patients have also been observed in other AML studies[9,25,50]. Importantly, *NPM1* and *FLT3* are both established prognostic markers in routine clinical practice[47]. *FLT3* mutations (ITD: internal tandem duplication of the juxtamembrane region, TKD: point mutations in the second tyrosine kinase domain) have been associated with increased relapse risk and poor outcome of AML patients[51,52]. The frequency of *FLT3-ITD* and *FLT3*-TKD mutations was nearly identical in the short-lived subgroup and an additional stratification according to the specific type of *FLT3* mutation did not further improve our classification of *DNMT3A*-mutant AML patients from TCGA. *NPM1* mutations frequently co-occur together with *FLT3*-ITD mutations, which counteracts a favorable prognosis that is observed for AML patients that only have a *NPM1* mutation but no *FLT3* mutation[12,47,53]. This is also supported by our two subgroups. The majority of short-lived patients had co-mutations of *NPM1* and *FLT3*, whereas long-lived patients did not show *NPM1* mutations in the background of *FLT3* mutations. Thus, our study clearly indicates that *NPM1* and/or *FLT3* mutations are likely to contribute to the prognosis of *DNMT3A*-mutant patients. This is supported by the previous findings that *DNMT3A* mutations jointly act with *FLT3* and *NPM1* mutations to promote resistance to anthracycline chemotherapy[54] and that concurrent mutations of *DNMT3A*, *FLT3*, and *NPM1* have also been associated with poor prognosis of AML patients[55]. In addition, all our survival analyses in combination with the presence or absence of *DNMT3A* mutations further support that *DNMT3A* mutations have an additional negative impact on survival that is independent of *FLT3* and/or *NPM1* mutations or co-mutations of both genes.

This is supported by findings for the presence or absence of *DNMT3A* mutations in AML patients with *FLT3* mutations[50]. Additional experiments should be done to elucidate whether the *DNMT3A* mutation cooperates with *FLT3* and *NPM1* co-mutations.

Since an increased rate of *DNMT3A*-R882 mutations was observed for our short-lived subgroup, we also analyzed a large independent cohort of AML patients[34] and observed an enrichment of *DNMT3A*-R882 and *NPM1* co-mutations and an enrichment of concurrent *DNMT3A*-R882, *NPM1*, and *FLT3* mutations compared to AML patients with *DNMT3A* mutations that did not affect the R882 codon. Interestingly, the blood composition of these groups differed in dependency of the type of the *DNMT3A* mutation indicating an impact on the differentiation capabilities of AML cells. Additional experiments are required to validate the accumulation of *NPM1* and/or *FLT3* mutations and to analyze the differentiation capabilities of AML cells in the background of specific *DNMT3A* mutations.

We further compared the gene expression profiles of the short- and long-lived subgroup revealing a molecular signature of 260 protein-coding genes that distinguished both subgroups. This signature included many transcription factors and genes of cancer-associated pathways like p53, VEGF and PI3K-Akt signaling and DNA replication. Importantly, a clustering of the patients based on these signature genes largely recapitulated the short-lived and long-lived subgroup and further revealed a set of patients with mixed expression levels. This indicates that at least three different transcriptional programs are associated with survival differences of *DNMT3A*-mutant AML patients. Further, it is important to note that *NPM1* or *FLT3* mutations or co-mutations of both genes that were observed for each short-lived patient also contribute to the observed expression differences. Therefore, our comparison of short- and long-lived gene expression profiles does not allow to disentangle the individual contributions of *DNMT3A*, *FLT3*, or *NPM1* mutations. Still, all our survival analyses comparing the presence or absence of *DNMT3A* mutations in the background of *NPM1* and/or *FLT3* mutations suggest an additional contribution of *DNMT3A* mutations. This additional contribution is also included in the gene expression signature and further supported by our gene expression-based classification of independent *DNMT3A*-mutant AML patients.

Alterations of miRNA expression profiles play an important role in AML[18,19]. We therefore compared the miRNA expression profiles of the short- and long-lived subgroup. We revealed a dominant trend of miRNA downregulation in the short-lived subgroup suggesting a wide-spread activation of otherwise repressed protein-coding genes, including known AML oncogenes and other oncogenes that were not associated with AML before. Further, associations with AML prognosis and/or mutation of *NPM1* and *FLT3* have already been reported for most miRNAs, but we also identified four miRNAs that have not been reported for AML so far. This included three miRNAs that were downregulated in the short-lived subgroup (i) *miR-153-2* implicated in brain, lung, liver and epithelial cancers[56–59], (ii) *miR-3065* for which an association with altered gene expression regulation in breast tumors was suggested[60], and (iii) *miR-95* known to be differentially expressed in different human cancers[61–63] with shown impacts on cell proliferation, invasion, migration, and apoptosis in a pancreatic tumor cell line and in hepatocellular carcinoma[61,63]. We did not find cancer-associated reports for the fourth miRNA *miR-6718*, but its strong 2.6-fold upregulation in the short-lived subgroup and the selection by our regulatory network approach suggests an association with prognosis. In addition, we discovered a downregulation of all three dynamin genes in the short-lived subgroup based on their co-localized miRNAs. This may have an impact on endocytosis, asymmetric cell divisions, and blockage of immune signals[64–67]. This suggests that these miRNAs could represent important biomarker candidates to discriminate between short- and long-lived *DNMT3A*-mutant AML patients. Additional experimental studies should be done to validate these potential markers and to better understand how they alter molecular mechanisms in *DNMT3A*-mutant AML patients.

We also learned gene regulatory networks to identify potential major regulators and to delineate modules of protein-coding and miRNA genes that were altered between the short- and long-lived subgroup. Due to the relatively small number of AML patients with *DNMT3A* mutations, our consensus network contained only relatively few genes compared to networks from similar studies of other cancers[68,69]. Still, those genes present in the network and the links between them were inferred with high confidence. It is important to note that the inferred links between genes can reflect direct or indirect regulatory dependencies or only represent correlations, because our network reconstruction method is based on correlations between gene expression levels. Yet, larger subnetworks can still point toward cellular pathways that are altered between both subgroups. Our revealed modules suggest alterations of several cellular processes in short-lived relative to long-lived patients. This included genes of the PI3K-Akt and p53 signaling pathway involved in AML[70,71] and an upregulation of HOX genes altered in leukemia[38,40]. In addition, we also identified genes that are expressed in different blood components. This included three genes downregulated in the short-lived subgroup - *SLC4A1*, *GYPA* and *RPS19* - that have previously been associated with anemia[72–74]. Notably, *SLC4A1* and its co-factor *GYPA* play a major role in oxygen and carbon dioxide exchange in erythrocytes[75,76] and their downregulation in the short-lived subgroup could be associated with less differentiated leukemic cells. Further, we found three gene modules with immunity-related functions downregulated in the short-lived subgroup and an increased number of differentially expressed cytokine receptor signaling pathway genes suggesting that immune evasion might be more effective in the short-lived subgroup, but immunosuppression in AML is still poorly understood[77]. The identified putative major regulators potentially represent important candidates for the development of biomarkers that could distinguish between short- and long-lived patients. Additional experimental validation studies are required to test their prognostic potential and to further characterize their functional role in *DNMT3A*-mutant AML patients.

Moreover, we also showed that the characteristic gene mutation and expression signatures that distinguished short- from long-lived *DNMT3A*-mutant TCGA AML patients contain relevant information that can be used to classify other independent *DNMT3A*-mutant AML patients as short- or long-lived. We demonstrated this for *DNMT3A*-mutant AML patients from the German-Austrian AML Study Group. Thus, our revealed molecular signatures could potentially provide a useful basis to enable a better stratification of *DNMT3A*-mutant AML patients to more precisely identify patients that are of high risk for a fast relapse. This is also supported by the

interpretation of our results with respect to the cytogenetic and molecular risk classification provided by TCGA, which assigned more than 82% of the *DNMT3A*-mutant patients to the intermediate risk group, whereas the remaining patients were assigned to the poor risk group, except one unclassified patient. Since our approach significantly improved the stratification of these TCGA patients, this also clearly indicates that our approach can improve this cytogenetic and molecular risk classification. The value of our approach is further supported by the significant improvement of the stratification of patients that were assigned to intermediate-1 risk category according to the ELN 2010 prognostic scoring system[46]. Further, we also observed potential benefits of our additional stratification for the ELN 2010 risk categories favorable and adverse, but more patients would have been required for a robust significance analysis. In addition, an analysis of the revised ELN 2017 risk categories[47] indicated that the favorable and intermediate risk groups could potentially benefit from our additional stratification, but this should be taken with caution, because this analysis was only possible for a subset of our validation patients. Additional validation studies are necessary to analyze how our findings generalize to other patient cohorts and how they impact on patient outcome. Future studies should include an extended comparison to the revised ELN 2017 scoring system. This was only partly possible in our study, because molecular data such as the *FLT3*-ITD-to-wild-type allelic ratio required for a reclassification were not publicly available for the patients considered in our study. However, a recent study has shown that *DNMT3A*-mutant AML patients have a worse prognosis than *DNMT3A* wild type patients for individual ELN 2017 risk categories[48]. Our study indicates that an improved stratification of individual risk categories might even be possible within the group of *DNMT3A*-mutant AML patients.

Our study represents the first in-depth computational approach to identify molecular factors associated with survival differences of *DNMT3A*-mutant AML patients. This may provide a basis to develop molecular markers for improved patient stratification. Future studies are required to further analyze and validate the findings of our computational study.

Methods

Molecular data. Gene and miRNA expression data and somatic mutations of patients from the TCGA AML cohort were obtained from the TCGA data portal (gdc.cancer.gov). After excluding lowly expressed genes with a counts per million value smaller one in two-thirds or more of the patients, we normalized the raw expression data using the R/Bioconductor package *limma* with normalization method cyclic loess[78]. By using information on the *DNMT3A* mutational status from the somatic mutation data, we determined 51 *DNMT3A*-mutated AML patients and derived corresponding gene expression (47 of 51 patients, 15,623 genes) and miRNA (42 of 51 patients, 514 miRNAs) data sets. Details to *DNMT3A*-mutations and processed data sets are provided in Supplementary Table 1.

Clustering based on somatic mutation data. We considered each of the 51 AML patients with a *DNMT3A* mutation and created for each patient its binary gene mutation profile by setting the entry of each gene to one (mutated) or to zero (not mutated) in dependency of the patient-specific gene mutation status. Next, we performed a hierarchical clustering of tumors based on binary profiles of the somatic mutation data using R with 1 minus Pearson correlation as distance measure with distances ranging from zero (two completely identical mutation profiles) to one (two completely different mutation profiles) in combination with Ward's clustering method (*ward.D2*)[79]. Note that the Pearson correlation coefficient of two binary variables is equal to the phi coefficient[80]. Hierarchical clustering initially considers each patient as a separate cluster and then repeats the following two steps until all clusters are merged together: (i) identification of the two clusters with the smallest distance followed by (ii) merging of these two clusters into a joint cluster. These iterative merging steps enable to reveal the hierarchical relationships between the clusters that are stored in a tree-structure called dendrogram. Two tumor subgroups were derived by cutting the resulting clustering dendrogram into two subtrees. These subgroups were named 'short-lived' and 'long-lived' according to survival differences between the subgroups (see below). The TCGA identifiers for patients of the short- and long-lived subgroup are provided in Supplementary Table 2. To assess the robustness of this patient clustering, we excluded k randomly selected patients, repeated the clustering into two groups as described above, and performed a log-rank test for survival differences between the groups (see below). We tested $k = 2, 4, 6, 8, 10$, and repeated the analysis 10,000 times for each k. We did not test larger values of k owing to the relatively small number of *DNMT3A*-mutated AML patients in the data set.

Survival analysis. Information about days to death (for patients with status 'Dead') or days to last follow-up (for patients with status 'Alive') was taken from the TCGA clinical data (Supplementary Table 2). Last follow-up events were considered as non-informative censoring events. We generated survival curves and performed log-rank tests using the R package *survival*[81].

Identification of differentially expressed genes and miRNAs. Differential gene and miRNA expression analysis between the short- and long-lived subgroup was done following *limma's* standard workflow[78]. Results of the gene and miRNA expression analysis are provided in Supplementary Table 3. Differentially expressed (signature) genes or miRNAs were selected using an FDR-adjusted p-value (q-value) cut-off of 0.1.

Gene and pathway annotation enrichment analysis. Gene, signaling pathway, and metabolome annotations were obtained from[35]. The number of signature genes per annotation category was counted separately for up- and downregulated genes and their significance of enrichment per category was calculated using Fisher's exact test.

Signature-specific regulatory network inference. We inferred transcriptional regulatory networks that model the expression of a signature gene as a linear combination of weighted expression values of the other signature genes and, optionally, of miRNAs. Mathematical details to the underlying linear model are provided in[35,82]. This approach has further been applied in similar studies of other human cancers[68,69,83,84]. We learned two types of networks using (i) the expression values of signature genes and (ii) the expression values of signature genes and miRNAs as predictors. miRNA expression values were set to zero for patients without available miRNA profiles. Lasso regression[85] in combination with a significance test for lasso[86] were used to estimate the coefficients and their corresponding significance of the predictors for each signature gene-specific linear model[82]. This sparse regression approach selects the most relevant predictors that best explain the observed expression levels of a signature gene across the *DNMT3A*-mutant AML patients.

Both network approaches were validated through cross-validation by repeated random sub-sampling. To this end, the data was randomly partitioned into a training set constituting three-quarter of the *DNMT3A*-mutated AML patients and a test set constituting the remaining one-fourth of patients. A network was constructed on the training data, and the expression of the signature genes was predicted and compared to the experimentally measured expression for the test data. This procedure was repeated 100 times. To assess prediction accuracy, we calculated Pearson correlation coefficients of predicted and measured gene expression averaged over the 100 networks. A consensus network was constructed by including all links with q-values of 0.1 or smaller that were predicted in at least two-thirds of the 100 networks.

Validation based on independent *DNMT3A*-mutant AML patients. To validate the separation capability of the characteristic gene mutation profiles of short- and long-lived *DNMT3A*-mutant AML patients from TCGA, we downloaded publicly available gene mutation profiles and clinical data of AML patients from https://github.com/gerstung-lab/AML-multistage/tree/master/data[34]. We considered all 208 *DNMT3A*-mutated AML patients from the German-Austrian AML Study Group (AMLSG)[43–45] that received a bone marrow transplantation to obtain a large validation cohort of patients that were treated similarly. The majority of these patients (204 of 208) were part of two clinical trials (AMLHD98A: 77[44]; AMLSG0704: 127[45]) focusing on AML patients in the age range of 18 to 65. The other four patients were part of the AMLHD98B trial that considered AML patients of age 61 or older[43]. Considered patients from AMLHD98A received an induction chemotherapy with idarubicin, cytarabine and etoposide (ICE) followed by allogeneic transplants. Treatment of considered patients form AMLSG0704 and AMLHD98B was similar, but patients were randomly assigned to receive ICE or ICE plus all-trans retinoic acid (ATRA) as induction therapy before transplantation[12]. We computed the most similar *DNMT3A*-mutated TCGA AML patient for each of these 208 patients by counting mismatches between each corresponding pair of gene mutation profiles. We had to focus on 31 genes that overlapped with the mutated genes of *DNMT3A*-mutated TCGA AML patients, because the data from[34] was obtained by targeted sequencing of selected cancer genes. We assigned each of the 208 patients either to the short- or to the long-lived group based on the class label of the most similar TCGA patient and performed a survival analysis as described in the section 'Survival analysis' above (Supplementary Table 6). Further, we also considered the European LeukemiaNet (ELN) 2010 risk classification[46] available for 192 of 208 patients to analyze if an additional stratification of each individual ELN 2010 risk category based on our short- and long-lived classification can improve this prognostic scoring system (Supplementary Table 6). We realized this by an extended survival analysis for the patients of an individual risk category in comparison to our corresponding short- and long-lived classifications of these patients. Similarly, we also analyzed our stratification into short- and long-lived patients considering the revised ELN 2017 risk classification[47]. This was only possible for 134 of 208 validation patients that were reclassified in[48] (Supplementary Table 6). The other validation patients could not be considered, because *FLT3*-ITD-to-wild-type allelic ratios required for a reclassification were not publicly available.

To validate the separation capability of the gene expression signature of short- and long-lived *DNMT3A*-mutant AML patients from TCGA, we considered a cohort of 218 AML patients from the University Hospital of Ulm of which 63 had a *DNMT3A* mutation. The majority of these 63 patients were part of the AMLSG0704 clinical trial[45] (59) and the remaining 4 patients were part of the AMLHD98A clinical trial[44] of the German-Austrian AML Study Group. The majority of these patients received a bone marrow transplantation (47 of 63). The AML gene expression profiles of these patients were measured on Affymetrix HG-U133 Plus 2 microarrays. We normalized the gene expression data set using GCRMA[87] in combination with a BrainArray design file (HGU133Plus2_Hs_ENTREZG 15.0.0). We focused on the 257 signature genes of the 260 signature genes from our TCGA analysis (Fig. 2A, Supplementary Table 3, $q < 0.1$) that were measured on the Affymetrix arrays. We computed for each of the 63 *DNMT3A*-mutated patients rank-based correlations (Kendall's tau) between its signature gene expression profile and the average short-lived and long-lived signature gene expression profiles of the *DNMT3A*-mutated AML patients from TCGA. We assigned each patient either to the short-lived or to the long-lived group based on the maximum of both correlations and performed a survival analysis as described above (Supplementary Table 7). We also repeated this analysis only focusing on the 47 patients that received a bone marrow transplantation. Further, we again considered the ELN 2010 risk classification[46] available for 62 of 63 patients (Supplementary Table 7) and performed an additional survival analysis to analyze if our short- and long-lived classification can improve the individual risk categories. Similarly, we analyzed our short- and long-lived stratification considering the revised ELN 2017 risk classification[47] for the subset of 37 of 63 validation patients that could be reclassified in[48] (Supplementary Table 7).

Ethical approval and informed consent. Not applicable. No ethical approval was required for this study. All utilized public omics data sets were generated by otherswho obtained ethical approval.

Data availability

Molecular data and meta-information of all considered TCGA AML patients are publicly available from The Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Additional files attached to this manuscript contain considered molecular data, survival information, and learned network links. Basic implementations of the algorithms considered for network inference are publicly available from GitHub (https://github.com/seifemi/regNet).

Received: 27 February 2019; Accepted: 13 July 2020
Published online: 29 July 2020

References

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545–1602 (2016).
2. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1459–1544 (2016).
3. Chabner, B. & Dan, L. *Harrisons Manual of Oncology* (McGraw-Hill Professional, New York, 2014). http://www.myilibrary.com?id=546719.OCLC:879790976.
4. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
5. Fialkow, P. J., Janssen, J. W. & Bartram, C. R. Clonal remissions in acute nonlymphocytic leukemia: evidence for a multistep pathogenesis of the malignancy. *Blood* **77**, 1415–1417 (1991).
6. Byrd, J. C. *et al.* Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood* **100**, 4325–4336 (2002).
7. Slovak, M. L. *et al.* Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. *Blood* **96**, 4075–4083 (2000).
8. Wheatley, K. *et al.* A simple, robust, validated and highly predictive index for the determination of risk-directed therapy in acute myeloid leukaemia derived from the MRC AML 10 trial. United Kingdom Medical Research Council Adult and Childhood Leukaemia Working Parties. *Br. J. Haematol.* **107**, 69–79 (1999).
9. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
10. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
11. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **526**, 526–531 (2018).
12. Papaemmanuil, E. *et al.* Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
13. Patel, J. P. *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1079–1089 (2012).
14. Ribeiro, A. F. T. *et al.* Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood* **119**, 5824–5831 (2012).
15. Bullinger, L. & Valk, P. Gene expression profiling in acute myeloid leukemia. *J. Clin. Oncol.* **23**, 6296–305 (2005).
16. Verhaak, R. *et al.* Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94**, 131–4 (2009).
17. Wouters, B. J., Löwenberg, B. & Delwel, R. A decade of genome-wide gene expression profiling in acute myeloid leukemia: flashback and prospects. *Blood* **113**, 291–298 (2009).
18. Marcucci, G., Mrózek, K., Radmacher, M. D., Garzon, R. & Bloomfield, C. D. The prognostic and functional role of microRNAs in acute myeloid leukemia. *Blood* **117**, 1121–1129 (2011).
19. Liao, Q., Wang, B., Li, X. & Jiang, G. miRNAs in acute myeloid leukemia. *Oncotarget* **8**, 3666–3682 (2017).
20. Shah, M. Y. & Licht, J. D. DNMT3A mutations in acute myeloid leukemia. *Nat. Genet.* **43**, 289–290 (2011).
21. Xu, F. *et al.* Molecular and enzymatic profiles of mammalian DNA methyltransferases: structures and targets for drugs. *Curr. Med. Chem.* **17**, 4052–4071 (2010).
22. Jurkowska, R. Z., Jurkowski, T. P. & Jeltsch, A. Structure and function of mammalian DNA methyltransferases. *Chembiochem* **12**, 206–222 (2011).
23. Challen, G. A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2011).
24. Yang, L., Rau, R. & Goodell, M. A. DNMT3A in haematological malignancies. *Nat. Rev. Cancer* **15**, 152–165 (2015).
25. Gaidzik, V. I. *et al.* Clinical impact of DNMT3A mutations in younger adult patients with acute myeloid leukemia: results of the AML Study Group (AMLSG). *Blood* **121**, 4769–4777 (2013).
26. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
27. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
28. Shivarov, V., Gueorguieva, R., Stoimenov, A. & Tiu, R. DNMT3A mutation is a poor prognosis biomarker in AML: results of a meta-analysis of 4500 AML patients. *Leuk. Res.* **37**, 1445–1450 (2013).
29. Kumar, D., Mehta, A., Panigrahi, M. K., Nath, S. & Saikia, K. K. DNMT3A (R882) mutation features and prognostic effect in acute myeloid leukemia in Coexistent with NPM1 and FLT3 mutations. *Hematol. Oncol. Stem Cell Ther.* **11**, 82–89 (2018).
30. Renneville, A. *et al.* Prognostic significance of DNA methyltransferase 3a mutations in cytogenetically normal acute myeloid leukemia: a study by the Acute Leukemia French Association. *Leukemia* **26**, 1247–1254 (2012).
31. Thol, F. *et al.* Incidence and prognostic influence of DNMT3A mutations in acute myeloid leukemia. *J. Clin.Oncol.* **29**, 2889–2896 (2011).
32. Ploen, G. G. *et al.* Persistence of DNMT3A mutations at long-term remission in adult patients with AML. *Br. J. Haematol.* **167**, 478–486 (2014).
33. Sun, Y. *et al.* Persistent DNMT3A mutation burden in DNMT3A mutated adult cytogenetically normal acute myeloid leukemia patients in long-term remission. *Leuk. Res.* **49**, 102–107 (2016).
34. Gerstung, M. *et al.* Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49**, 332–340 (2017).
35. Seifert, M., Friedrich, B. & Beyer, A. Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. *Genome Biol.* **17**(1), 204 (2016).
36. Blau, O. *et al.* DNMT3A Mutations in AML patients: prognostic impact and comparative analysis of mutations burden in diagnostic samples, after standard therapy, and after allogeneic stem cell transplantation. *Blood* **128**, 2891 (2016).
37. Wang, X., Chen, H., Bai, J. & He, A. MicroRNA: an important regulator in acute myeloid leukemia. *Cell Biol. Int.* **41**, 936–945 (2017).

38. Alharbi, R. A., Pettengell, R., Pandha, H. S. & Morgan, R. The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia* **27**, 1000–1008 (2013).
39. Celetti, A. *et al.* Characteristic patterns of HOX gene expression in different types of human leukemia. *Int. J. Cancer* **53**, 237–244 (1993).
40. De Braekeleer, E. *et al.* Hox gene dysregulation in acute myeloid leukemia. *Future Oncol.* **10**, 475–495 (2014).
41. Favreau, A. J. & Sathyanarayana, P. miR-590-5p, miR-219-5p, miR-15b and miR-628-5p are commonly regulated by IL-3, GM-CSF and G-CSF in acute myeloid leukemia. *Leuk. Res.* **36**, 334–341 (2012).
42. Garzon, R. *et al.* Distinctive microRNA signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. *Proc. Natl. Acad. Sci. USA* **105**, 3945–3950 (2008).
43. Schlenk, R. F. *et al.* Phase III study of all-trans retinoic acid in previously untreated patients 61 years or older with acute myeloid leukemia. *Leukemia* **18**, 1798–803 (2004).
44. Schlenk, R. F. *et al.* Prospective evaluation of allogeneic hematopoietic stem-cell transplantation from matched related and matched unrelated donors in younger adults with high-risk acute myeloid leukemia: German-Austrian trial AMLHD98A. *J. Clin. Oncol.* **28**, 4642–8 (2010).
45. Schlenk, R. F. *et al.* All-trans retinoic acid as adjunct to intensive treatment in younger adult patients with acute myeloid leukemia: results of the randomized AMLSG 07–04 study. *Ann. Hematol.* **95**, 1931–1942 (2016).
46. Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on Behalf of the European LeukemiaNet. *Blood* **115**, 453–74 (2010).
47. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
48. Herold, T. *et al.* Validation and refinement of the revised 2017 European LeukemiaNet genetic risk stratification of acute myeloid leukemia. *Leukemia* (2020).
49. Banaszak, L. G. *et al.* Crispr/Cas9-induced DNMT3A mutations in the K562 human leukemic cell line as a model of DNMT3A-mutated leukemogenesis. *Blood* **128**, 2704 (2016).
50. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–33 (2010).
51. Thiede, C. *et al.* Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood* **99**, 4326–35 (2002).
52. Yanada, M., Matsuo, K., Suzuki, T., Kiyoi, H. & Naoe, T. Prognostic significance of FLT3 internal tandem duplication and tyrosine kinase domain mutations for acute myeloid leukemia: a meta-analysis. *Leukemia* **19**, 1345–9 (2005).
53. Thiede, C. *et al.* Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood* **107**, 4011–20 (2006).
54. Guryanova, O. A. *et al.* Dnmt3a mutations promote anthracycline resistance in acute myeloid leukemia via impaired nucleosome remodeling. *Nat. Med.* **22**, 1488–1495 (2016).
55. Loghavi, S. *et al.* Clinical features of de novo acute myeloid leukemia with concurrent DNMT3A, FLT3 and NPM1 mutations. *J. Hematol. Oncol.* **7**, 74 (2014).
56. Ghasemi, A., Fallah, S. & Ansari, M. MiR-153 as a tumor suppressor in glioblastoma multiforme is downregulated by DNA methylation. *Clin. Lab.* **62**, 573–580 (2016).
57. Chen, W.-J. *et al.* MicroRNA-153 expression and prognosis in non-small cell lung cancer. *Int. J. Clin. Exp. Pathol.* **8**, 8671–8675 (2015).
58. Xia, W. *et al.* miR-153 inhibits epithelial-to-mesenchymal transition in hepatocellular carcinoma by targeting Snail. *Oncol. Rep.* **34**, 655–662 (2015).
59. Xu, Q. *et al.* Downregulation of miR-153 contributes to epithelial-mesenchymal transition and tumor metastasis in human epithelial cancer. *Carcinogenesis* **34**, 539–549 (2013).
60. Persson, H. *et al.* Identification of new microRNAs in paired normal and tumor breast tissue suggests a dual role for the ERBB2/Her2 gene. *Cancer Res.* **71**, 78–86 (2011).
61. Li, W.-G., Yuan, Y.-Z., Qiao, M.-M. & Zhang, Y.-P. High dose glargine alters the expression profiles of microRNAs in pancreatic cancer cells. *World J. Gastroenterol.* **18**, 2630–2639 (2012).
62. Ma, W., Ma, C. N., Li, X. D. & Zhang, Y. J. Examining the effect of gene reduction in miR-95 and enhanced radiosensitivity in non-small cell lung cancer. *Cancer Gene Ther.* **23**, 66–71 (2016).
63. Ye, J. *et al.* Up-regulation of miR-95-3p in hepatocellular carcinoma promotes tumorigenesis by targeting p21 expression. *Sci. Rep.* **6**, 34034 (2016).
64. Urrutia, R., Henley, J. R., Cook, T. & McNiven, M. A. The dynamins: redundant or distinct functions for an expanding family of related GTPases?. *Proc. Natl. Acad. Sci. USA* **94**, 377–384 (1997).
65. Henley, J. R., Cao, H. & McNiven, M. A. Participation of dynamin in the biogenesis of cytoplasmic vesicles. *FASEB J.* **13**(Suppl 2), S243-247 (1999).
66. Fürthauer, M. & González-Gaitán, M. Endocytosis, asymmetric cell division, stem cells and cancer: Unus pro omnibus, omnes pro uno. *Mol. Oncol.* **3**, 339–353 (2009).
67. Cendrowski, J., Mamińska, A. & Miaczynska, M. Endocytic regulation of cytokine receptor signaling. *Cytokine Growth Factor Rev.* **32**, 63–73 (2016).
68. Seifert, M., Garbe, M., Friedrich, B., Mittelbronn, M. & Klink, B. Comparative transcriptomics reveals similarities and differences between astrocytoma grades. *BMC Cancer* **15**, 952 (2015).
69. Lauber, C., Klink, B. & Seifert, M. Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups. *BMC Cancer* **18**(1), 399 (2018).
70. Dos Santos, C., Récher, C., Demur, C. & Payrastre, B. The PI3k/Akt/mTOR pathway: a new therapeutic target in the treatment of acute myeloid leukemia. *Bull. Cancer* **93**, 445–447 (2006).
71. Quintás-Cardama, A. *et al.* p53 pathway dysfunction is highly prevalent in acute myeloid leukemia independent of TP53 mutational status. *Leukemia* **31**, 1296–1305 (2017).
72. Leddy, J. P. *et al.* Erythrocyte membrane proteins reactive with IgG (warm-reacting) anti-red blood cell autoantibodies: II Antibodies coprecipitating band 3 and glycophorin A. *Blood* **84**, 650–656 (1994).
73. Matsson, H. *et al.* Truncating ribosomal protein S19 mutations and variable clinical expression in Diamond-Blackfan anemia. *Hum. Genet.* **105**, 496–500 (1999).
74. Fawaz, N. A. *et al.* dRTA and hemolytic anemia: first detailed description of SLC4A1 A858D mutation in homozygous state. *Eur. J. Haematol.* **88**, 350–355 (2012).
75. Hsu, L. & Morrison, M. A new variant of the anion transport protein in human erythrocytes. *Biochemistry* **24**, 3086–3090 (1985).
76. Pang, A. J. & Reithmeier, R. A. F. Interaction of anion exchanger 1 and glycophorin A in human erythroleukaemic K562 cells. *Biochem. J.* **421**, 345–356 (2009).
77. Teague, R. M. & Kline, J. Immune evasion in acute myeloid leukemia: current concepts and future directions. *J. Immunother. Cancer* **1**, (2013).
78. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**(7), e47 (2015).

79. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?. *J. Classification* **31**, 274–95 (2014).
80. Cramér, H. *Mathematical Methods of Statistics. Princeton Llandmarks in Mathematics and Physics* 19th edn. (Princeton University Press, Princeton, 1999).
81. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health* (Springer, New York, 2000).
82. Seifert, M. & Beyer, A. regNet: an R package for network-based propagation of gene expression alterations. *Bioinformatics* **34**, 308–11 (2018).
83. Gladitz, J., Klink, B. & Seifert, M. Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. *Acta Neuropathol. Commun.* **6**, 49 (2018).
84. Seifert, M. *et al.* Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse. *PLoS Comput. Biol.* **15**(11), e1007460 (2019).
85. Tibshirani, R. Shrinkage and Selection via the Lasso. *J. R Stat. Soc. B* **58**, 267–288 (1996).
86. Lockhart, R. *et al.* A significance test for the lasso. *Ann. Stat.* **42**, 413–468 (2014).
87. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–17 (2004).

Acknowledgements
This study would not have been possible without the comprehensive data sets made publicly available by the TCGA Research Network. This project was supported by the German Cancer Aid (SyTASC / 70111969). We thank the SyTASC members for valuable discussions. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the SLUB/TU Dresden to cover the article processing charge. We thank PD Dr. Klaus H. Metzeler (LMU Munich) for providing the ELN 2017 reclassification of our validation patients. Open access funding provided by Projekt DEAL.

Author contributions
M.S. designed the study. C.L. and M.S. conducted the computational analyses of the data and wrote the manuscript. M.S. performed all studies for the revisions of the manuscript and revised the manuscript four-times. A.D. and L.B. provided the gene expression and survival data of the independent Ulm validation cohort. N.C., A.T., M.A.R. and I.R. supported the discussion of the results. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69691-8.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.4 Publication:

Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis

Journal: Genome Biology

Received: 14 June 2016; **Accepted**: 6 September 2016; **Published**: 3 October 2016

Citation: Michael Seifert, Betty Friedrich and Andreas Beyer (2016): Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis, Genome Biology, 17:204.

Placement and summary of the publication

Copy number alterations (CNAs) of large genomic regions are frequent in many tumor types, but only few of them are assumed to be relevant for the cancerous phenotype. It had proven exceedingly difficult to ascertain rare mutations that might have strong effects in individual patients. At the time when we developed the computational framework that is underlying this study, most existing methods focused on the analysis of somatic single nucleotide variants (SNVs) and either considered mutation frequencies in a cancer population or the distribution of mutations along the gene body to predict cancer driver genes. Virtually none of the existing approaches for the identification of driver mutations was able to actually quantify the clinical risk associated with individual DNA copy number alterations (e.g. Vogelstein et al. (2013); Hofree et al. (2013); Davoli et al. (2013); Tamborero et al. (2013); Ding et al. (2015)). In addition, the vast majority of theses studies only contained little validation of findings on independent patient cohorts leaving the clinical relevance of large-scale predictions in doubt.

In this study, we demonstrated that a genome-wide transcriptional regulatory network learned from gene expression and gene copy number data of 768 human cancer cell lines can be used to quantify the impact of individual patient-specific gene CNAs on cancer-specific survival

signatures. This network was highly predictive for gene expression in 4,548 clinical samples originating from 13 different tissues. Focused analysis of tumors from six tissues revealed that in an individual patient a combination of up to 100 gene CNAs directly or indirectly affect the expression of clinically relevant survival signature genes. Importantly, rare patient-specific gene CNAs (less than 1% in a given cohort) often had stronger effects on signature genes than frequent gene CNAs. Such novel insights cannot be gained using basic single-gene tests or CNA-frequency driven approaches. Subsequent integration with genomic data suggested that frequency variation among high-impact genes was mainly driven by gene location rather than gene function. Survival analyses on independent tumor cohorts revealed tumor-type specific trends indicating that rare gene CNAs can be as important as frequent gene CNAs for the prediction of patient survival. Further, an in-depth comparison to a related network-based approach showed that the integration of indirectly acting gene CNAs significantly improved the survival analysis.

The developed computational framework contributes to a personalized quantification of cancer risk, along with determining individual key risk factors and their downstream targets. In addition, the key computational concepts for network inference and network propagation developed in this study formed the basis of my R package regNet (Seifert and Beyer (2018); see Section 4.5). This paved the way to realize the search for driver gene candidates within the region of the 1p/19q co-deletion of oligodendrogliomas (Gladitz et al. (2018)) and the search for potential drivers involved in the regulation of radioresistance of prostate cancer (Seifert et al. (2019)). Both studies are also part of this habilitation thesis (see Sections 4.6 and 4.7). The value of our novel computational approach was also highlighted by the selection for a late breaking research presentation at the joint conference on Intelligent Systems for Molecular Biology (ISMB) / European Conference on Computational Biology (ECCB) 2015 in Dublin given by Prof. Dr. Andreas Beyer. Moreover, I received a prize for the best poster out of more than 130 posters at the conference of Systems Biology of Mammalian Cells (SMBC) 2016 in Munich.

Author contribution

I developed the concept of the study together with Andreas Beyer. I implemented all algorithms and performed the computational analysis of the different cancer data sets. I performed the initial quality control of the publicly available microarray data sets of the cancer cell lines to exclude samples with strong hybridization artifacts. Betty Friedrich supported the quality control of the microarrays and helped to obtain the different gene annotations used in the study. I wrote and revised the manuscript together with Andreas Beyer.

Seifert *et al. Genome Biology* (2016) 17:204
DOI 10.1186/s13059-016-1058-1

Genome Biology

METHOD Open Access

Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis

Michael Seifert[1,2,4*], Betty Friedrich[3] and Andreas Beyer[4]

Abstract

It has proven exceedingly difficult to ascertain rare copy number alterations (CNAs) that may have strong effects in individual tumors. We show that a regulatory network inferred from gene expression and gene copy number data of 768 human cancer cell lines can be used to quantify the impact of patient-specific CNAs on survival signature genes. A focused analysis of tumors from six tissues reveals that rare patient-specific gene CNAs often have stronger effects on signature genes than frequent gene CNAs. Further comparison to a related network-based approach shows that the integration of indirectly acting gene CNAs significantly improves the survival analysis.

Keywords: Cancer genomics, Bioinformatics, Computational systems biology, Network biology, Network inference, Network propagation, Gene copy number mutations

Background

Tumor cells harbor combinations of mutations that together impair molecular pathways, which results in neoplastic transformation. Although only a relatively small fraction of all mutations in any given cancer cell contributes to tumorigenesis, it is emerging that many more genes than previously thought determine clinically relevant endpoints such as proliferation rates, metastatic potential, or drug resistance [1, 2]. Clearly, hundreds of genes have the potential to contribute to tumor phenotypes [3], but we are still far from being able to quantify individual cancer risks. The frequency at which specific genes are mutated in a certain cancer cohort is an indicator of clinical importance. Even though frequent mutations (i.e. mutations that are more frequent than expected by chance in a specific cohort) are more likely to have tumor-related effects, individual cancer risks are most likely not fully explained by frequent mutations alone [1, 2]. Rare mutations could act in combination with frequent mutations or they may, entirely independent from frequent mutations, establish a significant risk for the patient on their own. Quantifying the risks associated with rare mutations has been complicated by the following reasons: (1) by definition, only a few patients carry these mutations, which reduces the probability of observing them in clinical studies, (2) even if they are observed, it is often difficult to quantify cancer risks statistically by comparing carriers with non-carriers due to insufficient statistical power, (3) complex interactions with other mutations (epistasis) may hide effects when analyzing single mutations in isolation, and (4) rare mutations of individual genes may have weak effects, but the co-occurrence of a sufficient number of such mutations in the same cell could significantly increase cancer risks. For example, a set of oncogenes with small individual effects but residing on the same chromosomal arm may establish a significant selective advantage if this chromosomal arm is amplified [3]. Essentially, we do not know how important rare mutations are in comparison to frequently observed mutations, simply because we are lacking the means to quantify their effects. The specific pattern of small mutations (single nucleotide variations or SNVs and small indels) in candidate genes can be used to prioritize putative driver genes without using epidemiological information [2–5]. Also, it has been shown that molecular networks can be used to

*Correspondence: michael.seifert@tu-dresden.de
[1]Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Institute for Medical Informatics and Biometry, Fetscherstr. 74, 01307, Dresden, Germany
[2]National Center for Tumor Diseases (NCT), Dresden, Germany
Full list of author information is available at the end of the article

better stratify patient populations by considering frequent and rare mutations together [1].

Apart from SNVs, DNA copy number alterations (CNAs) and chromosomal instability are a hallmark of cancer [6–8]. Further, CNA-affected genes with altered expression levels are more likely to be involved in tumorigenesis than affected genes with unchanged expression levels [9]. This has been exploited in previous studies to identify driver genes [10]. However, since CNAs frequently alter the expression levels of directly affected genes [9], these methods typically make many false positive predictions and require a large number of samples for a reliable prediction of potential key drivers. Other model-based approaches for the integrative analysis of gene copy number and gene expression data have been developed utilizing genetic linkage analysis [11] or network-based approaches [12–14] to identify major regulators driving tumorigenesis. All these methods (and many others) have greatly contributed to the identification of potential CNA tumor driver mutations and a better understanding of tumorigenesis, but none of these methods allows us to quantify the impact of rare gene CNAs.

Hence, novel computational methods are required to explore the long tail of rare mutations in cancer. An important step in this direction was done by [1], which enables the stratification of tumors that rarely share the same mutational profile into clinically relevant subtypes. Recently, another study proposed a network-based method that enables the identification of rare mutations involved in the perturbation of pathways and protein complexes involved in tumorigenesis [15]. This study predicted significantly mutated sub-networks containing dozens of genes rarely affected by mutations across different cancer types. Importantly, a common feature of [1] and [15] is the use of specifically designed network propagation algorithms to identify rarely mutated, but potentially relevant genes. However, we are still lacking methods for directly quantifying the impact of rarely affected genes on clinical endpoints such as survival.

Here, we present an approach exploiting the additional information contained in gene expression data to quantify potential effects of rare CNAs on clinically relevant endpoints. Our framework rests on the notion that regulatory relationships between genes are fairly robust across tumors, whereas the specific mutational pattern of a given tumor is virtually private [1, 16]. Put differently, most CNAs increase or decrease the activity of genes, while potentially only a small fraction of them alter the regulatory relationships between genes. Hence, by using large compendia of expression and mutation data sets, we can establish regulatory relationships between genes in cancer cells and quantify the effects of CNAs on gene expression. Such a model can subsequently be used to analyze individual tumors with known mutational patterns to quantify the impact of specific CNAs on global expression. Further, by relating those expression changes to clinical endpoints, we are able to quantify the effects of single CNAs on the survival of an individual patient. Using this framework, we can quantify direct (cis) effects and indirect (trans) effects of CNAs, we can identify key regulators in CNA regions (driver genes) with a particularly strong impact on the expression of clinically relevant genes, we can compare the importance of rarely mutated genes with frequently mutated genes, and we can quantify the combined effects of all CNAs on survival risk for an individual patient. Our analysis shows that usually many gene CNAs together influence individual patient survival by together impacting on common molecular pathways. At the individual level, it turns out that rare gene CNAs (less than 1 % frequency in a given cancer cohort) can be as important as frequent gene CNAs and we are able to specifically pinpoint potential candidate genes that are the most risky rare and frequent gene CNAs in individual patients.

Results and discussion

Cancer cell transcriptional network

To predict the potential effects of gene CNAs in the specific environment of tumor cells, we computationally inferred a genome-wide transcriptional regulatory network from human cancer cell lines of 24 different tumor sites (Additional file 1: Figure S1) [17]. We termed this model the cancer cell transcriptional network (CCTN, Fig. 1). The input data for CCTN, consisting of genome-wide gene copy number and gene expression data, were strongly quality controlled for hybridization artifacts (e.g. Additional file 1: Figure S2): each microarray of the 991 cell lines was manually checked for potential artifacts and 768 cell lines were kept after this filtering step (Additional file 2: Table S1). To identify putative regulator genes for each of the considered 15,942 genes, we modeled the expression level of each gene (target gene) as a linear combination of the gene-specific copy number and the expression levels of all other potential regulator genes. Sparse regression based on lasso (least absolute shrinkage and selection operator) [18] was used to select those variables (target gene-specific copy number and expression levels of other regulator genes) that best predict the expression level of a specific target gene, while keeping the number of variables small. This approach has previously been shown to perform well in similar tasks [14, 19, 20]. We quantified the significance of the selected predictors of each target gene [21] and kept only edges with p values below 5×10^{-5} (unless stated otherwise). Further, we removed potentially spurious regulator genes in the chromosomal proximity of target genes that actually just reflect the copy number state of the target (see 'Methods' for details). This resulted in a sparse transcriptional regulatory network (CCTN) comprising 36,786

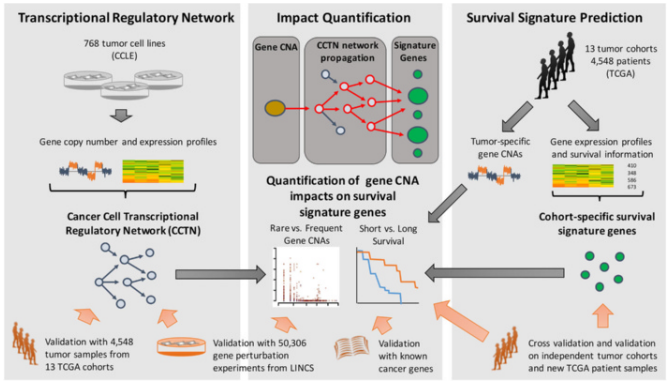

Fig. 1 Methodological overview. *Left* A cancer cell transcriptional regulatory network (CCTN) was inferred from gene expression and corresponding gene copy number data of 768 cancer cell lines of the Cancer Cell Line Encyclopedia (CCLE) and validated using data of thousands of tumor patients from The Cancer Genome Atlas (TCGA) and thousands of gene-specific perturbation experiments from the Library of Integrated Network-based Cellular Signatures (LINCS). *Right* Signature genes whose expression correlated with patient survival were determined for individual TCGA cohorts and validated on independent test data. *Center* CCTN was applied to gene copy number profiles of individual tumor patients of TCGA cohorts to predict the impacts of individual gene CNAs on cohort-specific survival signature genes and to separate short- from long-lived patients. The impact prediction was validated using LINCS data, known cancer genes, and data from two independent clinical cohorts and new TCGA patients. *CCLE* Cancer Cell Line Encyclopedia, *CNA* copy number alteration, *CCTN* cancer cell transcriptional regulatory network, *LINCS* Library of Integrated Network-Based Cellular Signatures, *TCGA* The Cancer Genome Atlas

directed trans-acting edges between regulator and target genes (Additional file 1: Figure S3; Additional file 3: Table S2). We refer to all genes affecting the expression of at least one other gene in CCTN as regulator genes (i.e. genes with at least one outgoing edge in CCTN). Note that this regulator definition is driven by the network inference approach that selects the most relevant predictors of each response gene. Not every regulator gene is necessarily a direct transcriptional regulator of a corresponding response gene. Genes affected by at least one regulator gene are regarded as target genes (at least one incoming edge in CCTN; see 'Methods' for details).

In total, 88 % of the genes (14,029 of 15,942) in CCTN were target genes, 60.6 % of the genes (9654 of 15,942) were selected as trans-acting regulators, and 27.3 % of the genes (4356 of 15,942) had a direct copy number effect that was always positively correlated with the underlying gene expression level (Additional file 3: Table S2). We further characterized the genes in CCTN based on their number of outgoing and incoming regulatory edges and found that the number of activator edges (32,521 of 36,786) is much greater than the number of repressor edges (4265 of 36,786) (Fig. 2a and b). In addition, CCTN is characterized by a few central hub genes that have a large number of incoming and outgoing edges. Well-known cancer genes [2, 22] (e.g. TNFRSF17, FUS, IKZF1, GATA1, PAX8, SFPQ, IRF4, KLK2, COL1A1, MSL2, HSP90AB1, PHOX2B, CD79B, and LYL1) were significantly overrepresented among the 219 hub genes with more than 20 trans-acting regulatory edges to or from other genes (Fisher's exact test: $p < 0.006$; Additional file 4: Table S3). Further, regulator genes with a large number of outgoing edges (i.e. major regulators) were enriched for known transcription factors and signaling pathway genes (Fig. 2c and d).

CCTN was derived from cancer cell lines, i.e. in vitro data. To test the validity of CCTN for in vivo tumor cells, we used independent data of 13 different cancer cohorts from The Cancer Genome Atlas (TCGA) [23]. We downloaded gene expression and corresponding gene copy number data of 4548 tumor patients (Additional file 5: Table S4) and tested the predictive power of CCTN on each TCGA cohort separately by predicting the expression level of each gene for each tumor using its corresponding copy number and expression data. To quantify the quality of the prediction, we computed the correlation between the originally measured TCGA gene expression levels and the corresponding expression levels predicted by CCTN for each gene across all patients in a cohort. A strong positive correlation between originally measured and CCTN-predicted expression levels suggests that the respective gene is well predictable by CCTN. The vast majority of genes had a positive median correlation (median across the 13 TCGA cohorts) between the predicted and measured expression levels (Fig. 2e): 94.7 % when using CCTN with all

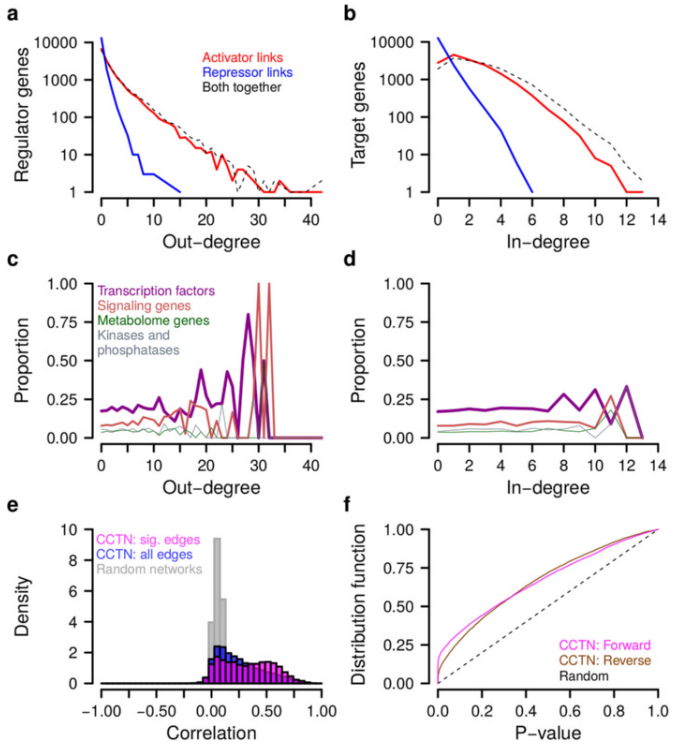

Fig. 2 Cancer cell transcriptional network (CCTN) characteristics and validation. **a**, **b** Node degree distributions. **c**, **d** Functional annotation of network genes with respect to their node degrees. **a**, **c** Regulator genes. **b**, **d** Target genes. **e** Median gene-specific correlations between predicted and originally measured gene expression levels of individual genes in 13 TCGA cancer cohorts for CCTN including only significant edges (*pink*), CCTN using all edges (*blue*), and for random networks with the same complexity as CCTN with significant edges (*gray*). CCTN with significant edges predicts gene expression levels significantly better than CCTN with all edges ($p < 6 \times 10^{-169}$) and random networks ($p < 2.2 \times 10^{-308}$, Wilcoxon test). CCTN with significant edges was used for all subsequent analyses. **f** Cumulative *p* value distributions correlating experimentally measured and computationally predicted single-gene perturbations pooling results from all 13 TCGA cancer cohorts. *Forward*: *p* values of correlations between computed impacts flowing from a perturbed regulator to its targets and the corresponding experimentally measured impacts. The forward model specifies the basic CCTN properties that were used to make impact predictions (one-sided correlation test quantifying for each single-gene perturbation if the observed correlation between predicted and measured impacts is significantly greater than zero). *Reverse*: *p* values of correlations between computed impacts flowing in the reverse direction from the responding targets to their perturbed regulator and experimentally measured forward impacts. *Random*: Baseline for non-significant enrichment of small *p* values. See 'Results and discussion' and 'Methods' for details of the forward and backward models. The forward model predicted responses of single-gene perturbations significantly better than the reverse model ($p < 0.015$ for each cohort) and than randomly expected ($p < 2.1 \times 10^{-23}$ for each cohort, one-sided Kolmogorov–Smirnov test). *CCTN* cancer cell transcriptional regulatory network, *sig.* significant, *TCGA* The Cancer Genome Atlas

edges and 95.1 % when reducing CCTN to significant edges. Restricting CCTN to significant edges had an even more dramatic effect on the magnitude of the correlation between predicted and observed expression (Fig. 2e; Wilcoxon–Mann–Whitney test: $p < 6 \times 10^{-169}$, Fig. 3). An additional comparison of CCTN to random networks with the same complexity showed that CCTN makes significantly better predictions of expression levels for the vast majority of genes (Fig. 2e; Wilcoxon–Mann–Whitney test: $p < 2.2 \times 10^{-308}$). We further confirmed that both target gene-specific direct copy number effects and trans-acting regulator genes contributed to the correct prediction of expression levels (Additional file 1: Figure S4). Although the predictive power of CCTN was variable

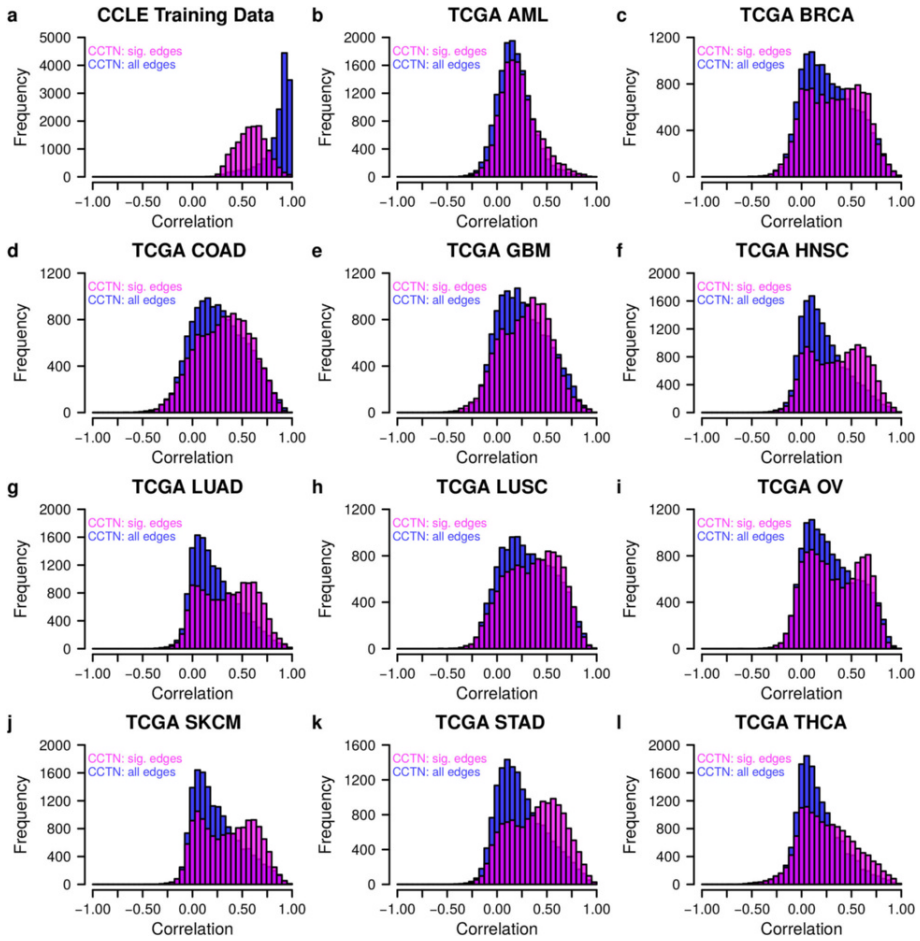

Fig. 3 CCTN-based prediction of gene expression levels for cancer cell lines and tumor patients. Gene-specific correlations between predicted and originally measured gene expression levels of individual genes comparing CCTN including only significant edges (*pink*) to CCTN using all edges (*blue*). A greater proportion of positive correlations reflects a better predictive power. **a** Prediction quality for human cancer cell lines used to train CCTN. As expected, CCTN using all learned edges is better than CCTN with significant edges only. **b–l** Prediction quality of CCTN for tumor patients of 11 independent TCGA cohorts. CCTN including only significant edges reaches strongly improved predictions for the vast majority of cohorts in comparison to CCTN with all learned edges. See Additional file 1: Figure S5 for all cohorts. *AML* acute myeloid leukemia, *BRCA* breast invasive carcinoma, *CCLE* Cancer Cell Line Encyclopedia, *CCTN* cancer cell transcriptional regulatory network, *GBM* glioblastoma multiforme, *HNSC* head and neck squamous cell carcinoma, *LUAD* lung adenocarcinoma, *LUSC* lung squamous cell carcinoma, *OV* ovarian serous cystadenocarcinoma, *sig.* significant, *SKCM* skin cutaneous melanoma, *TCGA* The Cancer Genome Atlas, *COAD* Colon adenocarcinoma, *STAD* Stomach adenocarcinoma, *THCA* Thyroid carcinoma

between individual genes and between tumor types, our model resulted in significant predictions for all considered patient cohorts (Fig. 3; Additional file 1: Figure S5) and was also very robust with respect to different p value cutoffs for including significant edges (Additional file 1: Figure S6).

We additionally compared CCTN, which was derived from in vitro cancer cell line data, to two network models

derived from in vivo data of specific tumor types. These tumor type-specific network models tended to reach a slightly or moderately improved predictive power compared to CCTN on independent test data cohorts of the same tumor type (Additional file 1: Figure S7a and b). This is expected, because CCTN was trained on a mixture of cancer cell lines and is, therefore, not specific for a certain tumor type. However, CCTN reached nearly identical or slightly improved predictive power in comparison to non-tumor type-specific network models (Additional file 1: Figure S7c and d). This again suggests that CCTN can be generalized to different tumor entities.

In conclusion, CCTN works well on independent data and correctly captures the majority of potential regulatory relationships between genes in the in vivo tumor situation.

Quantifying CNA impact on gene expression

Next we devised a method to quantify the impact of individual regulator genes on all other genes in the network (Fig. 1). This framework creates an impact matrix quantifying for each gene pair (a, b) the direct and indirect effect of gene a on the expression of gene b according to all existing directed regulatory network paths that link a to b in CCTN. The scoring also accounts for how well CCTN can predict the effects of mutations, i.e. CNA–target gene relationships that are poorly predicted get lower weights. Here, we operationally define the impact of a copy number change of gene a as its contribution to expression changes of gene b. That is, the impact is the (predicted) fraction of variance in the expression of a target gene caused by a specific gene CNA (see 'Methods' for details). The resulting impact matrix also accounts for the possibility of feedback cycles in CCTN, which could amplify (or dampen) the CNA effects.

We validated the correct prediction of impacts using individual gene perturbation data (LINCS L1000; see 'Methods' for details) [24, 25]. In these experiments, 933 genes (representatives of the human transcriptome) overlapped with CCTN genes and were perturbed on average 54 times and the expression responses of all other representative genes were measured, resulting in a total of 50,306 perturbation experiments (Additional file 6: Table S5). Note that the perturbations were repressing (knock-down) and increasing (overexpression) the transcript levels, which functionally mimics the effects of CNAs. We determined the significance of positive correlations between predicted and observed impacts across all 13 TCGA cancer types (see 'Methods') and found a strong enrichment of small p values (Fig. 2f; Additional file 1: Figure S8), confirming that the impact score is predictive for direct and indirect effects (one-sided Kolmogorov–Smirnov test comparing the p value distribution under the forward model to a uniform distribution expected under a random model: p values across TCGA cohorts ranging from 1.2×10^{-45} for thyroid carcinoma to 2.1×10^{-23} for skin cutaneous melanoma). The perturbation-expression data also allowed us to validate the direction of predicted effects: the correlated expression of two genes does not reveal which of the two genes is affecting which or if they are together under the control of a third gene. Since for perturbation experiments the direction of effects is known, we can use these data to assess the correct prediction of directional effects. We, therefore, compared the quality of CCTN predictions using a forward model (i.e. a model with correctly pointing interactions) with those of a reverse model (i.e. a model with inverted interactions). If the directionality information in CCTN is not meaningful, we would expect both models to perform equally well on the LINCS L1000 data. However, we observed that the forward model performed significantly better than the reverse model (Fig. 2f; Additional file 1: Figure S8: forward model contains more small p values than the reverse model: p values across TCGA cohorts ranging from 0.0004 for stomach adenocarcinoma to 0.015 for skin cutaneous carcinoma), suggesting that CCTN mostly correctly predicts the direction of effects.

Identification of tumor type-specific survival signatures

The CCTN-derived impact matrix has the ability to predict how a CNA of a gene affects the expression of all other genes in the network. To quantify the clinical relevance of individual gene CNAs, we determined genes whose expression levels are predictive of patient survival (Fig. 1). We developed an approach based on a random forest (RF) [26] to determine genes whose expression levels were significantly correlated with patient survival in individual TCGA cohorts (see 'Methods' for details). We chose RF for this task, because RF is particularly robust against overfitting, can handle complex non-additive relationships between predictor variables, and is able to exploit the molecular heterogeneity within a tumor cohort, which is essential for robust survival prediction and the characterization of survival-associated genes. In addition, an in-depth model comparison has previously shown that RF is among the best methods for the prediction of patient survival from gene expression data [27].

We initially tested our RF approach on all cohorts with more than 20 patients with survival information (8 of 13 TCGA cohorts; Additional file 5: Table S4). Testing of the resulting models on held-out patient samples (cross-validation) revealed that at least 100 patients with survival information were required to reach modest or more significant survival predictions (Additional file 1: Figure S9), which is in good accordance with previous findings for selected TCGA cohorts [28]. Correlations between RF-predicted and real patient survival on held-out samples were in the range of 0.12 to 0.35 for six TCGA cohorts (Additional file 1: Figure S9) with corresponding modest

significance ($p < 0.1$) for acute myeloid leukemia (AML) and skin cutaneous melanoma (SKCM), and stronger significance ($p < 0.013$) for head and neck squamous cell carcinoma (HNSC), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD), and ovarian serous cystadenocarcinoma (OV). The RF approach was not predictive for breast invasive carcinoma (BRCA) and lung squamous cell carcinoma (LUSC) (Additional file 1: Figure S9), possibly due to limited numbers of tumor samples or inadequate follow-up time. In addition, we also compared our RF approach to random survival forest (RSF) [29]. RSF can handle right-censored data to gain additional information from patients that were alive. However, our RF approach consistently reached better predictions of patient survival on held-out patient samples than RSF with and without censoring except for slightly improved survival predictions for AML (Additional file 1: Figure S10). RSF was also not predictive for BRCA and LUSC (Additional file 1: Figure S10). We further validated the RF-based survival prediction on GBM data from an independent patient cohort that was not part of the TCGA initiative [30]. The prediction of survival was highly significant, indicating that our RF model can make robust, potentially clinically relevant predictions (Additional file 1: Figure S11, $r = 0.52$, $p < 0.0006$, 36 patients). Thus, we focused on our RF approach and only kept the six cohorts (AML, GBM, HNSC, LUAD, OV, and SKCM) in all subsequent analyses, but note that the performance on held-out patients indicates a potentially greater clinical utility for HNSC, GBM, LUAD, and OV than for AML and SKCM.

Next, we ranked all genes based on their importance for predicting patient survival and filtered for the most important genes (signature genes) in each cohort by considering gene-specific contributions to the average correlation between RF-predicted and real patient survival (see 'Methods' and Additional file 1: Figure S12). The number of selected signature genes for the six cohorts ranged from eight for AML to 199 for GBM for a correlation cutoff of greater than 0.1 (Additional file 7: Table S6; Additional file 1: Figure S12). As expected, a complex clinical endpoint such as survival cannot be predicted from a small number of genes. Accordingly, the correlation of individual gene expression levels with survival was weak, thus underlining the need to consider multiple marker genes in combination to obtain significant predictions of patient survival (Additional file 1: Figure S13, $p < 0.004$ for all cohorts except for a more modest significance for genes positively correlated with HNSC survival reaching $p < 0.043$).

We further analyzed the obtained survival signatures for known cancer genes [22]. We found, for example, that the tumor suppressor NF1, a marker for mesenchymal GBMs [31], and the oncogene DNMT3A, a DNA methyltransferase impacting on proliferation and cell survival under hypoxic conditions [32], were part of the GBM survival signature. Interestingly, the transcription factor HOXD13 [33], which has been recently associated with poor survival of breast cancer patients [34], was part of the LUAD survival signature. Further, the tumor suppressor CAMTA1, a transcription factor involved in the regulation of cell growth and differentiation of neuroblastomas [35], and the tumor suppressor NIPBL, a cohesin regulator involved in developmental regulation, growth delay, and DNA repair [36], were part of the OV survival signature. We finally note that signature genes are not necessarily directly affected by CNAs. They should rather be considered as targets of driver mutations.

Impact of individual gene CNAs on survival signature genes

Using the CCTN-derived impact measures as defined above, we next quantified the contribution of each gene's copy number state on the expression of survival signature genes in each individual tumor (Fig. 1). First, we performed an integrated validation of the entire impact computation pipeline observing a significant positive correlation between patient-specific cumulative impacts of all individual gene CNAs and patient survival using an independent GBM cohort [30] that was not used for learning of any of our models for network effect quantification and survival signature prediction (Additional file 1: Figure S14a, Spearman rank correlation test: rho = 0.33, $p = 0.024$, 36 patients, see 'Methods' for details). This significant correlation between our impact scores and survival was not necessarily expected, as it does not account for mutations other than CNAs. In addition, these patient-specific impact scores further enabled a significant classification into short and long survival groups using Kaplan–Meier curves (Additional file 1: Figure S14b, $p < 0.02$).

After validating our impact scoring, we focused on the TCGA cohorts. For each mutated gene, we averaged its corresponding impact scores across all signature genes, yielding a single impact score that quantifies the contribution of this specific gene CNA on the expression of all survival signature genes. We selected high-impact gene CNAs for each of the six TCGA cohorts and corrected for multiple testing by comparing the originally obtained gene CNA-specific impact scores against corresponding gene-specific impact scores obtained under ten random networks of the same complexity as CCTN (Additional file 1: Figure S15, $q < 0.006$ for all cohort-specific selected genes, see 'Methods' for details). We further confirmed that these genes were enriched for known cancer genes [22] (Fisher's exact tests: $p < 0.03$ except for AML and SKCM).

In addition, our impact scoring identified many genes with established roles in the respective tumor classes (Fig. 4). For example, TAL1 had the greatest impact score among all LUAD-associated genes and had previously been identified as a hub transcriptional regulator in LUAD

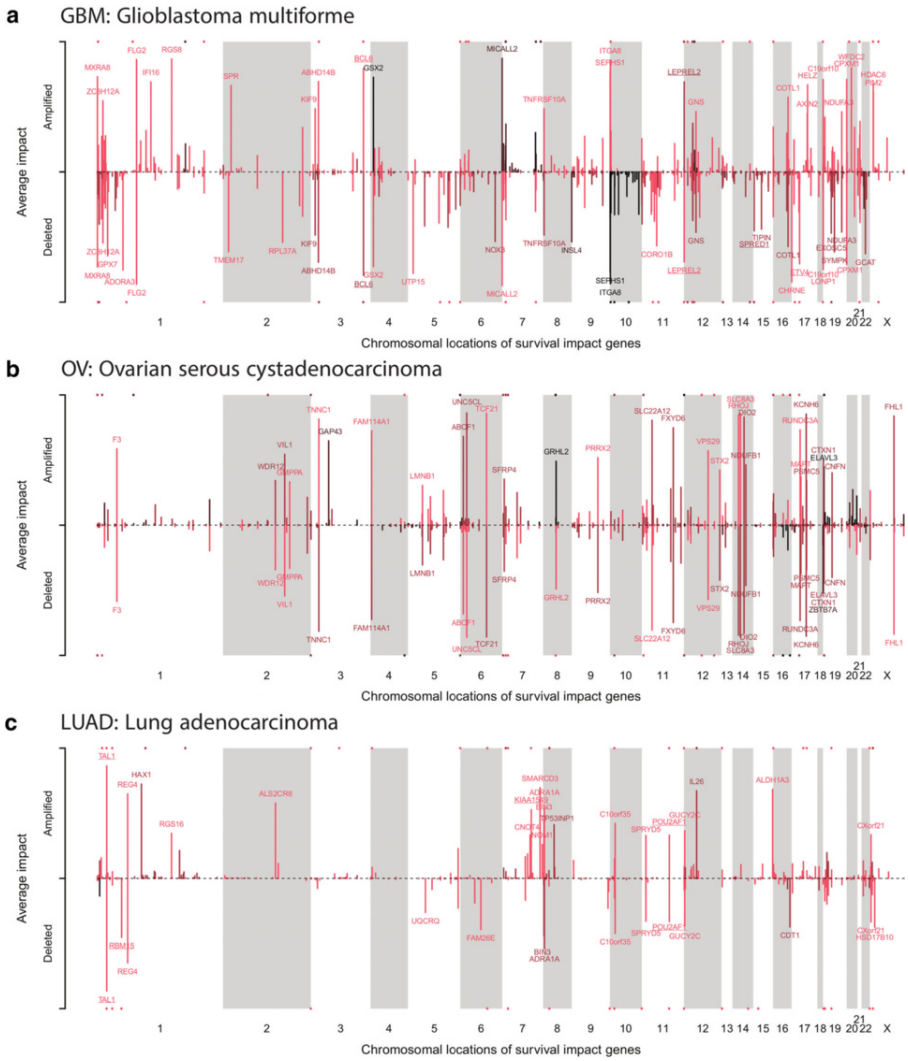

Fig. 4 Chromosomal locations of gene CNAs with high impact on survival signatures. **a** Glioblastoma multiforme (GBM). **b** Ovarian serous cystadenocarcinoma (OV). **c** Lung adenocarcinoma (LUAD). Genes are colored based on their observed mutation frequency (*light red*: rare, *black*: frequent). Heights of the peaks reflect the impact on survival signature genes. Selected gene names are displayed depending on their impact strength. Gene names of known cancer genes are underlined. *Colored dots* above or below genes additionally indicate the positions of previously reported cancer genes that were predicted as high-impact genes by CCTN. Only genes with significant impacts on survival signatures are shown ($q < 0.006$ for all cohort-specific genes, one-sided Wilcoxon test comparing the gene-specific impact score obtained under CCTN against corresponding impact scores obtained under random networks of the same complexity as CCTN, see 'Methods' and Additional file 1: Figure S15 for details). The cohort-specific genes shown are further significantly enriched for known cancer genes ($p < 0.03$ for each cohort, Fisher's exact test). *GBM* glioblastoma multiforme, *LUAD* lung adenocarcinoma, *OV* ovarian serous cystadenocarcinoma

with effects on TGF-beta signaling [37]. Another example is TNNC1, which is involved in the metastatic potential of ovarian cancer cells [38] and was among the top-ranking OV impact genes. Further, histone deacetylases (HDAC) have a well-established role in tumorigenesis and serve as important cancer drug targets. We correctly detected HDAC6 as a high-impact gene in GBM [39, 40]. The predicted high-impact gene CNAs impacting on AML, HNSC, and SKCM survival signatures are shown in Additional file 1: Figure S16. Apart from just confirming well-known tumor markers, our CCTN approach also provides supporting evidence for previously reported candidate genes and has the potential to reveal novel candidate genes impacting on survival. Both are demonstrated by the following examples.

Different gene CNAs putatively impact on the same survival signature gene

For example, HAX1 has been suggested to be involved in lung cancer [41]. We confirm that an increased HAX1 copy number contributes to an increased HAX1 expression level with downstream effects on the expression of TSEN15 (Additional file 3: Table S2). TSEN15, a LUAD survival signature gene, is involved in the tRNA splicing required for cell growth and division [42]. Our impact analysis further predicts TSEN15 as a downstream target of two other high-impact gene deletions of PLXNB2 and CHAC1 that both strongly impact on the expression of TSEN15. PLXNB2 is involved in cell proliferation and migration [43]. CHAC1 is a negative regulator of Notch signaling [44], involved in apoptosis [45] and known to function in other cancers [46, 47]. Thus, these three genes impact on a common molecular endpoint that is correlated with patient survival.

Duplication of chromosome 7 in GBM suggests further driver genes in addition to EGFR

It has previously been suggested that the clustering of driver genes on chromosomal arms may explain frequent amplifications or deletions of large chromosomal regions [3]. Our results support this notion and assist in the understanding of specific large deletions and amplifications. For example, the duplication of chromosome 7 is one of the most prominent chromosomal mutations found in GBMs [48] (Fig. 4a). Despite the frequency of this event, we have only an incomplete understanding about the genes in this region driving the cancerous phenotype. The amplification of the oncogene EGFR [49] on chromosome 7 is involved in GBM etiology. However, most likely additional genes on chromosome 7 contribute to GBM development and prognosis [50]. This is supported by our finding that patient-specific cumulative impact scores of all genes on chromosome 7 explain survival significantly better than the EGFR impact score alone (Meng's t-test on independent GBM cohort [30]: $p < 0.005$). We identified additional candidate genes on chromosome 7 with a high impact on GBM survival signature genes (Additional file 8: Table S7), including genes involved in (1) cell adhesion and migration, cytoskeletal organization, and neurite outgrowth (ARHGEF5, BAIAP2L1, MICALL2, SEMA3C, and TNS3), (2) transcriptional regulators and chromatin remodelers (ACTL6B, EZH2, H2AFV, IKZF1, and MLL3), and (3) cell proliferation, apoptosis, and DNA damage response (GIMAP6, HBP1, MCM7, PAXIP1, PPIA, SAMD9, and TBRG4). That EZH2 and MLL3 were found to be affected by small somatic mutations further supports their potential role in GBM etiology [48, 51].

Amplifications of tumor suppressors can contribute to longer survival

Interestingly, we also observed several high-impact genes that were amplified in some patients and deleted in others. The effect of an amplification or deletion may be conditional on other concurrent mutations, which is one possible explanation for this observation. However, we also detected some instances of positive gene CNAs where the respective CNA was associated with increased survival. For example, amplifications of the tumor suppressor genes WAC in GBM (97 tumors with amplifications vs 218 tumors with normal gene copy number, Fig. 4a, chromosome 10, p-arm) and CDH1 in OV (61 tumors with amplifications vs 174 tumors with normal gene copy number, Fig. 4b, chromosome 16, q-arm) were associated with significantly prolonged survival (t-test p values 0.0005 and 0.009, respectively).

Rare patient-specific gene CNAs strongly impact on survival signatures

After having established confidence in the impact scoring, we next determined the number of genes that have to be considered in combination to explain a certain fraction of survival risk in a given patient (Additional file 1: Figure S17). We revealed considerable variation between patients with respect to how many and to what extent gene CNAs affect survival signature genes. Up to 100 gene CNAs contributed together to the individually explained risk. Next, we focused on the relationship between impact and frequency at which gene CNAs occur in a tumor cohort (Fig. 5 for GBM, OV, and LUAD; Additional file 1: Figure S18 for AML, HNSC, and SKCM). As expected, more frequently mutated genes are more likely high-impact genes (Fig. 5a and b, correlation tests: $p < 0.03$) and accordingly, the median impact of frequently mutated genes tends to be higher than that of rarely mutated genes (Fig. 5c and d). Also, known tumor suppressor and oncogenes are enriched among more frequently mutated genes with high impact (Fig. 5e and f, correlation tests: $p < 0.005$ except LUAD deletions). However, even though frequently

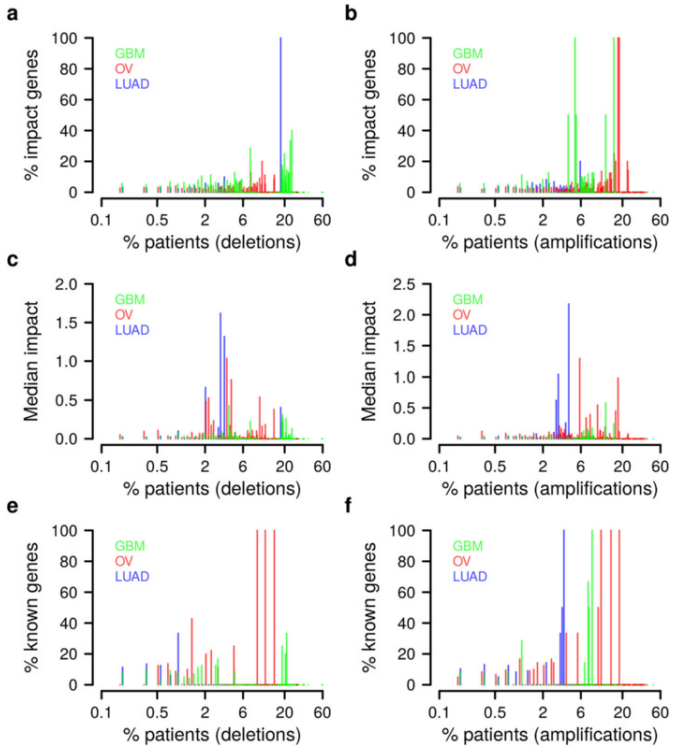

Fig. 5 Cohort-specific characterization of gene CNAs with a high impact on survival signature. All gene deletions and amplifications in each TCGA cohort (GBM, OV, and LUAD) with a high impact on the corresponding survival signature were grouped based on their mutation frequency shown in log-scale along the *x*-axes. **a**, **b** Percentage of genes in each bin belonging to the cohort-specific genes with a high impact on survival signature genes. More frequently mutated genes are more likely high-impact genes ($p < 0.03$ for each cohort, one-sided correlation test). **c**, **d** Median impact of high-impact genes in each frequency bin. The median impact quantifies the contribution of all high-impact gene CNAs in a bin to the variation of the expression levels of all cohort-specific survival signature genes. Average percentages of explained variance of survival signature expression computed for all high-impact gene CNAs were used to determine the median impact per bin. **e**, **f** Proportion of known cancer genes among the high-impact genes in each bin. Known tumor suppressor and oncogenes are enriched among more frequently mutated genes with high survival impact ($p < 0.005$ except for LUAD deletions, one-sided correlation test). *CNA* copy number alteration, *GBM* glioblastoma multiforme, *LUAD* lung adenocarcinoma, *OV* ovarian serous cystadenocarcinoma

mutated genes had on average larger impacts on signature genes, a substantial number of rarely mutated genes (frequency <1 %) also had strong impacts (Fig. 6 for GBM, OV, and LUAD; Additional file 1: Figure S19 for AML, HNSC, and SKCM). Importantly, some of these genes with CNAs in only one, two, or three individuals per cohort had impacts that were larger than those of many frequently mutated genes (Fig. 6a–f; Additional file 1: Figure S19a–f; Additional file 8: Table S7). In addition, a significant fraction of the low-frequency high-impact genes in GBM, OV, and LUAD have previously been reported as cancer genes [22] in other tissues (Fisher's exact tests: $p < 0.009$). In conclusion, the patient-specific expression pattern of survival signature genes can substantially be driven by individual rare gene CNAs, which is consistent with recent findings that patient-specific mutation patterns impact on survival [1].

Number of gene CNAs alone or single-gene tests do not allow to quantify survival impacts

The previous examples have shown that CCTN allows us to pinpoint rare and frequent gene CNAs that act on

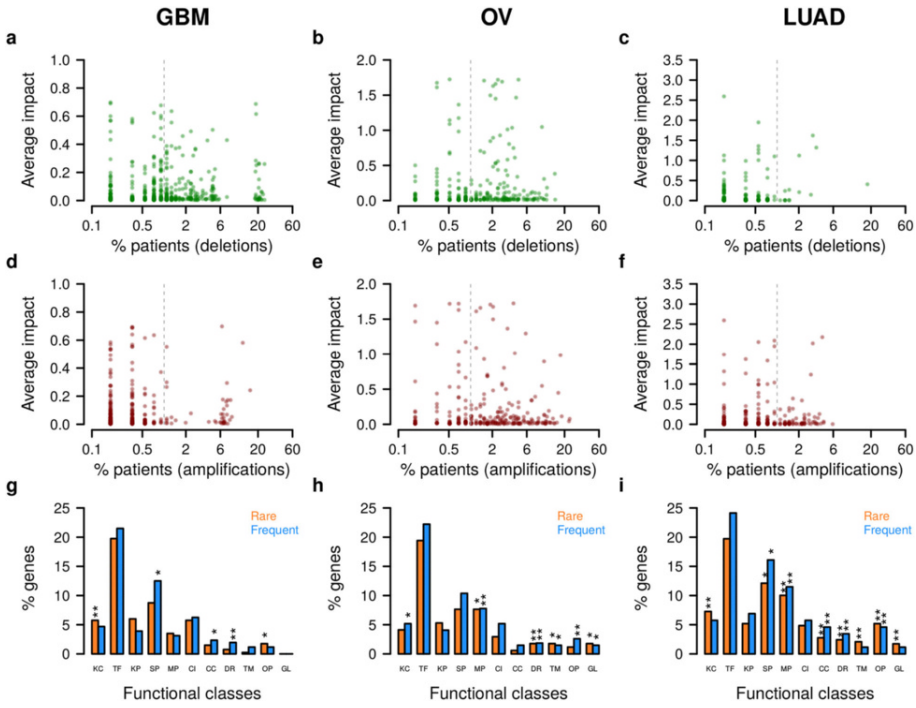

Fig. 6 Individual impact of rare and frequent gene CNAs with high impact on survival signatures. Characterization of TCGA cohorts GBM, OV, and LUAD. **a–f** Average impact of gene deletions (**a–c**) and amplifications (**d–f**) on the expression of cohort-specific survival signature genes. The average impact value of each specific gene CNA quantifies the average contribution of this gene CNA to the variation of the expression levels of all cohort-specific survival signature genes in percentage of explained variance. The *vertical gray dashed line* at 1 % of patients defines the cutoff used to separate rare from frequent CNAs. The *x*-axes are in logarithmic scale. **g–i** Corresponding joint functional annotation of deleted and amplified impact genes. Genes were classified as rarely (CNA frequency <1 %) or frequently (CNA frequency ≥1 %) mutated. The proportion of genes in selected functional classes are shown. Significant enrichment of an individual category is shown above bars as * ($p < 0.05$) and ** ($p < 0.01$, Fisher's exact test). Note that the height of the bars doses not necessarily correlate with significance due to the different sizes of the functional classes. *CC* cell cycle, *CI* cell–cell interaction, *CNA* copy number alteration, *DR* DNA replication, *GBM* glioblastoma multiforme, *GL* glycolysis, *KC* known cancer genes, *KP* kinases and phosphatases, *LUAD* lung adenocarcinoma, *MP* metabolic pathways, *OP* oxidative phosphorylation, *OV* ovarian serous cystadenocarcinoma, *SP* signaling pathways, *TCGA* The Cancer Genome Atlas, *TF* transcription factors, *TM* telomere maintenance

patient survival. We further analyzed if similar results can also be obtained using two alternative approaches. First, we considered the gene CNA burden of each patient, but we did not find any significant correlation between the number of CNA-affected genes (rare, frequent, or both together) and survival in any of the six TCGA cohorts (see Additional file 1: Text S2 for details). Second, we considered single-gene tests to determine if patients with a specific gene CNA had significant differences in survival compared to patients without this gene CNA. Considering the six TCGA cohorts, we were able to detect only some gene CNAs for AML that were significantly associated with survival, but as expected there were no rare gene CNAs among those genes (Additional file 1: Figure S20; see Additional file 1: Text S3 for details). Thus, our CCTN-based impact scoring approach allows us to gain novel insights into the putative impacts of specific gene CNAs.

Chromosomal location instead of gene function explains CNA frequency

Genes with very similar survival impact scores can have very distinct CNA frequencies in the same tumor class (Fig. 6a–f). We sought to identify factors explaining why some of those gene CNAs are observed much more rarely

than others. A first hypothesis was that rare high-impact mutations occur later in the tumor etiology and affect different endpoints than frequent gene CNAs. For example, frequent mutations might primarily drive the neoplastic transformation and thus affect proliferation, DNA damage response, and apoptosis, whereas rare CNAs might affect angiogenesis, metastatic potential, or drug resistance. However, functional classification of rare and frequent gene CNAs did not yield striking differences between the two CNA groups (Fig. 6g–i). Instead of function, the chromosomal location of genes seems to explain variable gene CNA frequencies better. The close placement of two tumor-relevant genes with antagonistic effects reduces the frequency of observing the respective CNAs [3]. For example, an oncogene and a tumor suppressor gene located in close chromosomal proximity reduce the chance that a CNA in that region will be beneficial for the tumor. We observed similar effects that distinguished rare from frequent gene CNAs in our data (Fig. 7; Additional file 1: Figure S19). For example in LUAD, frequent gene deletions are on average significantly further away from oncogenes [2, 3] and essential genes [52] than rare gene deletions (Fig. 7a and b, one-sided Wilcoxon tests: $p < 0.003$, average distance from oncogenes: 14.5 vs 6.3 Mbp, average distance from essential genes: 3.8 vs 2.9 Mbp), while gene amplifications are typically significantly further away from tumor suppressor genes [2, 3] (Fig. 7c, one-sided Wilcoxon test: $p < 0.002$, average distance from tumor suppressor genes: 7.2 vs 3.7 Mbp). Our data further show that the distance to fragile genomic sites [53] is correlated with the observed frequency of gene CNAs impacting on survival signatures. For example, in GBM, frequently amplified genes are significantly closer to fragile sites than rarely amplified genes (Fig. 7d, one-sided Wilcoxon test: $p < 5 \times 10^{-5}$, average distance from fragile sites: 4.7 vs 10.7 Mbp). Finally, the distance to frequently observed germ-line copy number variations (CNVs) [54] is correlated with the observed frequency of high-impact gene CNAs acting on survival signatures. For example, frequently amplified genes in GBM and OV are significantly closer to known germ-line CNV sites than rarely amplified genes (Fig. 7e and f, one-sided Wilcoxon tests: $p < 0.016$, average distance from tumor germ-line CNV sites for GBM is 0.98 vs 1.7 Mbp and 1.4 vs 1.7 Mbp for OV). Interestingly, these correlations between CNA frequency and genomic positioning were independent of survival impact, but highly specific for tumor type (Fig. 4; Additional file 1: Figure S21), suggesting that the molecular mechanisms leading to and maintaining CNAs are tissue-specific. Taken together, these analyses support that the chromosomal location of a gene rather than its function determines variable CNA frequencies among genes with similar impact.

Indirectly acting tumor-specific gene CNAs clearly improve survival prediction

Our CCTN-based impact quantification approach utilizes all patient-specific gene CNAs that directly or indirectly

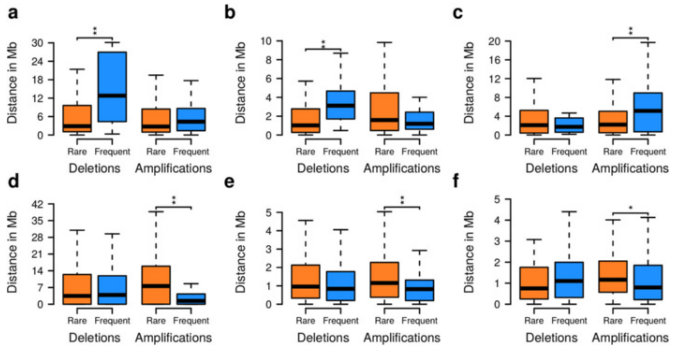

Fig. 7 Distances of rare and frequent high-survival-impact gene CNAs from genomic features. Selected examples for LUAD, GBM, and OV considering chromosomal distances of all gene CNAs with a high impact on cohort-specific survival signature genes from genomic features. See Additional file 1: Figure S21 for distance distributions of all tumor cohorts. **a** Distances of rare and frequent LUAD gene CNAs from known oncogenes. **b** Distances of rare and frequent LUAD gene CNAs from known essential genes. **c** Distances of rare and frequent LUAD gene CNAs from known tumor suppressor genes. **d** Distances of rare and frequent GBM gene CNAs from known fragile sites. **e** Distances of rare and frequent GBM gene CNAs from known frequently occurring germ-line CNVs. **f** Same as (**e**), but for OV. Significant differences in distances of rare and frequent CNAs from genomic features are represented by * ($p < 0.05$) and ** ($p < 0.01$, Wilcoxon test). *CNA* copy number alteration, *GBM* glioblastoma multiforme, *LUAD* lung adenocarcinoma, *OV* ovarian serous cystadenocarcinoma

act on patient survival to distinguish between short- and long-lived patients (Additional file 1: Figure S14). We further analyzed the value of integrating indirectly acting gene CNAs by comparing our approach to a basic version that only considers CNAs of genes in the direct network neighborhood of survival signature genes. Both impact scoring approaches utilize CCTN as the basis for enabling a fair comparison (see Additional file 1: Text S4 for details). To compare both approaches, we considered five independent test cohorts [Rembrandt: GBM [30]; Clinical Lung Cancer Genome Project (CLCGP): LUAD [55]; newly added TCGA patients: LUAD, SKCM, and HNSC; Additional file 5: Table S4]. First, we determined the numbers of patients that could be assigned to the short or long survival group based on their individual gene CNAs. We found that the integration of indirectly acting gene CNAs led to significantly increased numbers of classifiable patients for four out of five cohorts (Fig. 8a, $p < 2.2 \times 10^{-16}$, Fisher's exact test). This is explained by the observation that many patients did not have gene CNAs in the direct network neighborhood of survival signature genes, which prohibits a classification by the basic version. Second, we compared the separation quality between patients classified as short- and long-lived by both impact scoring approaches. In two out of five cohorts (Rembrandt and CLCGP), we did not find a significant difference in the separation between short- and long-lived patients (Fig. 8b). For all other cohorts (LUAD, SKCM, and HNSC), including indirect effects significantly improved the survival prediction compared to considering only direct effects (Fig. 8b). For example, this significant performance improvement is also observed when comparing the survival curves of short- and long-lived CLCGP and SKCM patients utilizing only frequent or all gene CNAs (Fig. 8c and d). Thus, the integration of indirectly acting gene CNAs into the prediction of short or long patient survival is an important factor to improve the classification of patients.

Frequent and rare tumor-specific gene CNAs contribute to survival prediction

We have already shown that individual frequent and rare tumor type-specific gene CNAs can have strong impacts on survival signature genes (Fig. 4). This motivated us

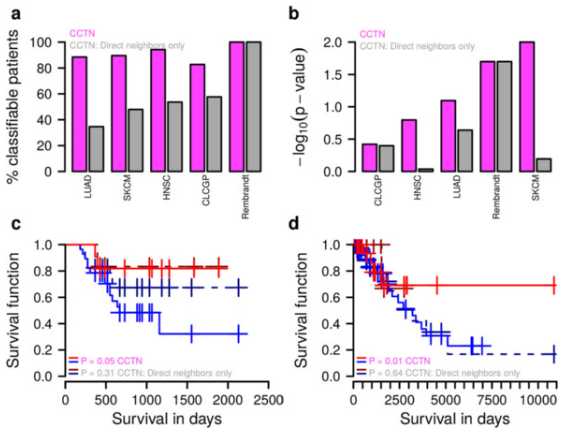

Fig. 8 Model comparison highlighting the importance of indirectly acting gene CNAs for survival analysis. Comparison of our standard survival impact quantification (CCTN) to a basic version (CCTN: direct neighbors only). The standard approach considers all gene CNAs that directly or indirectly impact on survival signature genes, whereas the basic version accounts only for directly acting gene CNAs to realize a classification into short- and long-lived patients. **a** Percentage of patients per cohort that can be classified by both approaches. CCTN reaches significantly increased numbers of classifiable patients compared to the basic version ($p < 2.2 \times 10^{-16}$, Fisher's exact test). **b** Separation quality of patients classified as short- or long-lived by both approaches. Corresponding *p* values quantify the separation between patients classified as short- or long-lived with respect to differences of the short- and long-lived Kaplan–Meier curve obtained for each approach under consideration of all tumor-specific gene CNAs. The *y*-axis is plotted in negative logarithmic scale. CCTN reaches a clearly improved patient separation for the majority of cohorts in comparison to the basic version. **c** Kaplan–Meier curves for separation of CLCGP patients into short and long survival based on frequent gene CNAs. CCTN reaches a significantly improved patient separation. **d** Like (**c**), but here for SKCM patients considering all patient-specific gene CNAs. *CNA* copy number alteration, *CCTN* cancer cell transcriptional regulatory network, *HNSC* head and neck squamous cell carcinoma, *LUAD* lung adenocarcinoma, *SKCM* skin cutaneous melanoma

to analyze further if tumor-specific gene CNAs of individual patients can be used to distinguish between short and long survival. A slight modification of our impact quantification algorithm enabled us to compute personalized impacts for each gene CNA in a patient-specific tumor (see 'Methods' and Additional file 1: Text S1 for details). This personalized impact score quantifies if the corresponding gene CNA has an inhibitory impact (negative impact value) or an activating impact (positive impact value) on a tumor type-specific survival signature gene. To account for the direction of the survival association of each signature gene, we multiplied this regulatory impact with the corresponding sign of the correlation observed between the expression levels of the signature gene and the survival of patients. This resulted in a personalized score that quantifies the impact of each tumor-specific gene CNA on survival. The score captures, for example, that a gene CNA with an inhibitory impact on a signature gene that is negatively correlated with survival has a potential positive impact on survival (may increase survival), whereas a gene CNA with an inhibitory impact on a positively correlated survival signature gene has a potential negative impact on survival (may decrease survival). To get an integrated survival score for all gene CNAs of a patient-specific tumor, we summarized the tumor-specific gene CNA scores to an average patient-specific survival impact score. Based on the score derivation, negative scores are expected to be associated with shorter patient survival than positive scores. We used these patient-specific average survival impact scores to analyze if the CCTN-based impact quantification allows us to distinguish between short and long survival coupled with a systematic analysis to quantify how frequent and rare gene CNAs contribute to the discrimination. In total, we utilized data of 292 tumor patients from five independent tumor cohorts including GBM patients from Rembrandt [30], LUAD patients from CLCGP [55] and newly added TCGA patients from LUAD, SKCM, and HNSC (Additional file 5: Table S4; no new GBM and AML patients were added to TCGA and only too few new OV patients were available from TCGA) that were not involved in any step of the CCTN inference nor in any step of the survival signature gene prediction before. To analyze these new patient samples, we used CCTN as derived from the cancer cell lines in combination with the corresponding tumor type-specific survival signature genes derived for the TCGA cohorts representing the same tumor entity to perform patient-specific impact quantification. An analysis of the contributions of (1) all patient-specific gene CNAs, (2) only patient-specific frequent gene CNAs, and (3) only rare patient-specific gene CNAs to the separation of long- and short-lived patients is shown in Fig. 9 for selected cohorts (Rembrandt: GBM; CLCGP: LUAD and SKCM; new TCGA patients). Results obtained for the other cohorts are shown in Additional file 1: Figure S22 (new LUAD and HNSC patients from TCGA). Importantly, this patient stratification was better than using random networks of the same complexity as CCTN, which led to a collapse of the impact quantification system (Fig. 9; Additional file 1: Figure S23), implying that our scoring is able to prioritize successfully gene CNAs with strong impacts on individual patient survival.

In more detail, for Rembrandt GBMs and new SKCM patients from TCGA, we observed a significant stratification into long- and short-lived patients (Fig. 9a and i, $p < 0.02$). Interestingly, rare gene CNAs (frequency <1 % in the training cohort) strongly contributed to the correct impact scoring for GBM and SKCM (Fig. 9c and k). In contrast to GBM and SKCM, the full scoring based on all patient-specific gene CNAs was not predictive for LUAD samples from CLCGP (Fig. 9e), whereas a scoring based only on frequent gene CNAs (frequency ≥ 1 % in the training cohort) was predictive for long and short survival (Fig. 9f, $p < 0.05$). Rare gene CNAs did not improve the LUAD patient stratification (Fig. 9g). These trends were also confirmed by an independent analysis of new LUAD patients from TCGA (Additional file 1: Figure S22a–c, $p < 0.01$ for frequent gene CNAs). Finally, we note that our impact scoring was not predictive for HNSC patients (Additional file 1: Figure S22i–l, Figure S23q–t), possibly due to the great molecular heterogeneity of HNSC tumors containing subtypes with only very few CNAs [56, 57].

In summary, frequent and rare gene CNAs are both important for the prediction of survival impacts. Overall there is no general trend that frequent gene CNAs tend to be more important than rare gene CNAs for the prediction of patient survival. The contributions of patient-specific rare and frequent gene CNAs tend to be rather tumor type-specific.

Conclusions

Multiple mutational patterns can perturb molecular pathways in similar ways leading to clinically almost indistinguishable phenotypes [1]. Thus, although the number of cellular endpoints that have to be altered is limited [6], the space of possible mutational patterns affecting the aggressiveness of a tumor (and ultimately patient survival) is practically unlimited. As a corollary of that, frequency-based approaches for detecting clinically relevant mutations will be capable only of detecting the mountains, leaving much of the phenotypic variation unexplained [2]. This study demonstrates the feasibility of an alternative strategy: the impact of gene CNAs on the expression of signature genes can be predicted using large compendia of independent data. Importantly, gene–gene relationships inferred from such data are largely conserved across multiple tumor types and enable statistically significant predictions of in vivo expression levels of most genes.

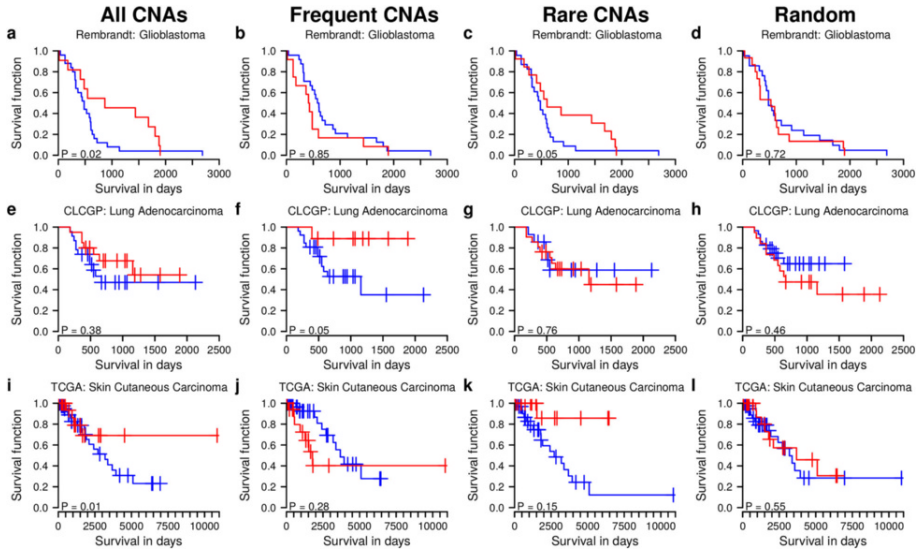

Fig. 9 Impact of rare and frequent patient-specific gene CNAs on survival. Kaplan–Meier curves for patients with negative (*blue*) and positive (*red*) average impact of their corresponding tumor-specific gene CNAs on cancer type-specific survival signature genes. Shown are results for independent tumor cohorts (Rembrandt: GBM; CLCGP: LUAD; TCGA: SKCM new patients) that were not involved in any step of CCTN network inference or RF-based prediction of survival signature genes. CCTN derived from cancer cell lines and cancer type-specific survival signature genes identified from TCGA cohorts were used to investigate the impact of rare and frequent gene CNAs on patient survival for these cohorts. Separation of long- and short-lived patients by CCTN is expected to be predictive if patients with positive average survival impact scores (*red*) tend to survive longer than patients with negative impact scores (*blue*). The corresponding *p* value quantifies if the *red curve* is significantly above the *blue curve* in comparison to random class label permutations. **a–d** GBM results. **a** Patients with positive average impact scores on survival signature genes (*red*) tend to survive significantly longer than patients with negative impact scores (*blue*). **b** Frequent patient-specific gene CNAs (frequency $\geq$1 % in corresponding training cohort) alone cannot explain the significant difference in (**a**). **c** Rare patient-specific gene CNAs (frequency <1 %) significantly contribute to the observed significant differences in (**a**). **d** Loss of patient separation into short and long survival for patient-specific survival impact scores computed based on all patient-specific gene CNAs under a random network. **e–h** LUAD results. Similar to GBM, but in contrast to GBM, frequent mutations alone contribute to a significant separation between long- and short-lived patients. **i–l** SKCM results. Overall trends are comparable with those observed for GBM. *CLCGP* Clinical Lung Cancer Genome Project, *CCTN* cancer cell transcriptional regulatory network, *CNA* copy number alteration, *GBM* glioblastoma multiforme, *LUAD* lung adenocarcinoma, *RF* random forest, *SKCM* skin cutaneous melanoma, *TCGA* The Cancer Genome Atlas

Thus, although the expression variation of individual regulators changes the activity of molecular sub-networks, the topology of regulatory relationships as such turns out to be remarkably robust across cell types [58]. Because of that, we were able to quantify the importance of gene CNAs for individual tumor risks leading to the observation that rare variants can be as important as frequent variants. Although this observation is not unexpected in light of recent research [1–3, 15], our framework allows us to specifically identify individual CNA-affected genes with a potentially high impact on survival. Importantly, the frequency at which a high-impact gene gets mutated seems to be determined by factors that are independent of its function or impact. Thus, the fact that some high-impact genes have higher CNA frequencies may simply be due to their placement in genomic regions that are more amenable for CNAs than others. Because of the higher CNA frequency in those regions, these genes will preferentially be selected during tumor evolution leading to increased average impacts of high-frequency CNAs. In short, impact does not affect frequency, but high frequency still correlates with high impact.

In addition, we noticed striking differences between tissues and between tumor types. For example, the correlation between CNA frequencies and genomic features was highly dependent on the tumor type. In addition, the importance of rare and frequent gene CNAs to distinguish between short and long patient survival was also

highly tumor type-specific. Further, we found many survival impact genes that are well-established cancer genes in one tissue to be also mutated (with a large predicted impact) in other tumors. However, the CNA frequency in those new tissues was mostly low, explaining why many of these genes have not been detected as being relevant in those tumors before. These observations imply that tissue-specific factors such as chromatin state, cell-cycle rates, exposure to DNA-damaging agents, number of stem cell divisions, or even the expression of specific genes could considerably impact on mutational mechanisms [59–61] that in the end affect patient survival.

Our conclusions rest on two computational models: the first, CCTN, describes transcriptional regulatory relationships between genes in a tumor context, i.e. in fast proliferating cells, but independent of a specific tumor type. The second model predicts signature genes associated with patient survival given cohort-specific expression and survival data. Three lines of evidence suggest that these models are robust and predictive. First, CCTN was predictive on a large set of in vitro perturbation-expression measurements. Second, CCTN was predictive on in vivo tumor data from all TCGA cohorts that we tested. Third, the impact scoring (which integrates both models) was predictive for survival in four out of five independent clinical cohorts that were not used for any of the previous analyses, revealing the tumor type-specific contributions of rare and frequent gene CNAs for the separation into long- and short-lived patients. However, despite our efforts to validate the models using a wide range of external data, this study is just a proof of principle. Obviously, improved models will have to account for a much wider range of mutation types, consider epigenetic effects, and include non-coding genes. Further, CCTN was learned from cancer cell lines to exclude variations in tumor cell purity between tumor samples that may have caused spurious dependencies between genes. Clearly, the usage of cancer cell line data has also disadvantages in comparison to tumor samples. Cell lines may not always correctly reflect the in vivo situation in tumors due to limitations set by cell cultures. We have designed the CCTN-based impact computation such that only those genes whose expression can adequately be predicted in the respective tumor entity contribute to the impact estimate. Thus, our framework makes no statement about genes whose regulatory networks differ significantly between cell lines and tumors. The list of genes that can be included in this analysis is further restricted by the use of different experimental platforms that did not cover identical gene sets. In addition, the quality of the predicted survival signatures varied greatly between the different tumor types, which is in agreement with previous reports on the limited usability of TCGA gene expression data for the prediction of patient survival [28]. This variability is in part due to the different sizes of patient cohorts or inadequate follow-up time. Further, the complexity of the mutational patterns and the relevance of CNAs in particular for the etiology of a tumor entity may further contribute to the differences in the predictive power of CCTN.

Our study provides clear indications that personalized analyses of patient-specific gene CNA profiles are feasible. The potential impacts of each patient-specific rare and frequent gene CNA on clinically relevant signature genes can be determined. So far we have only analyzed the impacts of rare and frequent gene CNAs on survival, but our framework is much more general, enabling, for example, other studies that may focus on impacts of rare and frequent gene CNAs on cancer-relevant signaling pathways or molecular signatures associated with treatment resistance. In addition, our framework also allows us to pinpoint potential high-impact genes in large chromosomal regions or on chromosomes that are recurrently affected by deletions or amplifications. We have demonstrated this potential for the recurrent duplication of chromosome 7 in glioblastomas, suggesting additional driver genes apart from the known role of EGFR. Further, comparative analyses of single-gene tests and a related network-based approach clearly demonstrated the value of our approach. Thus, our framework enables us to study the impacts of rare and frequent gene CNAs. Since copy number changes play a role in many other diseases or genetic disorders (e.g. trisomy 21), we anticipate applications of our framework beyond other interesting applications in cancer research.

Future work yet has to establish the value of accounting for rare gene CNAs to improve diagnostic and therapeutic measures. Fortunately, the availability of a regulatory model facilitates the detection of genes that are commonly affected by different rare gene CNAs, which might open a window of opportunity for developing therapeutic strategies against such rare mutations.

Methods

Cancer cell line data for CCTN inference

We initially considered all 991 human cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [17] and reconstructed hybridization images of corresponding gene expression and aCGH microarrays to systematically screen for and remove all cancer cell lines with hybridization artifacts. This resulted in a cancer cell line data set of 768 cell lines from 24 primary tumor sites (Additional file 2: Table S1). We normalized the gene expression experiments using GCRMA [62] in combination with a BrainArray design file (HGU133Plus2_Hs_ENTREZG 15.0.0). The resulting gene expression levels of each cancer cell line were further standardized by subtracting the corresponding average gene-specific expression level of all cell lines leading to log-ratios. We removed genes that did

not show any variation in gene expression across all cancer cell lines leading to 15,942 genes that were finally considered. The corresponding gene copy number data of all cancer cell lines were taken from CCLE. The copy number of a gene in a cell line was given by the log-ratio of the gene-specific copy number measured in the cell line in comparison to a normal reference.

CCTN inference

We divided the genome-wide transcriptional regulatory network inference problem into independent gene-specific sub-network inference tasks to obtain CCTN. For each target gene $i \in \{1,\ldots,N\}$, we assume that the expression level e_{id} of gene i in a cancer cell line $d \in \{1,\ldots,D\}$ can be predicted by a linear combination,

$$e_{id} = a_{ii} \cdot c_{id} + \sum_{j \neq i} a_{ji} \cdot e_{jd}, \quad (1)$$

of the gene-specific CNA c_{id} and the expression levels e_{jd} of other potential regulator genes $j \neq i$. The unknown parameters of this gene-specific linear model are specified by $\vec{a}_i := (a_{1i},\ldots,a_{Ni})^{\mathrm{T}} \in \mathbb{R}^N$. Here, a_{ii} quantifies the direct local gene copy number effect and a_{ji} with $j \neq i$ specifies the impact of the expression level of gene j on the expression level of gene i. The integration of gene-specific copy number data into the linear model extends gene expression-based correlation network inference approaches [19] and contributes to predicting the directionality of regulatory effects. We assume that a CNA of a regulator gene can lead to an altered expression of the regulator. This altered regulator expression can further lead to expression changes of target genes of the mutated regulator. Thus, each model parameter a_{ji} has a straightforward interpretation: (1) $a_{ji} < 0$ implies that the putative regulator j is associated with the repression of target i, (2) $a_{ji} > 0$ implies that the putative regulator j is associated with the activation of target i, and (3) $a_{ji} = 0$ implies that no putative regulatory edge between j and i exists.

All unknown parameters of the gene-specific linear model can be learned from the gene expression and gene copy number data of the 768 curated CCLE cancer cell lines. The use of cancer cell line data (which is free of normal cells) circumvented the variation in tumor cell purity between tumor samples that could lead to a spurious correlation between CNAs and expression levels of affected genes. Obviously, using cell line data also has disadvantages compared to data from tumors. For example, cell lines may incorrectly reflect the in vivo situation. However, our in-depth validation on TCGA tumor data suggests that the regulatory relationships are strongly conserved between cancer cell lines and patient-specific tumors (Fig. 2e and f; Additional file 1: Figures S4, S5). We utilized lasso regression [18] to compute a sparse solution for the linear model in Eq. (1). Lasso minimizes the residual sum of squares,

$$\vec{a}_i^{\,*} = \underset{\vec{a}_i}{\operatorname{argmin}} \sum_{d=1}^{D} \left(e_{id} - \left(a_{ii} \cdot c_{id} + \sum_{j \neq i} a_{ji} \cdot e_{jd} \right) \right)^2 + \lambda_i \sum_{j=1}^{N} |a_{ji}|, \quad (2)$$

of the measured expression e_{id} of gene i and the model-based predicted expression of gene i under consideration of all cancer cell lines in dependency of a fixed complexity parameter $\lambda_i \geq 0$. The complexity parameter λ_i determines the amount of shrinkage of the individual model parameters a_{ji} toward zero, where larger values of λ_i lead to greater shrinkage. This also enables us to select relevant predictors (gene-specific copy number impact and regulator genes) that best explain the expression of the response gene, because irrelevant model parameters can be shrunk to zero. The values of the fitted model parameters depend on the choice of the gene-specific complexity parameter. We utilized the R package glmnet [63] to determine an optimal gene-specific complexity parameter and corresponding optimal model parameters. We determined λ_i by averaging the optimal complexity parameters (cv.glmnet: lambda.min) obtained from ten independent repeats of a tenfold cross-validation across all cancer cell lines. We then used this gene-specific complexity parameter λ_i to compute the corresponding optimal model parameters $\vec{a}_i^{\,*}$ considering all cancer cell lines. We further determined the significance of model parameters when they first enter the lasso model in Eq. (2) using a recently developed significance test for lasso [21]. This provides an efficient way to get p values instead of using computationally expensive permutation strategies. To realize this, we first computed the lasso solution paths for the active predictors (model parameters in $\vec{a}_i^{\,*}$ that are unequal to zero) with respect to all cancer cell lines using the R package lars [64]. These results were then evaluated using the R package covTest [65] to obtain p values that characterize the importance of individual active predictors in the gene-specific linear model.

The p value distributions of active predictors and a quantile–quantile plot are shown in Additional file 1: Figure S24a and b for ten learned CCTN instances. We observed a strong enrichment of non-significant p values close to one and a smaller peak for highly significant p values with values close to zero. p values between these two extremes tended to be uniformly distributed. This highly left-skewed p value distribution (strong enrichment of non-significant p values) favors the parsimony of the model and is expected from the mathematical theory behind the significance test for lasso [21] (see Additional file 1: Text S5 and Figure S24c and d for details). Thus, as expected for lasso-based network inference, only very few predictor genes are required for the prediction of

the expression levels of specific response genes, whereas the majority of predictors shrink to zero. Note that the selected predictors remained significant after correction for multiple testing (Additional file 1: Figure S24e). Thus, regularization via lasso (reduction of the potential predictor test space) followed by additional filtering based on the significance of individually selected predictors represents an appropriate strategy to account for multiple testing.

We further removed all potentially selected local chromosomal regulator genes that were 50 genes upstream or downstream of each target gene to avoid the inclusion of genes that may simply reflect the copy number state of the target gene rather than regulatory dependencies. The choice of the local predictor cutoff is motivated by the observation that local chromosomal correlations of gene expression levels quickly approach zero with increasing distance between genes (Additional file 1: Figure S25a). Further, the structure of CCTN was hardly affected by varying local gene predictor cutoffs considering 20, 50, or 80 genes upstream or downstream of each response gene (Additional file 1: Figure S25b and c). Importantly, removing local chromosomal predictors did not affect the CCTN prediction accuracy, which was stable for the varying local predictor cutoffs (Additional file 1: Figure S25d–f). We just note that one could replace the fixed cutoff by a nucleotide distance cutoff to account for differences in local gene density, but as shown in Additional file 1: Figure S25, this will not have a strong influence on the results of our study.

We further tested if our network inference approach was affected by the multicollinearity of predictors by computing variance inflation factors (Additional file 1: Figure S26). Collinearity is present when two or more of the response gene-specific predictors have highly correlated measurements. The vast majority of variance inflation factors were close to one. Only 0.16 % of the predictor combinations had a variance inflation factor greater than ten, which is considered as an indicator of high multicollinearity [66]. Thus, CCTN is not affected by multicollinearity.

We repeated the learning of each gene-specific linear model ten times to evaluate the stability of our approach. We observed only very little variation of the gene-specific optimal complexity parameter, the gene-specific root mean square error, and the selected gene-specific predictors across the ten independent runs (Additional file 1: Figure S27a–c). We further selected for each target gene only those gene-specific predictors that had $p < 5 \times 10^{-5}$ (standard numerical precision limit of the R package covTest) in all ten runs (Additional file 1: Figure S25d and e). This corresponds to a q value cutoff of 0.0032. Note that also other cutoffs can be used, but we specifically focused on the resulting most parsimonious network, which reached substantially better predictive power than more complex network instances. In more detail, the prediction accuracy of the resulting ten instances of the gene-specific linear model was highly similar considering the CCLE data (Additional file 1: Figure S27f). The resulting reduced gene-specific linear models also showed significantly improved prediction accuracies for all independent TCGA patient cohorts compared to the initially obtained linear models, which also included non-significant predictors (Fig. 2e; Additional file 1: Figures S4–S6). Further, these predictions were also significantly better than the predictions of ten random networks of the same complexity as CCTN derived by degree-preserving permutations obtained by randomly exchanging predictors between the reduced gene-specific linear models while keeping the number of incoming and outgoing regulatory links constant for each gene (Fig. 2e). All subsequent analysis was based on average predictions done by an ensemble of ten CCTN instances focusing on significant predictors. The computation of a CCTN instance was computationally demanding and could not be realized on a standard desktop computer. It took on average 13.03 ± 3.06 min to learn the parameters of a gene-specific linear model from the 768 CCLE cancer cell lines (AMD Opteron 6274, 2.2 GHz, 2 GB RAM). Thus, it would take more than 140 days to obtain the whole network for the 15,942 genes in a sequential approach. We, therefore, solved the independent regression problems in parallel on a high-performance computing cluster (HPC Atlas Cluster TU Dresden, AMD Opteron 6274).

Generally, the network inference is very time-consuming because of the large number of potential gene-specific regulators and the large number of samples that should be considered to obtain robust networks. So far, we have removed only genes with constant expression levels among all cancer cell lines to reduce the number of potential predictors. This could be further extended by removing genes that show only little variation of expression levels for all cancer cell lines. Additionally, a preselection of potential gene-specific predictors via a correlation analysis could further help to reduce the predictor space to reduce the global computation time. However, such potential future preselection steps should be done carefully to avoid the loss of predictive power, because in our final network, about 61 % of all genes were selected as potential regulators of other genes.

Tumor data for validation, survival signature prediction and CNA impact studies

We downloaded gene expression and gene copy number data of 13 different tumor cohorts (4548 tumor patients in total) from TCGA [23]. Additional file 5: Table S4 contains all patient identifiers and dates of data freezes for the individual cohorts. We reorganized these data sets to obtain for each patient the corresponding gene expression

levels and gene copy numbers for the 15,942 genes considered in the CCLE cancer cell line data set. To obtain gene-specific copy number log-ratios for each tumor patient, we mapped the tumor-specific aCGH segments to the corresponding genes. If segment breaks occurred within a gene, we used the average log-ratio of the involved segments as a gene-specific copy number measurement. If a gene was not covered by at least one aCGH segment, we assumed that this gene was not affected by a copy number change and set its corresponding gene copy number measurement to zero. Note that personal normal aCGH controls were not available from TCGA. Instead, copy number signals were normalized against a universal reference. Thus, it is not possible to distinguish germ-line gene CNVs from somatic gene CNAs. This, however, does not affect our impact estimates, since our estimates do not require any enrichment of somatic mutations at driver genes. Microarray gene expression data were already reported by TCGA as log-fold changes against a universal reference. For RNA-seq data, we computed log-fold changes by normalizing to the average expression of the given gene in a cohort. Generally, genes that were measured in the CCLE data set used for CCTN inference, but which were not measured in some TCGA cohorts due to different experimental platforms, were always included with artificial measurements of zero, which did not provide any information for CCTN. This was done to enable a standardized application of CCTN to the different cohorts. We finally added corresponding patient survival information (status: dead or alive; survival time; and follow-up time) from TCGA.

We further downloaded gene expression, gene copy number, and survival data of five additional tumor cohorts (292 tumor patients in total) to validate the whole CCTN impact scoring pipeline based on tumor data that were not used in any analysis before. We considered independent GBM patients from the Rembrandt repository [30], curated and standardized in [67]. We downloaded processed data of independent LUAD patients form the CLCGP cohort [55]. We further downloaded newly added patients for the TCGA cohorts HNSC, LUAD, and SKCM and processed them as described above. Corresponding patient identifiers and dates of data freezes of all cohorts are provided in Additional file 5: Table S4.

CCTN-based impact computation

We developed a two-step approach to predict the impact of a specific gene CNA on the expression of a target gene of interest (here, signature genes) using CCTN, which represents regulatory relationships between genes learned from CCLE data. We now use CCTN to infer a cohort- or patient-specific impact matrix by propagating effects through CCTN using its regulatory paths between genes. Importantly, the resulting impact score is corrected for the variance that can be explained by CCTN at each node (gene) on the paths from the CNA gene to the target. An alternative naive approach would have been to correlate CNA states of genes directly with the expression of target genes of interest. Such an approach, however, has several disadvantages. First of all, such a model would be unable to predict the effects of CNAs that were not already contained in the training data, rendering it basically useless to investigate the effects of rare CNAs. Our approach can predict the effects of CNAs that were not seen in the specific patient cohort before. Second, the naive correlation model would lack mechanistic detail about how effects are propagated through the network, which is important for the interpretation of the results.

Basic network propagation algorithm

For all these reasons, we developed a network propagation algorithm that utilized CCTN to compute the information flow between genes in the network. This allowed us to compute the impact of patient-specific gene CNAs on survival signature genes. We considered a given TCGA cancer cohort of D patients for which gene expression and gene copy number profiles were measured for N genes. For each patient $d \in \{1, \ldots, D\}$, we took its gene expression and gene copy number profile to predict the expression level e_{id} of each gene $i \in \{1, \ldots, N\}$ using the corresponding gene-specific linear model in Eq. (1) with optimal parameters $\vec{a}_i^*$ from CCTN. Next, we computed each gene-specific correlation coefficient r_i between the predicted and the originally measured expression levels of gene i across all D patients of that cohort. Subsequently, we analyzed only genes with a positive correlation between predicted and observed expression levels ($r_i > 0$), and we termed those genes predictable. The fraction of predictable genes varied between tumors types (Additional file 1: Figure S5). Note that poorly predictable genes (i.e. genes with small positive r_i) will contribute only very little to the total impact score (see below). Thus, it is not necessary to further increase the minimal r_i for calling predictable genes. Next, we computed the corresponding variance $R_i^2 = r_i \cdot r_i$ explained for predictable genes that was covered by the underlying linear model in Eq. (1) and set $R_i^2 := 0$ for unpredictable genes ($r_i < 0$). Thus, R_i^2 directly reflects the network-based prediction accuracy for the expression level of gene i under CCTN by quantifying to what extent CCTN can explain the variance of gene i in a specific cancer cohort. Next, we considered each regulator gene j of gene i and determined for each regulator its direct contribution to the observed explained variance R_i^2 of gene i. Therefore, we computed the average proportion of each regulator j on the prediction of the expression of target gene i by

$$p_{ji} = \frac{1}{D}\sum_{d=1}^{D}\frac{|a_{ji}\cdot e_{jd}|}{|a_{ii}\cdot c_{id}| + \sum_{v\neq i}|a_{vi}\cdot e_{vd}|}$$

and determined the direct average copy number contribution of target gene i by

$$p_{ii} = \frac{1}{D}\sum_{d=1}^{D}\frac{|a_{ii}\cdot c_{id}|}{|a_{ii}\cdot c_{id}| + \sum_{v\neq i}|a_{vi}\cdot e_{vd}|}$$

under consideration of the D patients. We used absolute values in the computation of p_{ij} (and p_{ii}) to account for regulator genes that act as either inhibitors or activators of target gene i. If a gene j is not a direct regulator of gene i ($a_{ji} = 0$), then p_{ji} is set to zero. In analogy, if target gene i does not have a direct copy number effect ($a_{ii} = 0$), then p_{ii} is set to zero. Based on that, we defined a basic network flow matrix,

$$F = \left(f_{ji}\right)_{1\leq j,i\leq N} := p_{ji}\cdot R_i^2,$$

by weighting the explained variance R_i^2 of target gene i with the average proportion p_{ji} of its direct predictors (gene copy number and regulator genes) j. Thus, each column i of F contains the explained variance of a target gene i split into average proportions according to the contributions of its copy number and its target gene-specific regulators. The prediction of gene expression levels in tumors is of good quality, but of course not perfect (Additional file 1: Figure S5). For that reason, the explained variance fulfills $0 \leq R_i^2 < 1$ and, thus, the column sum norm of F is strictly less than one. We utilized this to compute the indirect effects of gene CNAs on other genes (i.e. the network flow) via:

$$F^* = \sum_{k=1}^{\infty} F^k,$$

which sums over the contributions of all network paths of increasing length k. Here, F^k specifies the kth matrix power obtained by a k-fold matrix multiplication of F. An element f_{ji}^k of F^k represents the impact of a trans-acting regulator gene j on the explained variance of a target gene i via all directed network paths from j to i of length k. Since the basic network flow matrix F has a column sum norm that is strictly less than one, the network flow F^* will converge to its limit $(I-F)^{-1} - I$ (geometric series of matrix F starting at one), where I is the identity matrix and $(I-F)^{-1}$ specifies the inverse of matrix $I - F$. However, the computation of the inverse of a large matrix is very time-consuming ($I - F$ has dimension $N \times N$). In addition, due to the sparsity of F (the majority of entries are zero because CCTN utilizes only the most relevant predictors) and its entries in $[0, 1)$, we also know that the values of the elements in F^k quickly approach zero. Thus, it is more efficient to approximate F^* by adding only an additional F^k if the obtained difference of the sum over F^k up to k and the previous sum up to $k-1$ is greater than a predefined threshold. We stopped the approximation of F^* if the sum of the differences of the column sums of the current and the previous approximated matrix was less than 10^{-3}. Starting with a TCGA cohort-specific sparse initial basic flow matrix F, we typically reached convergence after less than 50 iterations for most of the 13 different TCGA cohorts. The resulting network flow matrix F^* represents the impact values for each gene pair. All impact values in F^* are equal to or greater than zero. We further standardized each column of F^* by dividing each column-specific impact entry by the total sum of column-specific impacts followed by multiplication by 100 to get impact values in percentages. The impact of a gene j on the variation of expression of a gene i is given by f_{ji}^*. By considering the corresponding entries of F^*, we were able to quantify the impact of each patient-specific gene CNA on the predicted TCGA cohort-specific survival signature genes.

Identification of gene CNAs with a high impact on survival

We applied the basic network propagation algorithm to all TCGA cohorts for which we obtained survival signature genes that were significantly associated with patient survival (AML, GBM, HNSC, LUAD, OV, and SKCM). This resulted in a cohort-specific impact matrix F^* for each cohort. We next determined for each patient in a cohort all of their tumor-specific gene CNAs (genes with absolute aCGH log-ratio $\geq$0.75; Additional file 1: Figure S17: results obtained for a more stringent absolute aCGH log-ratio cutoff $\geq$1 were highly similar) and computed the frequency of all gene-specific deletions or duplications in the whole cohort. We then took each mutated gene and considered the cohort-specific impact matrix F^* to compute the average impact that a mutated gene had on all cohort-specific survival signature genes (Additional file 1: Figure S12: selection of a stringent set of signature genes using a correlation cutoff >0.1; Additional file 1: Figure S13; and Additional file 1: Figure S17: results obtained for a less stringent correlation cutoff >0.05 were highly similar). We next considered for each cohort all genes that had at least one deletion or duplication, sorted all these genes in increasing order of their impacts, computed the cumulative impact across all mutated genes, and plotted this cumulative impact clearly highlighting cohort-specific gene CNAs with a high impact on patient survival (Additional file 1: Figure S15). We next used a cumulative impact cutoff of greater than one to select high-impact gene CNAs for each cohort (Additional file 1: Figure S15: black dashed line close to zero). We further ensured that the impact of each selected high-impact gene on the survival signature genes was significantly greater than the corresponding gene-specific impacts obtained under ten random networks of the same complexity as

CCTN (degree-preserving network permutations). Therefore, we computed for each CNA gene in a cohort the difference between its CCTN-based impact score and each corresponding impact score under a random network leading to ten gene-specific impact score differences per gene. We then tested for each gene if the gene-specific differences between the original and the random impact scores were greater than zero using a one-sided Wilcoxon test. We further corrected the resulting p values for multiple testing by computing false-discovery-rate-adjusted p values (q values) for all genes [68]. We recognized that also very small impacts close to zero can be highly significant, because the observed impacts obtained under random networks were even closer to zero. However, such genes with very small impact are less likely to be biologically or clinically relevant. Therefore, we decided to focus only on stringent selections of cohort-specific high-impact genes based on Additional file 1: Figure S15 as described above instead of using a fixed q value cutoff. The q values of the selected high-impact genes were less than 0.006 for all TCGA cohorts (q value cutoffs: AML < 0.0053, GBM < 0.0048, HNSC < 0.0058, LUAD < 0.0056, OV < 0.0046, and SKCM < 0.0049).

Extension to patient-specific impact scores

Further, we note that the proportions p_{ji} and p_{ii} used to construct the basic network flow matrix F are cohort-specific averages using the basic network propagation algorithm described above. One can easily modify the computation to get specific proportions for each individual tumor patient (Additional file 1: Text S1: Patient-specific absolute impact scores) to construct a patient-specific basic network flow matrix F, but these computations and the later network propagation steps are even more time- and resource-consuming because one now has to apply the network propagation algorithm to each individual patient. This takes about 24 hours on an AMD Opteron 6274 with 2.2 GHz, requiring up to and more than 80 GB RAM for one patient. A compressed basic network flow matrix F required about 1 MB of disk storage for one patient, but the resulting compressed final impact matrix F^* required about 1 GB of hard disk space. To compare both approaches, we randomly selected 100 patients from each of the six TCGA cohorts (AML, GBM, HNSC, LUAD, OV, and SKCM) and found that the obtained patient-specific impact values acting on survival signature genes or all network genes are strongly correlated with the corresponding cohort-specific impact values (Additional file 1: Figure S28). For that reason, we decided to work with cohort-specific impact scores in Figs. 4, 5, 6 and 7 and the corresponding Additional file 1: Figures S15–S20. We did notice, however, that in some cases the patient-specific impact matrix significantly deviated from the cohort average (Additional file 1: Figure S28), suggesting that in the future, it might even be worthwhile to use personalized impact matrices.

In addition to this absolute quantification of impacts of patient-specific gene CNAs, one can further slightly modify the computation of the specific proportions p_{ji} and p_{ii} to obtain relative proportions that enable us to propagate patient-specific repressive and activating impacts through the network (Additional file 1: Text S1: Patient-specific relative impact scores). We used these scores to compute patient-specific survival impact scores considering all corresponding tumor-specific gene CNAs as described in 'Results and discussion.' These patient-specific survival impact scores enabled us to distinguish between long- and short-lived patients and to investigate the contributions of all, frequent, or rare tumor-specific gene CNAs on patient survival (Fig. 9; Additional file 1: Figures S14, S22, and S23). This approach was as time and resource intensive as described above.

Perturbation data for CCTN-based impact validation

We used the L1000 data set of the Library of Integrated Network-based Cellular Signatures (LINCS) [24] to validate our CCTN-derived impact scores. The L1000 data set provides information about gene expression changes of different human cell lines in response to chemical (small molecule) or genetic (shRNA) perturbations. We focused on perturbation experiments done for the about 1000 landmark genes defined by the LINCS consortium as representatives of the human transcriptome. We found that 933 of these landmark genes were part of CCTN. We next considered all gold standard perturbation experiments performed for these 933 genes and downloaded for each perturbation experiment the corresponding accessible top 100 response genes (top 50 up- and top 50 down-regulated landmark genes) via the application programming interface accessible under http://api.lincscloud.org/. Overall, we obtained the top 100 response genes of 50,306 perturbation experiments leading to on average 54 perturbation experiments for each of the 933 genes (Additional file 6: Table S5). We used this information to create a response gene frequency statistic for each perturbed gene by taking into account all corresponding gene-specific perturbation experiments, i.e. we counted how frequently each of the 933 landmark genes was observed among the top 100 response genes. Next, we compared the ranks of the corresponding impact scores from the CCTN-derived impact matrix with these independently obtained response scores. CCTN-derived impact scores and LINCS-derived response scores were correlated gene-wise. The distribution of p values resulting from a pan-cancer analysis of the individual impact matrices obtained for the 13 different TCGA cohorts was significantly shifted towards small values [Fig. 2f, one-sided Kolmogorov–Smirnov test comparing the p value

distribution of the forward model (see below) to a uniform distribution representing the baseline for non-significant enrichment: $p < 2.1 \times 10^{-23}$ for each TCGA cohort], confirming the overall significant predictive power of our impact scores. Importantly, such a significant shift towards small p values was also observed for each individual impact matrix of a TCGA cohort (Additional file 1: Figure S8).

In addition, for the perturbation experiments, the directionality of effects is known. Thus, we utilized the LINCS data to validate the correct prediction of the directionality of effects by CCTN. Therefore, we compared the standard forward model, which quantifies the significance of correlations between computed impacts flowing from a perturbed regulator to its targets and the corresponding experimentally measured impacts, to the reverse model, which quantifies the significance of correlations between computed impacts flowing in the reverse direction from the responding targets to their perturbed regulator and experimentally measured forward impacts. That means that in the forward model, both compared impacts flow in the same direction, whereas in the reverse model, the compared impacts flow in opposite directions. If CCTN contained only information about pairwise correlations of gene expression levels, we would expect that the forward and the reverse models would perform equally well on the LINCS data. We found that the forward model reached a stronger enrichment of small p values than the reverse model for a pan-cancer analysis of the individual impact matrices obtained for the 13 different TCGA cohorts (Fig. 2f, one-sided Kolmogorov–Smirnov test comparing the p value distribution of the forward model to the p value distribution of the reverse model: $p < 0.015$ for each TCGA cohort). This was also found for the impact matrix of each individual TCGA cohort (Additional file 1, Figure S8a–m) and further supported by direct gene-specific comparisons of the forward and backward models (Additional file 1: Figure S8o). This suggests that CCTN is mostly able to correctly predict the directionality of effects.

Identification of survival signature genes

We used random forest (RF) [26] to identify genes that were associated with the survival of patients in TCGA cohorts. RF was previously found to be one of the best performing methods for the prediction of patient survival based on gene expression data [27]. All analyses were performed on uncensored data using the R package randomForest [69] with standard settings. We initially applied RF to patient-specific gene expression profiles of each TCGA cohort containing more than 20 patients with survival information (Additional file 5: Table S4: AML, BRCA, GBM, HNSC, LUAD, LUSC, OV, and SKCM) to evaluate how many patients are required for significant predictions of patient survival. Validations of each cohort-specific RF on corresponding out-of-the-bag data (patient-specific gene expression profiles that a specific tree of the RF has not seen during its construction) showed that for six TCGA cohorts with more than 100 patients (AML, GBM, HNSC, LUAD, OV, and SKCM), significant predictions of patient survival were possible (one-sided correlation tests: $p < 0.1$; Additional file 1: Figure S9).

Next, we focused on these six cohorts and developed an RF-based approach to determine genes that are associated with patient survival. For each of the selected TCGA cohorts, we standardized the expression levels of each gene to a mean of zero and a standard deviation of one across all patients. We next randomly selected 90 % of the patients for the training of an RF and utilized the remaining patients as independent test sets for the evaluation of the performance of survival prediction and the characterization of relevant genes. We trained an RF on the training set and determined the corresponding gene-specific selection frequencies (SFs) that quantify how frequently each gene was chosen as a relevant survival predictor. We repeated the separation into training and test data 100 times and trained the corresponding RFs to evaluate the stability of the obtained SFs. We found that the standard deviations of the SFs were close to zero also for genes with SFs clearly greater than zero. Thus, the RF-based association of genes with patient survival was robust. Next, we computed the average SF for each gene based on the 100 RFs and corrected them for selection biases. This was done by subtracting average gene-specific SFs obtained from 100 corresponding RFs that were trained using randomly permuted survival information. To obtain a ranking of genes with respect to their strength of association with patient survival, we ranked all genes in decreasing order of their average corrected SFs. This allowed us to quantify their importance for the prediction of patient survival utilizing the independent test data set that we had initially put aside. Therefore, we considered each of the 100 RFs and its corresponding test data set and predicted the survival of the test patients with respect to successively increasing numbers of permuted expression levels (permutation of gene-specific expression levels across all test patients) for the previously determined top-ranking predictor genes. For each of these successive permutation steps, we computed the correlation between the originally observed test patient survival and the RF-predicted test patient survival to quantify the importance of the top-ranking genes associated with survival. We repeated this procedure ten times for each of the 100 RFs leading to 1000 permutation runs in total. We did this in steps of single genes for the first 1000 top-ranking predictors followed by steps of 100 genes for the remaining top-ranking predictors. We finally averaged the obtained correlation profiles for successively permuted top-ranking predictors

across the 1000 permutations. We found that the average correlation profile of the top-ranking predictors quickly approached zero, enabling us to set a cutoff to select the most relevant genes associated with survival (Additional file 1: Figure S12). We subsequently considered all predictor genes above a stringent correlation cutoff of 0.1 (also later used in our in-depth studies) and a less stringent cutoff of 0.05 as TCGA cohort-specific survival signature genes and confirmed that the expression of these genes was correlated with patient survival.

Therefore, we used standard hierarchical clustering to group the top-ranking predictor genes revealing two major groups: (1) survival signature genes negatively associated with survival and (2) survival signature genes positively associated with survival. We finally computed average patient-specific gene expression levels for these two clusters and confirmed that these average expression profiles are significantly correlated with patient survival (one-sided correlation tests: $p < 0.05$; Additional file 1: Figure S13), suggesting that our RF approach is well suited for the identification of survival signature genes.

In addition, we also compared our RF approach to random survival forest (RSF) [29], which can handle right-censored data to gain additional information for the prediction of patient survival. We used the corresponding R package randomForestSRC to determine RSFs. We found that our RF approach reached clearly better predictions of patient survival than RSF without and with censoring for the initially considered TCGA cohorts (Additional file 1: Figure S10). See 'Results and discussion' for more details.

Gene annotations and genomic features

Lists of human transcription factors and co-factors, phosphatases, kinases, signaling and metabolic pathway genes, essential genes, tumor suppressor and oncogenes, and known cancer genes were compiled from different public resources (see Additional file 9: Table S8 for genes and references). Fragile genomic sites [53] were extracted and lifted over to hg19 (Additional file 10: Table S9). Frequently observed CNV sites [54] were available for hg19 (Additional file 11: Table S10).

Additional files

Additional file 1: Texts S1–S5 and Figures S1–S28. (PDF 6430 kb)
Additional file 2: Table S1. Cancer cell lines used. (TXT 62 kb)
Additional file 3: Table S2. CCTN gene interaction table. (TXT 966 kb)
Additional file 4: Table S3. CCTN node degrees. (TXT 1177 kb)
Additional file 5: Table S4. Patient samples used. (XLS 278 kb)
Additional file 6: Table S5. LINCS perturbation experiments used. (TXT 1054 kb)
Additional file 7: Table S6. Predicted TCGA cohort-specific survival signature genes. (XLS 58 kb)
Additional file 8: Table S7. Gene CNA rates and survival impacts. (XLS 4259 kb)
Additional file 9: Table S8. Gene annotations table. (XLS 3420 kb)
Additional file 10: Table S9. Fragile sites used. (TXT 2 kb)
Additional file 11: Table S10. Germ-line CNV sites used. (TXT 17 kb)

Abbreviations

AML: Acute myeloid leukemia; BRCA: Breast invasive carcinoma; CCLE: Cancer cell line encyclopedia; CCTN: Cancer cell transcriptional network; CLCGP: Clinical lung cancer genome project; CNA: Copy number alteration; CNV: Copy number variation; COAD: Colon adenocarcinoma; GBM: Glioblastoma multiforme; HNSC: Head and neck squamous cell carcinoma; lasso: Least absolute shrinkage and selection operator; LINCS: Library of Integrated Network-based Cellular Signatures; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; RF: Random forest; RSF: Random survival forest; SKCM: Skin cutaneous melanoma; SNV: Single nucleotide variation; STAD: Stomach adenocarcinoma; TCGA: The cancer genome atlas; THCA: Thyroid carcinoma

Acknowledgments

We thank Michael Kuhn (EMBL Heidelberg) for the recommendation of the LINCS resource and Martin Peifer (University of Cologne) for critically reading the manuscript and for providing his list of frequent germ-line-affected CNV regions. We thank Martin Garbe for supporting the compilation of lists of transcription factors, kinases, and phosphatases. We are grateful to the Broad LINCS Center for providing their data prior to publication. Further, this work would have been impossible without the TU Dresden's Center for Information Services and High Performance Computing (ZIH) providing us with significant computational and storage resources. We thank the reviewers and the editor for their very valuable comments.

Funding

MS, BF, and AB were supported by GlioMath-Dresden funded by the European Social Fund and the Free State of Saxony. MS was further supported by the Center for Information Services and High Performance Computing (ZIH) TU Dresden, by the Cluster of Excellence in Cellular Stress Responses in Aging-associated Diseases (CECAD) University of Cologne, and by the Institute of Medical Informatics and Biometry (IMB) TU Dresden.

Availability of data and materials

The cancer cell lines from [17] are listed in Additional file 2: Table S1. All analyzed tumor patients are listed in Additional file 5: Table S4. CCTN, all data sets, R source code for network inference and network propagation licensed under GNU GPLv3, and usage examples are available from https://zenodo.org/record/58793 (DOI: 10.5281/zenodo.58793).

Authors' contributions

MS and AB developed the analysis. MS implemented and performed the analysis. MS and BF performed the quality control of the cell line data for CCTN inference. MS and AB wrote the manuscript. All authors contributed to the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

No ethical approval was required for this study. All utilized public omics data sets were generated by others who obtained ethical approval.

Author details

[1]Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Institute for Medical Informatics and Biometry, Fetscherstr. 74, 01307, Dresden, Germany. [2]National Center for Tumor Diseases (NCT), Dresden, Germany. [3]Institute of Molecular Systems Biology, Auguste-Piccard-Hof 1, 8093, Zurich, Switzerland. [4]Cellular Networks and Systems Biology, CECAD, University of Cologne, Joseph-Stelzmann-Str. 26, 50931, Cologne, Germany.

Received: 14 June 2016 Accepted: 6 September 2016
Published online: 03 October 2016

References

1. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10(11):1108–15.

2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
3. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948–62.
4. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238–44.
5. Ding J, McConechy MK, Horlings HM, Ha G, Chun Chan F, Funnell T, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Commun. 2015;6:8554.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
7. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2015;47(10):1127–33.
8. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.
9. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci. 2002;99:12963–8.
10. Louhimo R, Lepikhova T, Monni O, Hautaniemi S. Comparative analysis of algorithms for integration of copy number and expression data. Nat Methods. 2012;9(4):351–5.
11. Adler AS, Lin M, Horlings H, Nuyten DSA, van de Vijver MJ, Chang HY. Genetic regulators of large-scale transcriptional signatures in cancer. Nat Genet. 2006;38(4):421–30.
12. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Synder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463(21):318–25.
13. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143(6):1005–17.
14. Jörnsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling T, et al. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. Mol Syst Biol. 2011;7:486.
15. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106–14.
16. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108–13.
17. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
18. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B. 1996;58(1):267–88.
19. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796–804.
20. Seifert M, Garbe M, Friedrich B, Mittelbronn M, Klink B. Comparative transcriptomics reveals similarities and differences between astrocytoma grades. BMC Cancer. 2015;15:952.
21. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. 2014;42(2):413–68.
22. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83.
23. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.
24. Library of integrated network-based cellular signatures (LINCS). Broad LINCS center U54hg006093, pre-publication data communicated by Aravind Subramanian, L1000 data set. http://api.lincscloud.org/. Accessed 2 Jul 2014.
25. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42(Web Server issue):W449–60.
26. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
27. van Wieringen WN, Kun D, Hampel R, Boulesteix A-L. Survival prediction using gene expression data: a review and comparison. Comput Stat Data Anal. 2009;53(5):1590–603.
28. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol. 2014;32(7):644–52.
29. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841–60.
30. Madhavan S, Zenklusen J-C, Kotliarov Y, Sahni H, Fine HA, Buetow K. Rembrandt: helping personalized medicine become a reality through integrative translational research. Mol Cancer Res. 2009;7(2):157–67.
31. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17(1):98–110.
32. Lachance G, Uniacke J, Audas TE, Holterman CE, Franovic A, Payette J, et al. DNMT3a epigenetic program regulates the HIF-2α oxygen-sensing pathway and the cellular response to hypoxia. Proc Natl Acad Sci. 2014;111(21):7783–8.
33. Shah N, Sukumar S. The Hox genes and their roles in oncogenesis. Nat Rev Cancer. 2010;10(5):361–71.
34. Zhong Z, Shan M, Wang J, Liu T, Xia B, Niu M, et al. HOXD13 methylation status is a prognostic indicator in breast cancer. Int J Clin Exp Pathol. 2015;8(9):10716–24.
35. Henrich KO, Bauer T, Schulte J, Ehemann V, Deubzer H, Gogolin S, et al. CAMTA1, a 1p36 tumor suppressor candidate, inhibits growth and activates differentiation programs in neuroblastoma cells. Cancer Res. 2011;71(8):3142–51.
36. Vrouwe MG, Elghalbzouri-Maghrani E, Meijers M, Schouten P, Godthelp BC, Bhuiyan ZA, et al. Increased DNA damage sensitivity of Cornelia de Lange syndrome cells: evidence for impaired recombinational repair. Hum Mol Genet. 2007;16(12):1478–87.
37. Meng X, Lu P, Bai H, Xiao P, Fan Q. Transcriptional regulatory networks in human lung adenocarcinoma. Mol Med Rep. 2012;6(5):961–6.
38. Leung CS, Yeung T-L, Yip K-P, Pradeep S, Balasubramanian L, Liu J, et al. Calcium-dependent FAK/CREB/TNNC1 signalling mediates the effect of stromal MFAP5 on ovarian cancer metastatic potential. Nat Commun. 2014;5:5092.
39. Sharma V, Koul N, Joseph C, Dixit D, Ghosh S, Sen E. HDAC inhibitor, scriptaid, induces glioma cell apoptosis through JNK activation and inhibits telomerase activity. J Cell Mol Med. 2010;14(8):2151–61.
40. Wu Y, Song SW, Sun J, Bruner JM, Fuller GN, Zhang W. IIp45 inhibits cell migration through inhibition of HDAC6. J Biol Chem. 2010;285(6):3554–60.
41. Trebinska A, Rembiszewska A, Ciosek K, Ptaszynski K, Rowinski S, Kupryjanczyk J, et al. HAX-1 overexpression, splicing and cellular localization in tumors. BMC Cancer. 2010;10:76.
42. Paushkin SV, Patel M, Furia BS, Peltz SW, Trotta CR. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3′ end formation. Cell. 2004;117(3):311–21.
43. Saha B, Ypsilanti AR, Boutin C, Cremer H, Chédotal A. Plexin-B2 regulates the proliferation and migration of neuroblasts in the postnatal and adult subventricular zone. J Neurosci. 2012;32(47):16892–905.
44. Chi Z, Byrne ST, Dolinko A, Harraz MM, Kim M-S, Umanah G, et al. Botch is a γ-glutamyl cyclotransferase that deglycinates and antagonizes Notch. Cell Rep. 2014;7(3):681–8.
45. Mungrue IN, Pagnon J, Kohannim O, Gargalovic PS, Lusis AJ. CHAC1/MGC4504 is a novel proapoptotic component of the unfolded protein response, downstream of the ATF4-ATF3-CHOP cascade. J Immunol. 2009;182(1):466–76.
46. Joo NE, Ritchie K, Kamarajan P, Miao D, Kapila YL. Nisin, an apoptogenic bacteriocin and food preservative, attenuates HNSCC tumorigenesis via CHAC1. Cancer Med. 2012;1(3):295–305.
47. Goebel G, Berger R, Strasak AM, Egle D, Müller-Holzner E, Schmidt S, et al. Elevated mRNA expression of CHAC1 splicing variants is associated with poor outcome for breast and ovarian cancer patients. Br J Cancer. 2012;106(1):189–98.
48. Sturm D, Bender S, Jones DTW, Lichter P, Grill J, Becher O, et al. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. Nat Rev Cancer. 2014;14(2):92–107.

49. Talasila KM, Soentgerath A, Euskirchen P, Rosland GV, Wang J, Huszthy PC, et al. EGFR wild-type amplification and activation promote invasion and development of glioblastoma independent of angiogenesis. Acta Neuropathol. 2013;125(5):683–98.
50. Ohgaki H, Kleihues P. The definition of primary and secondary glioblastoma. Clin Cancer Res. 2013;19(4):764–72.
51. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155(2):462–77.
52. Solimini NL, Xu Q, Mermel CH, Liang AC, Schlabach MR, Luo J, et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science. 2012;337(6090):104–9.
53. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews M, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893–8.
54. Lu X, Thomas RK, Peifer M. CGARS: cancer genome analysis by rank sums. Bioinformatics. 2014;30(9):1295–6.
55. Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM). A genomics-based classification of human lung tumors. Sci Transl Med. 2013;5(209):209ra153.
56. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11(1):9–22.
57. Zhang P, Mirani N, Baisre A, Fernandes H. Molecular heterogeneity of head and neck squamous cell carcinoma defined by next-generation sequencing. Am J Pathol. 2014;184(5):1323–30.
58. Ernst J, Kellis M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. 2013;23(7):1142–54.
59. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.
60. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015;347(6217):78–81.
61. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518(7539):360–4.
62. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Ass. 2004;99(468):909–17.
63. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
64. Hastie T, Efron B. LARS: Least angle regression, lasso and forward stagewise. 2013. https://cran.r-project.org/web/packages/lars/index.html. Accessed 24 Sept 2013.
65. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. covTest: Computes covariance test for adaptive linear modelling. 2013. https://cran.r-project.org/web/packages/covTest/index.html. Accessed 24 Sept 2013.
66. Kutner MH, Nachtsheim CJ, Neter J. Applied linear regression models, volume 4. New York: McGraw-Hill Education; 2004.
67. Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. Autoregressive higher-order hidden Markov models: exploiting local chromosomal dependencies in the analysis of tumor expression profiles. PLoS ONE. 2014;e100295:.
68. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57:289–300.
69. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.

4.5 Publication:

regNet: an R package for network-based propagation of gene expression alterations

Journal: Bioinformatics

Received: 12 June 2017; **Accepted**: 30 August 2017; **Published**: 31 August 2017 (online), 15 January 2018 (print)

Citation: Michael Seifert and Andreas Beyer (2018): regNet: an R package for network-based propagation of gene expression alterations, Bioinformatics, 34(2), 308-311.

Placement and summary of the publication

Hundreds or even thousands of genes are typically altered in their expression comparing disease to normal tissue. Gene expression changes and potentially underlying gene copy number alterations can be measured routinely by wet lab experiments (e.g. Pollack et al. (2002); Hastings et al. (2009); Henrichsen et al. (2009)), but the quantification of individual impacts of altered genes on clinically relevant characteristics (e.g. cell proliferation, altered signaling pathways, patient survival) is still very challenging. Frequently altered genes can be determined by comparing disease to normal samples using standard statistical tests (e.g. Ritchie et al. (2015)), but contributions of individual sample-specific gene expression alterations on clinically relevant characteristics cannot be determined by such approaches. A promising strategy to address this is the analysis of altered genes with the help of protein or gene interaction networks utilizing network propagation (Hofree et al. (2013); Leiserson et al. (2015); Seifert et al. (2016)).

I have developed the R package regNet to provide user-friendly implementations of our network inference and network propagation algorithms that we established in our prior work (Seifert et al. (2016)). regNet utilizes gene expression and gene copy number data to learn gene regulatory networks to quantify potential impacts of individual gene expression alterations on user-defined target genes via network propagation. regNet provides an excellent starting point for the analysis of transcriptome data in the context of gene regulatory networks.

We demonstrated the value of regNet by identifying putative major regulators that distinguish pilocytic astrocytomas from diffuse astrocytomas using data of my prior study (Seifert et al. (2015)). We revealed that especially the downregulation of TBR1, a transcription factor ex-

pressed in post-mitotic cells and required for normal brain development (Bulfone et al. (1995)), could strongly contribute to the increased malignancy of diffuse astrocytomas. Further, we used regNet to predict putative impacts of glioblastoma-specific gene copy number alterations on known cell cycle pathway genes and patient survival. Interestingly, long-lived glioblastoma patients tended to show more gene copy number alterations that impact on the cell cycle than short-lived patients. Several of these genes had important functions in the regulation of cell growth, migration and proliferation indicating that at least some of these genes may counteract fast tumor growth.

regNet contributes to the quantification of individual impacts of sample-specific gene expression alterations on user-defined target genes. regNet can identify potential key drivers and can quantify combined impacts of altered genes on clinically relevant characteristics. Moreover, regNet has been used to realize my joint study with Gladitz et al. (2018) and my study Seifert et al. (2019) that are both part of this habilitation thesis (see Sections 4.6 and 4.7).

Author contribution

I developed the R package and designed its key structure. I implemented all methods for data handling, network inference and network flow analysis. I designed the two application studies. I wrote the manuscript and performed the revision of the manuscript. I trained the testers Josef Gladitz, who was a MD student in my group, and Xiaohui Wu, who was a visiting postdoc in the group of Andreas Beyer, on how to use regNet and how to implement algorithmic extensions. Andreas Beyer supported the writing and the revision of the manuscript.

Bioinformatics, 34(2), 2018, 1–4
doi: 10.1093/bioinformatics/btx544
Advance Access Publication Date: 31 August 2017
Applications Note

Systems biology

regNet: an R package for network-based propagation of gene expression alterations

Michael Seifert[1,2]* and Andreas Beyer[3]

[1]Carl Gustav Carus Faculty of Medicine, Institute for Medical Informatics and Biometry (IMB), Technische Universität Dresden, D-01307 Dresden, Germany, [2]National Center for Tumor Diseases (NCT), Dresden, Germany and [3]Cellular Networks and Systems Biology, CECAD, University of Cologne, D-50931 Cologne, Germany

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on June 12, 2017; revised on July 31, 2017; editorial decision on August 29, 2017; accepted on August 30, 2017

Abstract

Summary: Gene expression alterations and potentially underlying gene copy number mutations can be measured routinely in the wet lab, but it is still extremely challenging to quantify impacts of altered genes on clinically relevant characteristics to predict putative driver genes. We developed the R package regNet that utilizes gene expression and copy number data to learn regulatory networks for the quantification of potential impacts of individual gene expression alterations on user-defined target genes via network propagation. We demonstrate the value of regNet by identifying putative major regulators that distinguish pilocytic from diffuse astrocytomas and by predicting putative impacts of glioblastoma-specific gene copy number alterations on cell cycle pathway genes and patient survival.

Availability and implementation: regNet is available for download at https://github.com/seifemi/regNet under GNU GPL-3.

Contact: michael.seifert@tu-dresden.de

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

1 Introduction

Gene expression alterations play a central role in many complex genetic diseases. Molecular alterations such as DNA copy number mutations can trigger expression changes of directly affected genes that further influence the expression of other genes (Hastings *et al.*, 2009; Henrichsen *et al.*, 2009). Typically, many genes (hundreds or even thousands) are differentially expressed between disease and healthy samples. This still puts great challenges on the identification of potential key disease drivers. Standard statistical tests only allow to identify frequently altered genes, but contributions of individual sample-specific gene expression alterations on clinically relevant signatures cannot be quantified routinely so far. Generally, network-based analyses represent promising strategies to overcome this (Hofree *et al.*, 2013; Leiserson *et al.*, 2015; Seifert *et al.*, 2016).

Here, we present the R package regNet that utilizes gene regulatory networks to quantify impacts of gene-specific expression changes on clinically relevant signature genes. regNet propagates expression changes through networks. Thereby it becomes possible to predict the impact of alterations of specific genes (e.g. mutations) on other, disease-related genes. The mathematical framework behind regNet has been described in great detail along with in-depth validation studies in Seifert *et al.* (2016). The basic regNet workflow is illustrated in Figure 1a. First, regNet learns a regulatory network from a large collection of paired gene expression and copy number profiles. Next, regNet utilizes this network to quantify impacts of sample-specific gene expression changes (source genes, e.g. differentially expressed genes with underlying copy number mutations in a tumor specified by the user) on other clinically relevant target genes (e.g. known disease markers) using network propagation. This enables to identify potential key drivers. regNet enables user-friendly access to the network inference algorithm and the network propagation algorithms described in Seifert *et al.* (2016).

After providing details to the implementation, we demonstrate the basic usage and value of regNet in two case studies. First, we

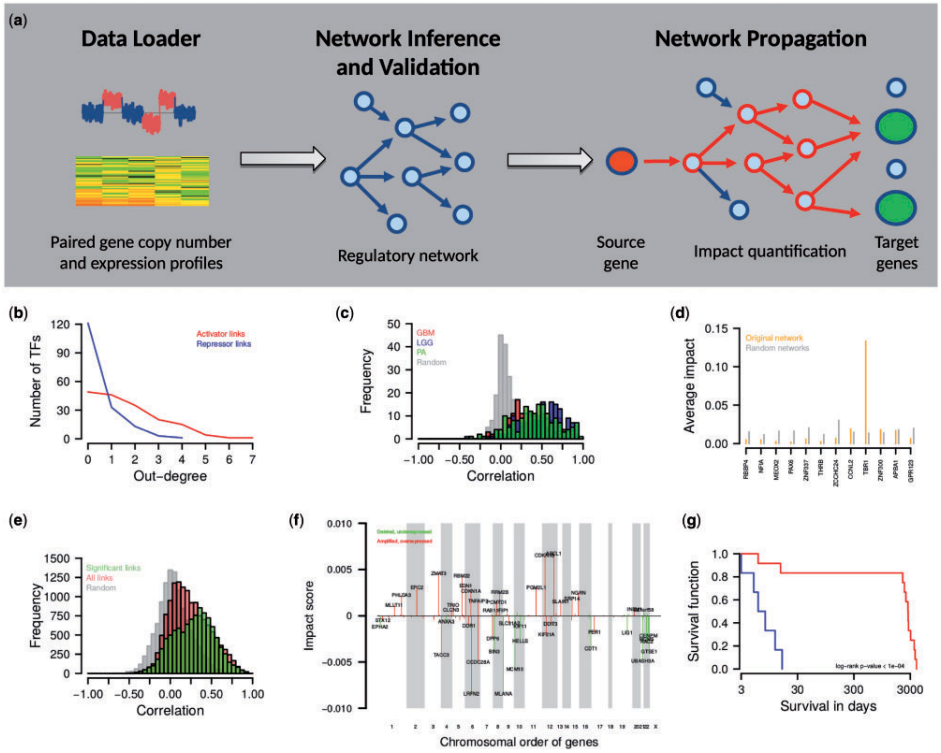

Fig. 1. regNet workflow (**a**) and case study results (**b–g**). (**a**) Basic regNet workflow (grey box). Paired gene expression and copy number data are loaded (data loader) and used to learn a regulatory network (network inference) followed by testing the predictive power of the network on different samples (network validation) and the quantification of impacts (network propagation) of sample-specific gene expression alterations (source genes, e.g. differentially expressed genes with underlying copy number mutations in a tumor specified by the user) on user-defined target genes (e.g. clinically relevant signature genes). (**b**) Out-degree distribution of TF network considering activator (red) and inhibitor (blue) links. (**c**) Correlations between TF network predicted and originally measured gene-specific expression levels of three independent astrocytoma test cohorts (GBM, LGG, PA) in comparison to average predictions obtained by ten random networks of the same complexity (Random). (**d**) Average impacts of major regulators on other TFs comparing the learned (orange) TF network to 100 random networks of the same complexity (grey). (**e**) Predictive power of the learned cancer cell line network on glioblastoma patients for different network instances. The network instance that only considers the most relevant links (green) is significantly better than the instance that considers all learned links (red) and than random networks of the same complexity (grey). (**f**) Genome-wide characterization of patient-specific differentially expressed genes with directly underlying gene copy number mutations on cell cycle pathway genes for a selected long-lived glioblastoma patient (TCGA-06-6693). Shown are relative average impacts that can be inhibitory (negative impact score) or activating (positive impact score) for underexpressed genes affected by deletions (green) or overexpressed genes affected by amplifications (red). (**g**) Kaplan–Meier curve distinguishing short from long-lived glioblastoma patients utilizing patient-specific survival impact scores

analyze the potential interplay of transcription factors that were differentially expressed between pilocytic and diffuse astrocytomas suggesting that underexpression of TBR1 may contribute to increased malignancy of diffuse astrocytomas. Next, we predict potential impacts of glioblastoma-specific gene copy number mutations on cell cycle pathway genes and show that tumor-specific gene copy number mutations enable to distinguish between short and long survival.

2 Implementation

regNet is divided into four main modules enabling data loading, network inference, network-based predictions of gene expression levels, and network propagation. regNet uses a fixed folder structure to enable a user-friendly storage and loading of results. A comprehensive summary about the underlying mathematical models is given in Text S1. Detailed information about the individual modules and the fixed folder structure are provided in Text S2. A basic demonstration of regNet code usage is outlined in Text S3.

2.1 Data loader

regNet requires gene expression profiles and corresponding gene copy number profiles as input data for network inference, network predictions and network propagation. regNet can handle tab-delimited datasets with a fixed column-structure. regNet also allows to transform datasets to apply a learned network for the analysis of

new datasets with different numbers of genes than the initial dataset that was used to learn the network. This is done by removing additional genes and by setting expression measurements of missing genes to zero (no contribution to analysis).

2.2 Network inference

regNet splits the global network inference problem into independent gene-specific sub-network inference tasks. regNet models the expression of each gene as a linear combination of its own copy number and the expression of other putative regulators. A detailed description of the underlying mathematical model and in-depth validation studies were done in Seifert *et al.* (2016). See Text S1 for an overview of the underlying mathematical model. Lasso (least absolute shrinkage and selection operator) regression (Tibshirani, 1996) followed by a significance test for lasso (Lockhart *et al.*, 2014) is used to learn for each gene those predictors (gene copy number and/or expression levels of other genes) that best predict the expression level of the gene in a given dataset. Corresponding FDR-adjusted *P*-values (Benjamini and Hochberg, 1995) can be used to obtain user-defined network instances. regNet stores all *P*-values for learned gene-gene associations (links) and enables the user to choose different *P*-value cutoffs to create and test network instances of variable confidence and complexity. A specific network instance consists of a subset of the links at a user-defined *P*-value cutoff. Network inference is usually very time-consuming. In order to speed up the computation, network inference can be split across multiple compute cores. Further, random network instances of a learned network can be obtained by degree-preserving network permutations.

2.3 Network-based prediction of gene expression levels

In order to evaluate the quality of a previously learned network, regNet can be used to predict the expression levels of genes in a given dataset. regNet quantifies the prediction quality of each individual gene by computing correlations between predicted and originally measured expression levels. See Text S1 for more details. Ideally, this evaluation should be done using an independent test dataset that was not used for network inference. Different instances of a network can be analyzed by setting a global *P*-value cutoff and a local gene cutoff (excludes genes in close chromosomal proximity as gene-specific predictors, because such predictors may only represent the underlying local DNA copy number instead of potential regulatory dependencies) to only consider the most relevant network links. The obtained gene-specific correlations enable to evaluate the predictive power of the underlying network. regNet stores these correlations and corresponding *P*-values in a tab-delimited file for further analysis. This functionality is also available for random networks enabling comparisons to baseline models. Further, correlations between predicted and measured expression levels provide the basis to integrate the quality of the predictions of individual genes into the impact computations via the network propagation module (Text S1).

2.4 Network propagation

regNet quantifies for a given dataset the impact of individual regulator genes on all other genes utilizing a previously learned network. This algorithm quantifies for each gene pair (*a*, *b*) the direct and indirect contribution of gene *a* on the expression of gene *b* under consideration of all existing network paths from *a* to *b*, the prediction quality of individual genes along the paths, and possibly existing feedback loops. These impact scores can be computed over all patients in a cohort or for each individual patient. regNet can also integrate contributions of genes acting as potential inhibitors or activators for individual patients. This is done by accounting for the sign of effects (activating, positive sign or inhibiting, negative sign) of individual network links. regNet implements different functions to utilize these possibilities (Suppelmentary Table S1). The statistical significance of individual impact scores can be determined by comparisons to impact scores obtained under corresponding random networks. This enables to identify those genes that have the greatest impact on user-defined target genes. An overview of the underlying mathematical models is provided in Text S1. Implementation details are given in Text S2. Code usage examples are shown in Text S3. Further details on the impact computation and an in-depth validation are provided in Seifert *et al.* (2016).

3 Application

We demonstrate the basic functionality and the potential of regNet in two case studies.

3.1 Identification of hub regulators distinguishing pilocytic from diffuse astrocytomas

We analyzed the potential interplay and activity of transcription factors (TFs) that distinguish pilocytic astrocytomas (PA) in children from diffuse astrocytomas (AS) in adults. This study requires less than ten minutes on a standard computer enabling to become familiar with the basic regNet functionality. Details of individual regNet function calls of this case study are provided in Text S3. We used regNet to learn a putative TF-TF interaction network of 171 TFs based on gene expression and copy number data of 124 different astrocytomas (47 PA and 77 AS samples) from Seifert *et al.* (2015). Characteristics of the network are summarized in Figure 1(b–d) and Supplementary Figure S2. The network contained more putative activator (269 of 341) than inhibitor links (72 of 341) (Fig. 1b). Only few TFs had more than four outgoing links (Supplementary Fig. S2b). Some of these potential major regulators are known to be involved in the development of the central nervous system (PAX6, THRB, TBR1), cell proliferation (MEOX2), apoptosis (CCNL2), or histone acetylation (RBBP4) (Safran *et al.*, 2010). We further used the network to predict the expression of genes in three independent cohorts. The network reached a significantly better prediction of gene expression levels than random networks of the same complexity (Fig. 1c, Text S3). Finally, we determined which potential major regulators had great impact on other TFs in the network. We found that TBR1 has by far the strongest impact on other TFs (Fig. 1d). TBR1 is expressed in post-mitotic cells and required for normal brain development (Bulfone *et al.*, 1995). We found that the expression of TBR1 was clearly reduced in AS compared to normal brain and PA suggesting that TBR1 may contribute to the strongly increased malignancy of AS compared to PA.

3.2 Impacts of glioblastoma gene copy number mutations on cell cycle pathway genes and patient survival

We used regNet to predict glioblastoma-specific gene copy number mutations that influence the expression of cell cycle genes and patient survival. This study required about 6.3 computing hours using 400 cores of a compute server [Bull HPC-Cluster (Taurus), Intel(R) Xeon(R) CPU E5-2680 v3 2.50 GHz, ZIH TU Dresden]. Most time was used for network inference, which generally scales linearly with the number of cores making such a study also feasible on a compute server with less cores. More details to this study are provided in

Text S4. We learned a genome-wide transcriptional regulatory network from gene expression and copy number data of 15 811 genes in 768 human cancer cell lines (Barretina *et al.*, 2012). This network contained much more activator (46 366 of 53 955) than inhibitor (7589 of 53 955) links and 4938 genes had a direct copy number effect. This observation was not unexpected (e.g. synchronous activation of target genes of a TF will lead to activator links). We used this network to predict gene expression of glioblastomas (TCGA, 2008). In accordance with Seifert *et al.* (2016), we found that the predictive power of this network was significantly better than for random networks of the same complexity and a corresponding more complex network that utilized all learned links without filtering for significant links (Fig. 1e, Supplementary Fig. S3, Supplementary Text S4). Next, we determined the impact of differentially expressed genes with underlying copy number mutations on the expression of cell cycle pathway genes for patients with very short (8 patients: less than 20 days) or very long (10 patients: more than 2000 days) survival. We found that genes with significant impact on cell cycle were not disjoint between short and long-lived patients (Fig. 1f, Supplementary Fig. S4). Interestingly, there was a tendency that long-lived patients tend to contain more gene copy number mutations that impact on the expression of cell cycle pathway genes than short-lived patients (Supplementary Fig. S5). Several of these genes play important roles in the regulation of cell growth, migration and proliferation (Text S4). This suggests that some of the observed gene copy number mutations in long-lived patients may counteract fast tumor growth with benefits for patient survival. Finally, we computed the impacts of differentially expressed genes with underlying copy number mutations on known survival signature genes for each of the short and long-lived patients. regNet was able to separate the selected glioblastoma patients into a short and long-lived group (Fig. 1g, Supplementary Fig. S6).

4 Conclusion

regNet predicts the impact of gene expression alterations on user-defined target genes, while accounting for direct and indirect network effects. regNet can identify potential key drivers and quantify combined impacts of altered genes on clinically relevant characteristics. Since network inference and propagation are typically very time and resource consuming for large datasets, we recommend to use regNet on a compute server. regNet currently exploits information contained in gene expression and copy number data for network inference, but the underlying mathematical model is flexible enough to enable the integration of additional omics layers. We finally note that applications of regNet are not necessarily limited to cancer.

Acknowledgements

We thank Josef Gladitz (IMB TU Dresden) and Xiaohui Wu (CECAD Cologne) for testing regNet. We thank the Center for Information Services and High Performance Computing (ZIH) TU Dresden for providing computational and storage resources. We thank the reviewers for their valuable comments.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

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

Bulfone,A. *et al.* (1995) T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron*, **15**, 63–78.

Hastings,P.J. *et al.* (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.

Henrichsen,C.N. *et al.* (2009) Copy number variants, diseases and gene expression. *Hum. Mol. Genet.*, **18**, R1–R8.

Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.

Leiserson,M.D.M. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.

Lockhart,R. *et al.* (2014) A significance test for the lasso. *Ann. Stat.*, **42**, 413–468.

Safran,M. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)*, **2010**, baq020.

Seifert,M. *et al.* (2015) Comparative transcriptomics reveals similarities and differences between astrocytoma grades. *BMC Cancer*, **15**, 952.

Seifert,M. *et al.* (2016) Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. *Genome Biol.*, **17**, 204.

TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Tibshirani,R. (1996) Shrinkage and Selection via the Lasso. *J. R Stat. Soc. B*, **58**, 267–288.

4.6 Publication:

Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion

Journal: Acta Neuropathologica Communications

Received: 19 April 2018; **Accepted**: 8 May 2018; **Published**: 11 June 2018

Citation: Josef Gladitz, Barbara Klink and Michael Seifert (2018): Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion, Acta Neuropathologica Communications, 6:49.

Placement and summary of the publication

Oligodendrogliomas represent a specific class of human brain tumors that are characterized by a joint loss of one copy of the p-arm of chromosome 1 and the q-arm of chromosome 19 (1p/19q co-deletion) in combination with a heterozygous point mutation of the isocitrate dehydrogenase gene (IDH1/2) (Labussiere et al. (2010); Louis et al. (2016)). The IDH mutation is known to induce an epigenetic reprogramming of the tumor cells (Noushmehr et al. (2010); Turcan et al. (2012)) and the 1p/19 co-deletion is of important clinical relevance to distinguish oligodendrogliomas from closely related astrocytomas (Louis et al. (2016)). The 1p/19q co-deletion develops most likely from an unbalanced translocation (Jenkins et al. (2006)). On the one hand, this suggested that driver genes could be located in close chromosomal proximity to the fusion point, but no oncogenic fusion genes have been discovered so far. On the other hand, the recurrently occurring 1p/19q co-deletion suggest that tumor suppressors are located on the 1p and 19q arm. Inactivating point mutations of FUBP1 on 1p and of CIC on 19q have been identified (Bettegowda et al. (2011); Eisenreich et al. (2013)), but they were not present in each oligodendroglioma (The Cancer Genome Atlas Research Network (2015)) suggesting

that both genes are not responsible for the initial development of oligodendrogliomas. Thus, despite the important clinical relevance of the 1p/19q co-deletion, the underlying pathology still remains elusive since many years.

This motivated us to search for potential driver genes by focusing on gene expression changes that were induced by the 1p/19q co-deletion. The loss of one allele of each gene on 1p and 19q could directly contribute to the development of oligodendrogliomas by reduced expression or indirectly by alterations of regulatory networks of the tumor cells. Since hundreds of genes on 1p and 19q are downregulated and because the 1p/19q co-deletions are nearly identical in different oligodendrogliomas, we could not utilize standard statistical approaches for differential gene expression analysis to distinguish between potential driver and passenger genes. We addressed this challenge by developing a network-based strategy for the identification of putative driver genes within the region of the 1p/19q co-deletion by utilizing my R package regNet (Seifert and Beyer (2018)).

In our study, we learned oligodendroglioma-specific gene regulatory networks based on publicly available gene expression and copy number data from The Cancer Genome Atlas Research Network (2015). We utilized these networks to quantify potential impacts of differentially expressed genes within the 1p/19q region on cancer-relevant signaling and metabolic pathways. We predicted 8 genes with strong impact on signaling pathways and 14 genes with strong impact on metabolic pathways. In-depth literature analysis suggested that many of these genes probably push and others may counteract oligodendroglioma development. Among these candidates was ELTD1, a key player of tumor angiogenesis (Masiero et al. (2013)) and functionally validated glioblastoma oncogene (Towner et al. (2013); Ziegler et al. (2017)), which was overexpressed despite the loss of one copy of the 1p arm. Further, we found that the glioblastoma tumor suppressor SLC17A7 (Lin et al. (2015)) on 19q was underexpressed. Moreover, we found that SDHB, which triggers epigenetic alterations in paragangliomas (Letouzé et al. (2013); Aspuria et al. (2014); Baysal and Maher (2015)), was underexpressed and may support and possibly enhance the epigenetic reprogramming of oligodendrogliomas that is induced by the IDH mutation (Cohen et al. (2013); Louis et al. (2016)). In addition, we analyzed other rarely observed chromosomal deletions and amplifications and identified putative drivers within these regions that could contribute to the development of specific oligodendroglioma subgroups.

Generally, our unique in-depth computational study contributes to a better understanding of the oligodendroglioma pathology and may open the possibility to develop new therapeutic strategies in the future. Unfortunately, functional validations of our findings by wet lab experiments were not possible, because oligodendroglioma cells do not grow in cell culture and mouse models of oligodendrogliomas did not exist, but future progress may enable this.

Author contribution

I developed the concept of the study and supervised the realization of the study by Josef Gladitz, who was an Else-Kröner MD-student in my group. I further drafted the structure of the manuscript, revised the figures and contributed significantly to the writing of the manuscript. I discussed our findings with Barbara Klink, who supported the biological interpretation of our results. Josef Gladitz supported the writing of the manuscript by providing key words and key results to each section. I performed the revision of the manuscript.

Gladitz *et al. Acta Neuropathologica Communications* (2018) 6:49
https://doi.org/10.1186/s40478-018-0544-y

Acta Neuropathologica Communications

RESEARCH Open Access

Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion

Josef Gladitz[1], Barbara Klink[2,3] and Michael Seifert[1,3*]

Abstract

Oligodendrogliomas are primary human brain tumors with a characteristic 1p/19q co-deletion of important prognostic relevance, but little is known about the pathology of this chromosomal mutation. We developed a network-based approach to identify novel cancer gene candidates in the region of the 1p/19q co-deletion. Gene regulatory networks were learned from gene expression and copy number data of 178 oligodendrogliomas and further used to quantify putative impacts of differentially expressed genes of the 1p/19q region on cancer-relevant pathways. We predicted 8 genes with strong impact on signaling pathways and 14 genes with strong impact on metabolic pathways widespread across the region of the 1p/19 co-deletion. Many of these candidates (e.g. *ELTD1*, *SDHB*, *SEPW1*, *SLC17A7*, *SZRD1*, *THAP3*, *ZBTB17*) are likely to push, whereas others (e.g. *CAP1*, *HBXIP*, *KLK6*, *PARK7*, *PTAFR*) might counteract oligodendroglioma development. For example, *ELTD1*, a functionally validated glioblastoma oncogene located on 1p, was overexpressed. Further, the known glioblastoma tumor suppressor *SLC17A7* located on 19q was underexpressed. Moreover, known epigenetic alterations triggered by mutated *SDHB* in paragangliomas suggest that underexpressed *SDHB* in oligodendrogliomas may support and possibly enhance the epigenetic reprogramming induced by the *IDH*-mutation. We further analyzed rarely observed deletions and duplications of chromosomal arms within oligodendroglioma subcohorts identifying putative oncogenes and tumor suppressors that possibly influence the development of oligodendroglioma subgroups. Our in-depth computational study contributes to a better understanding of the pathology of the 1p/19q co-deletion and other chromosomal arm mutations. This might open opportunities for functional validations and new therapeutic strategies.

Keywords: Oligodendrogliomas, 1p/19q co-deletion, Network biology, Network inference, Network propagation, Cancer genomics, Bioinformatics, Computational systems biology

Introduction

Between 4 and 8 percent of all primary human brain tumors are classified as oligodendrogliomas [80]. Oligodendrogliomas belong to the class of diffuse gliomas that typically show infiltrative growth into the surrounding brain tissue, relapse, and progression to more aggressive tumors [54]. Histological similarities to normal oligodendrocytes were used for many years to diagnose oligodendrogliomas [47], but pure histological classifications can vary considerably between neuropathologists [14, 76]. Therefore, molecular markers for a more robust classification of oligodendrogliomas have been explored. First, it has been revealed that the majority of oligodendrogliomas showed a recurrent loss of heterozygosity of the chromosomal arms 1p and 19q (1p/19q co-deletion) associated with improved chemotherapy response and longer relapse-free survival [9, 28, 60]. Further, the 1p/19q-co-deletion is always accompanied by heterozygous somatic point mutations of the isocitrate dehydrogenase gene (*IDH1/2*) [37]. These *IDH*-mutations are known to induce the glioma-CpG island methylator phenotype

*Correspondence: michael.seifert@tu-dresden.de
[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
[3]National Center for Tumor Diseases, Dresden, Germany
Full list of author information is available at the end of the article

(G-CIMP) [53, 75]. Both characteristic molecular markers (1p/19q co-deletion and *IDH* mutation) have recently been included into the new 2016 World Health Organization (WHO) classification system for tumors of the central nervous system [48]. This new classification utilizes histological features in combination with the co-occurrence of the 1p/19q co-deletion and the *IDH*-mutation to diagnose oligodendrogliomas.

So far, the clinical relevance of the oligodendroglioma-specific 1p/19q co-deletion has been well-studied [9, 10, 28, 33, 69, 78], but the pathogenesis of the recurrent 1p/19q co-deletion still remains elusive. The 1p/19q co-deletion is likely to emerge from an unbalanced translocation between the 1q and 19p arm [29]. This suggests that driver genes could be located in close proximity to the fusion points, but no oncogenic fusion genes have been reported. On the other hand, the recurrent 1p/19q co-deletion suggests that tumor suppressors could be located on the 1p and 19q arm. According to the classical two hit hypothesis, both alleles of a tumor suppressor must be mutated to contribute to oncogenesis [36]. The search for inactivating point mutations on the remaining copies of the 1p and 19q arm identified *FUBP1* located on 1p and *CIC* located on 19q as potential tumor suppressors [8, 20]. But *FUBP1* mutations are only observed in about 29% and *CIC* mutations in about 62% of oligodendrogliomas [69]. This implies that these mutations occur later during tumor development and are therefore not responsible for the initial development of oligodendrogliomas. Moreover, it is likely that haploinsufficiency [16, 63] induced by the 1p/19q co-deletion may contribute to the development of oligodendrogliomas. The loss of one allele of each gene on 1p and 19q could directly contribute to oncogenesis by reduced expression levels or indirectly by alterations of regulatory networks. However, standard statistical approaches are not suited to identify differentially expressed driver genes on 1p/19q, because hundreds of genes are down-regulated on both chromosomal arms due to the co-deletion making it impossible to distinguish between driver and passenger genes. Further, the recurrence of virtually identical 1p/19q co-deletions in different oligodendrogliomas does not allow to narrow down chromosomal regions on 1p and 19q where driver genes might be located.

Novel computational strategies are required to search for putative cancer candidate genes located within the region of the 1p/19q co-deletion. Generally, the analysis of gene mutations in the context of gene interaction networks represents a promising strategy to address this challenge [24, 39, 65]. Importantly, we recently showed that gene regulatory networks inferred from gene expression and copy number data can be used to quantify impacts of gene copy number mutations on cancer-relevant target genes [64, 65]. The key idea behind this approach is the propagation of gene expression alterations through a gene regulatory network to determine how individual gene copy number mutations influence the expression of other genes in the network. Utilizing such an approach, each individual gene located within the region of the 1p/19q co-deletion can be analyzed offering the unique possibility to search for novel cancer candidate genes that influence the development of oligodendrogliomas.

Here, we develop a network-based approach to identify novel putative cancer gene candidates for oligodendrogliomas (Fig. 1). We utilized gene expression and copy number data of 178 histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) to learn gene regulatory networks. We used these networks to determine impacts of differentially expressed genes with underlying copy number mutations on known cancer-relevant signaling and metabolic pathway genes utilizing network propagation. We screened the region of the recurrent 1p/19q co-deletion and other rarely mutated chromosomal arms revealing several interesting novel putative cancer candidate genes that have the potential to be involved in the development of oligodendrogliomas.

Materials and methods

Gene copy number and expression data

DNA copy number profiles (aCGH), gene expression data (RNA-seq), and clinical annotations of 178 histologically classified oligodendrogliomas (133 with and 45 without 1p/19q co-deletion) of the TCGA lower grade glioma (LGG, gdc.cancer.gov) cohort and gene expression data (RNA-seq) of three commercially available normal brain samples (StrataGen, BioChain, and Clonetech) from [38] were considered. The tumors represent the 1p/19q and IDHme subgroups described in [38]. Gene copy number profiles of individual tumors were determined from aCGH profiles as described in [65]. Gene expression counts of tumor and normal samples were jointly normalized with the cyclic loess method using the R function voom of the limma package [61]. We finally included gene copy number and gene expression measurements of 12,285 genes in our data set after excluding all genes with very low expression values (less than 1 read count per million reads mapped) in at least 50% of samples. Further, aCGH and gene expression data of 118 histologically classified oligoastrocytomas (34 with and 84 without 1p/19q co-deletion) of the TCGA LGG cohort were processed in the same way and considered for independent network validation. All processed data are contained in Additional file 1: Table S1 and Additional file 2: Table S2.

Identification of chromosomal aberrations and gene copy number mutations

Hierarchical clustering of gene copy number profiles of the 178 histologically classified oligodendrogliomas was

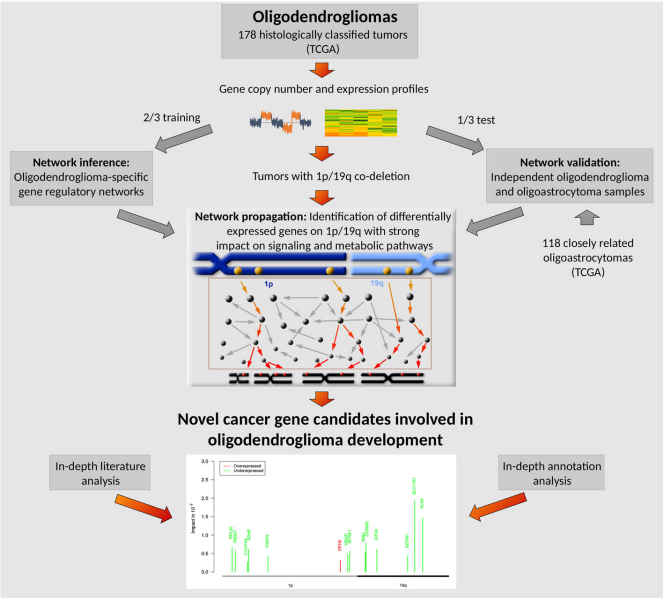

Fig. 1 Methodological overview. Gene copy number and expression data of 178 histologically classified oligodendrogliomas from The Cancer Genome Atlas (TCGA) were randomly split into training (two-thirds) and corresponding test (one-third) samples. Oligodendroglioma-specific gene regulatory networks were learned based on gene expression and copy number data of each training set. The obtained networks were validated on their corresponding oligodendroglioma test sets and independent data of closely related oligoastrocytomas from TCGA. Network propagation was applied to genes within the region of the 1p/19q co-deletion to identify those genes that had a strong impact on the expression of known cancer-relevant signaling and metabolic pathways

done using the R function heatmap.2 (euclidean distance, complete linkage) of the R package gplots [79]. We found that 133 tumors had the characteristic 1p/19q co-deletion (Additional file 3: Figure S1, see [38] for all tumors). We considered each of these tumors with 1p/19q co-deletion and determined deleted and duplicated genes. To realize this, we computed the average copy number $\log_2$-ratio r_i ($r_i < 0$) of tumor to normal DNA within the 1p/19q co-deletion region for tumor $i \in \{1, \cdots, 133\}$. We marked each gene in tumor i as deleted if its gene-specific copy number log-ratio was less than $0.5 \cdot r_i$. In analogy, we marked each gene as duplicated if its log-ratio was greater than $-0.5 \cdot r_i$. We considered a scaling factor of 0.5 to account for the fact that gene copy number measurements are typically noisy depending on the individual tumor content of the patient samples. This enabled us to specify for each tumor all genes and chromosomal regions (e.g. deletion of 4q, duplication of 7p) affected by deletions or duplications that were visible in the heatmap in addition to the 1p/19q co-deletion (Additional file 3: Figure S1).

Identification of differentially expressed genes

Differentially expressed genes between oligodendrogliomas with 1p/19q co-deletion and normal brain samples were derived by moderated t-tests using limma's standard workflow [61]. *P*-values were adjusted for multiple testing by computing *q*-values (R package qvalue) [68]. Under- and overexpressed genes in oligodendrogliomas in comparison to normal brain were selected using a *q*-value cutoff of 0.05 (Additional file 4: Table S3).

Gene and pathway annotation analysis

Gene annotations (transcription factors/cofactors, kinases, phosphatases, oncogenes, tumor suppressors) and genes included in cancer-relevant signaling and metabolic pathways were obtained from [65]. The number of differentially expressed genes per annotation category was determined separately for under- and overexpressed genes and the significance of gene enrichment in each category was quantified using Fisher's exact test.

Inference of gene regulatory networks

Gene expression ($\log_2$-ratios of tumor to average normal brain) and gene copy number ($\log_2$-ratios of tumor to normal DNA) data of all histologically classified oligodendrogliomas were used to learn gene regulatory networks using the R package regNet [64]. Histologically classified oligodendrogliomas without 1p/19q co-deletion were included to increase the variation of gene expression within the region of the 1p/19q co-deletion to support the selection of relevant links between genes. We randomly divided the oligodendrogliomas into a training set containing two-thirds of the tumors (119) for network inference and a test set containing the remaining one-third of tumors (59) for network validation. For each of the 12,285 genes in our data set, regNet models the expression of each gene as a linear combination of its own gene copy number and the expression of all other genes to determine the most relevant predictors (gene-specific copy number and expression of putative regulators) of each gene [64]. To solve each gene-specific linear model, regNet uses lasso regression [70] in combination with a significance test for lasso [46] to estimate the coefficient and corresponding significance (q-value) for each gene-specific predictor. Lasso regression selects the most relevant predictors of each gene and automatically shrinks the coefficients of other irrelevant predictors to zero. To avoid the inclusion of spurious predictors that only represent the local copy number state but not putative regulatory dependencies between genes, we removed local gene-specific predictors 50 genes down- and up-stream of each gene as done in [65]. We finally only considered the most significant predictors of each gene with a q-value equal or less than 0.01. Network inference was very time consuming (390h CPU time per network). Nevertheless, we repeated the genome-wide network inference ten times with different training sets to integrate evidences from different networks into the prediction of novel tumor gene candidates.

Identification of major regulators

To determine major regulators with many outgoing links to other genes, we defined a scoring scheme that integrates the learned networks. We assume that links that are present in more networks are also more relevant than links that are only found in some networks. First, we counted for each gene $g \in \{1, \ldots, 12285\}$ the number of outgoing links c_{gi} that were observed in $i \in \{1, \ldots, 10\}$ of the 10 networks resulting in a count matrix $C := (c_{gi})$. Next, we standardized each column sum of C to 1 to account for different numbers of outgoing links involved in counting. Finally, we determined for each gene its score by summing up the corresponding gene-specific row values of the standardized count matrix C. Genes with greater score values have more stable outgoing links across the learned networks than genes with lower scores. This ranking of genes enabled to determine major regulators across the networks and to test if genes of a specific annotation class have greater scores than genes that were not part of this class (Wilcoxon rank sum test).

Validation of learned networks

To assess the prediction quality of individual gene expression levels by each network, we computed correlations between network-based predicted and experimentally measured gene expression levels for each of the ten networks considering the corresponding network-specific oligodendroglioma test set. We further utilized each network to predict the expression levels of 118 histologically classified oligoastrocytomas (34 with 1p/19q co-deletion, 84 without 1p/19q co-deletion), a tumor type that is closely related to oligodendrogliomas [38, 48]. In addition, to have baseline models for the different validation data sets, we computed 25 random networks (degree-preserving network permutations) for each of the ten learned networks using regNet [64] to compare their prediction quality to those of the ten original networks. To summarize the prediction results of the different networks, we computed median correlations between predicted and measured expression levels and we further analyzed if the obtained median correlation distribution of the original networks was significantly shifted into the positive range compared to the correlation distribution of the random networks using a Wilcoxon rank sum test.

Network-based impact quantification of gene copy number mutations on signaling and metabolic pathways

We considered all oligodendrogliomas with 1p/19q co-deletion to analyze how differentially expressed genes between tumor and normal brain tissue located within the region of the 1p/19q co-deletion impact on cancer-relevant signaling and metabolic pathways. We used the network propagation algorithm implemented in regNet [64] to realize this. This algorithm considers a learned network and the prediction quality of individual genes to compute direct and indirect impacts between each pair of genes considering all possible network paths (Fig. 1). We have previously shown that this algorithm can correctly predict downstream impacts of gene perturbation experiments [65].

We first computed the total strength of impacts that flow from a differentially expressed gene located within the region of the 1p/19q co-deletion to individual signaling and metabolic pathway genes for each of the ten learned networks. To compare the obtained impacts to random baseline models, we considered the 25 random network instances computed for each of the ten networks to determine the corresponding average impacts of each differentially expressed gene of the 1p/19q region on all

signaling and metabolic genes. We next compared the median impact of each gene under the ten original networks to the corresponding average impacts of this gene under the random networks using a paired one-sided Wilcoxon rank sum test and further corrected for multiple testing by computing q-values [68]. We used a paired test to account for the fact that the random networks that belong to each of the ten individual networks were derived by degree-preserving network permutations. We considered a one-sided test because only genes with greater impact obtained under corresponding random models are of interest. We considered differentially expressed genes within the region of the 1p/19q co-deletion as high-impact genes if they had significantly greater impacts on signaling or metabolic pathways than under corresponding random networks using a q-value cutoff of 0.05.

Moreover, we also analyzed impacts of differentially expressed genes in chromosomal regions that were much less frequently affected by deletions or duplications in oligodendrogliomas. We specifically focused on aberrations of whole chromosomal arms in addition to the characteristic 1p/19q co-deletion. To account for noisy gene copy number measurements, we defined a chromosomal arm to be mutated if at least 80% of its genes were duplicated or deleted, respectively. To validate the considered mutated chromosomal arms, we compared our predictions to those reported for oligodendrogliomas of the POLA cohort [33] and found that they have been previously described (Table 1). We considered each chromosomal arm that was mutated in at least six oligodendrogliomas with 1p/19q co-deletion and defined a corresponding subcohort of oligodendrogliomas that showed these mutations. We considered each subcohort and computed for all differentially expressed genes located on the mutated chromosomal arm corresponding impacts on signaling and metabolic pathway genes as described above.

Results and discussion

Many under- and several overexpressed genes are observed within the region of the 1p/19q co-deletion

We considered all oligodendrogliomas with 1p/19q co-deletion and compared their gene expression profiles to normal brain references to identify differentially expressed genes. We observed 3,068 (23.8%) under- and 3204 (24.9% of genes) overexpressed genes in oligodendrogliomas (q-value < 0.05, Additional file 4: Table S3). Only few strongly underexpressed tumor suppressors (log-ratio < -2: *ANO3* and *CDH1*), but several strongly overexpressed oncogenes (log-ratio > 2: *MYC*, *EGFR*, *PDGFRA*, *PIK3CA*, *PRRX1*, *ASCC3*, *ZNF117*, *CRISPLD1*, *CSMD3*, *ALDH1L2*, *MDGA2*, *TSHR* and *H3F3A*) were among these genes.

Considering chromosomal locations, we found that 524 underexpressed genes (45.2% of genes on 1p/19q) and interestingly also 130 overexpressed genes (11.2% of genes on 1p/19q) were located within the region of the 1p/19q co-deletion. We observed strong underexpression for 74 of the 524 underexpressed genes on 1p/19q (log-ratio < -2). The ten most strongly underexpressed genes on 1p/19q were *LC17A7*, *PRKCG*, *RIMS3*, *KIAA1324*, *AK5*, *SLC6A17*, *CD22*, *HPCA*, *MAG* and *CHD5*. We also observed strong overexpression for 10 of the 130 overexpressed genes on 1p/19q (log-ratio > 2: *SAMD11*, *SLC35E2*, *HES5*, *GRHL3*, *RCC1*, *SPOCD1*, *HFM1*, *DLL3*, *IL4I1* and *CACNG6*). Interestingly, *DLL3* and *HES5* are part of the Notch signaling pathway involved in oligodendrocyte specification [56] restricting cell proliferation and tumor growth in glioma mouse models [22].

We further analyzed all differentially expressed genes in the context of known cancer-relevant signaling pathways (Fig. 2). We observed that especially the Notch and Hedgehog signaling were strongly enriched for overexpressed genes, whereas MAPK signaling was enriched for underexpressed genes (Fig. 2a). In addition, also ErbB signaling and the Adherens junction pathway tended to show an enrichment of underexpressed genes. Considering metabolic pathways, we found that the oxidative phosphorylation pathway was enriched for underexpressed genes (Fig. 2b). Also the pyrimidine, purine and pentose phosphate pathway tended to show some enrichment of differentially expressed genes.

Transcriptional regulatory networks predict tumor gene expression levels

To provide the basis for the impact quantification of gene copy number mutations on cancer-relevant pathways, we used regNet [64] to learn genome-wide transcriptional

Table 1 Statistics of rarely mutated chromosomal arms

	Deletions					Duplications		
Chromosomal arm	4q	9q	13q	15q	18q	7p	7q	11q
TCGA: OD II + III	9.8%	4.5%	10.5%	9.0%	15.0%	6.0%	9.0%	4.5%
TCGA: OD III	13.8%	5.2%	10.3%	12.1%	20.7%	6.9%	10.3%	6.9%
POLA: OD III	16.2%	14.7%	5.9%	14.7%	8.8%	4.4%	7.3%	19.1%

Chromosomal arms affected by deletions (4q, 9q, 13q, 15q and 18q) and duplications (7p, 7q and 11q) in subsets of oligodendrogliomas in addition to the characteristic 1p/19q co-deletion. Percentages of affected oligodendrogliomas are shown for the TCGA cohort (TCGA: OD II + III comprised 133 oligodendrogliomas of WHO grades II and III, TCGA: OD III comprised 58 oligodendrogliomas of WHO grade III) and the POLA cohort [33] (POLA: OD III comprised 68 oligodendrogliomas of WHO grade III)

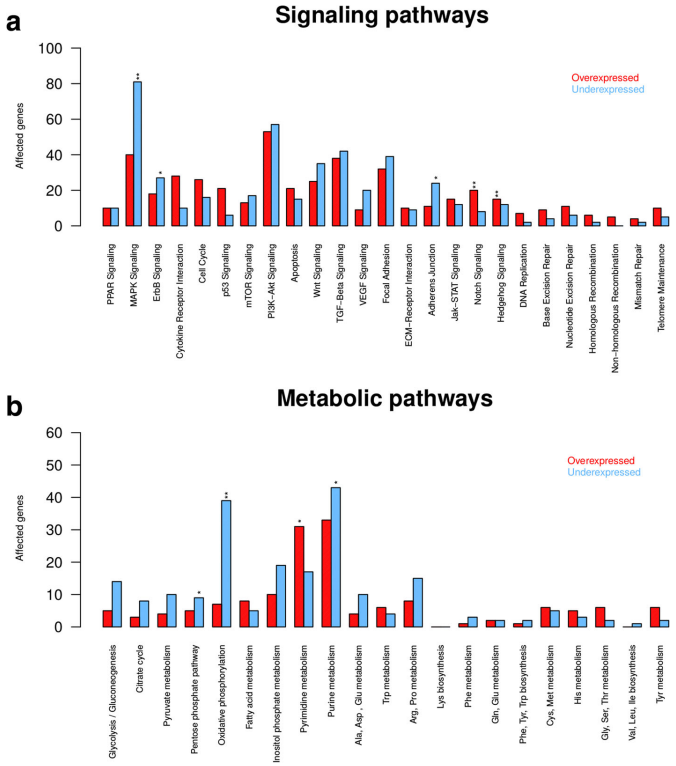

Fig. 2 Signaling and metabolic pathway analysis of differentially expressed genes. Differentially expressed genes between oligodendrogliomas and normal brain references (q-value < 0.05, Additional file 4: Table S3) were mapped to known cancer-relevant signaling (**a**) and metabolic pathways (**b**). The number of over- and underexpressed genes are shown for each pathway. Asterisks symbols highlight pathways enriched for over- or underexpressed genes (Fisher's exact test with '*' for $P < 0.1$ and '**' for $P < 0.05$)

regulatory networks based on gene copy number and expression data of 178 histologically classified oligodendrogliomas with and without 1p/19q co-deletion. We repeated the genome-wide network inference ten times utilizing different training and test data sets (see "Materials and methods" section for details). The resulting networks had on average 67,900 ± 1080 directed links between regulators and target genes (Additional file 3: Figure S3). More than three quarters of these links were activator links (78%) and the others were inhibitor links.

Next, we integrated the outgoing links of each gene across the ten networks to derive a connectivity score that accounts for the co-occurrence of links (see "Materials and methods" for details). This score is higher for genes with more stable outgoing links across n of networks than for genes with less co-occurring links. We utilized these scores and found that tumor suppressor genes, oncogenes, essential genes and signaling pathway genes had significantly greater connectivity scores than genes that were not included in these categories (Wilcoxon rank sum tests: $P = 0.035$ for tumor suppressors, $P = 0.028$ for oncogenes, $P = 5.39 \cdot 10^{-9}$ for essential genes, $P = 0.01$ for signaling pathway genes). The ten genes with the greatest connectivity score were (*GARS*, *CCDC85B*, *NDUFA1*, *SPRED2*, *BIRC6*, *MRPL45*, *EDA2R*, *HMGCS1*, *SLC17A7*, *RAB40B*; Additional file 3: Figure S4). *CCDC85B* is a known downstream target of p53 signaling with reported function as tumor suppressor [26]. Also *SPRED2* is a known tumor suppressor that induces autophagy [32]. *SLC17A7* has

been observed as a tumor suppressor in a glioblastoma stem cell line [44]. *BIRC6* can inhibit apoptosis in glioblastoma cell lines [11]. *RAB40B* is a member of the RAS oncogene family potentially involved in the remodeling of the extracellular matrix during invasion of breast cancer [27]. Other genes like *GARS*, *NDUFA1*, *HMGCS1*, and *SLC17A7* have known functions in cellular metabolism. This clearly indicates that major regulators in our networks are known to have important cancer-relevant functions.

We further tested the capability of each network to predict the expression level of each of the 12,285 genes in independent oligodendroglioma (59 randomly selected tumors left out from network learning) and closely related oligoastrocytoma (118 samples including 34 tumors with and 84 tumors without 1p/19q co-deletion) test sets that were not considered for network inference. To realize this, we computed correlations between originally measured gene expression levels and corresponding network-based predicted gene expression levels across all tumor samples in each test set for each of the ten networks to analyze the prediction quality. Corresponding median gene-specific correlations integrating the prediction results of the ten networks are summarized in Fig. 3 (see Additional file 3: Figure S5 for individual networks). Overall, the vast majority of genes showed strong positive correlations between measured and predicted expression values with a median correlation of 0.75 for the oligodendroglioma test sets and a median correlation of 0.73 for the oligoastrocytoma test set. We also compared these results to predictions of gene expression levels that were obtained from random networks of same complexity as the originally learned networks (degree-preserving network permutations). We found that our networks made significantly better predictions of originally measured gene expression levels than corresponding random networks (Fig. 3, Wilcoxon rank sum test: $P < 2.2 \cdot 10^{-16}$ for each of both test sets).

Genes directly affected by the 1p/19q co-deletion strongly impact on cancer-relevant signaling pathways

We utilized the learned networks to determine impacts of differentially expressed genes located within the region of the 1p/19q co-deletion on known cancer-relevant signaling pathway genes (see Fig. 1 for an illustration). To realize this, we considered the 654 differentially expressed genes observed within the 1p/19q region (524 under- and 130 overexpressed genes with q-values < 0.05, Additional file 4: Table S3) and applied regNet [64] to compute impacts of these genes on the expression of signaling pathway genes using network propagation. We did this independently for each network and computed corresponding impacts for each gene pair under random networks. We further integrated the scores of the ten networks and determined all differentially expressed 1p/19q-genes with significantly greater impacts on the expression of known cancer signaling pathway genes than under random networks (paired Wilcoxon rank sum tests, q-value < 0.05, see "Materials and methods" section for details). Predicted high-impact genes are shown in Fig. 4a and provided in Additional file 5: Table S4. We performed in-depth literature searches and analyzed gene annotations [62] to

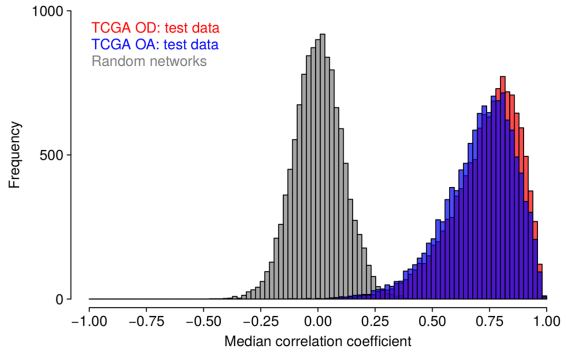

Fig. 3 Network-based prediction quality of gene expression levels. The ten learned gene regulatory networks were analyzed for their performance to predict the expression levels of the 12,285 genes in independent tumor test sets (TCGA OD: 59 network-specific oligodendrogliomas left out from network inference, TCGA OA: 118 closely related oligoastrocytomas). Corresponding histograms of gene-specific median correlations between predicted and measured gene expression levels are shown. The strong shift of both histograms (red, blue) into the positive range shows that the prediction quality of oligodendroglioma-specific networks was significantly better than for random networks (grey) of same complexity (Wilcoxon rank sum tests: $P < 2.2 \cdot 10^{-16}$)

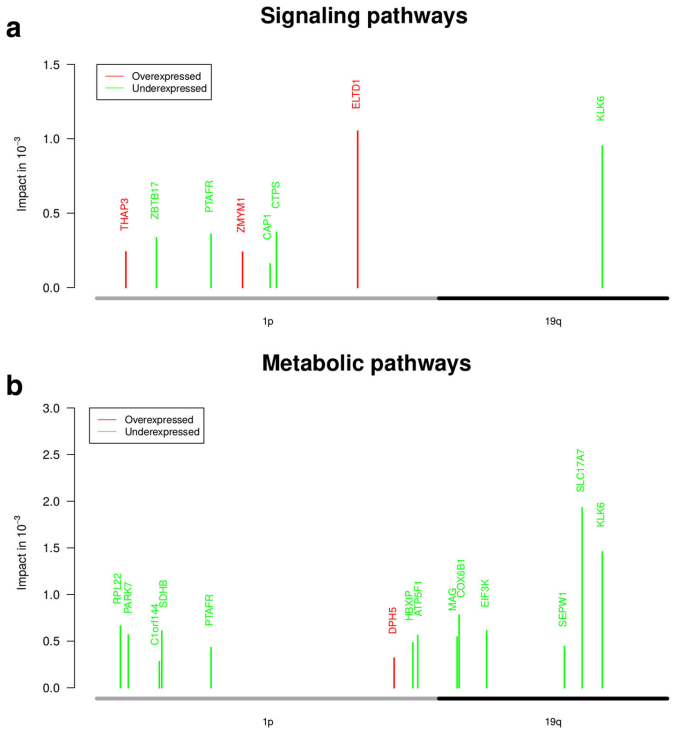

Fig. 4 Genes located in the region of the 1p/19q co-deletion with strong impact on signaling and metabolic pathways. Impacts of differentially expressed genes of the 1p/19q region on the expression of known cancer-relevant signaling pathway genes (**a**) and metabolic pathway genes (**b**). All shown genes had significantly greater impacts under the oligodendroglioma-specific gene regulatory networks than under corresponding random networks of same complexity (q-value ≤ 0.05). The high-impact genes are widespread across the region of the 1p/19q co-deletion. Genes colored in green were underexpressed and genes colored in red were overexpressed in oligodendrogliomas compared to normal brain tissue

characterize cellular functions and known cancer-relevant impacts of these genes.

The gene with the greatest impact on signaling pathway genes was *ELTD1* located on the 1p arm. *ELTD1* encodes for a G-protein coupled receptor. The deletion of one copy of the 1p arm in oligodendrogliomas did not lead to a reduced expression of *ELTD1*. We found *ELTD1* significantly overexpressed in oliogodendrogliomas compared to normal brain tissue. *ELTD1* has been identified to represent a key player of tumor angiogenesis [50]. *ELTD1* has also been functionally validated as oncogene in glioblastomas [72, 86]. The microRNA-139-5p has been reported to act as a tumor suppressor inhibiting *ELTD1* expression in glioblastoma cell lines [15].

The only detected high-impact gene located on the 19q arm with strong impact on signaling pathway genes was *KLK6*. *KLK6* encodes for a serine-protease and was strongly underexpressed in oligodendrogliomas compared to normal brain tissue. High expression levels of *KLK6* have been associated with poor prognosis of intracranial tumors [18] and resistance of glioblastomas to cytotoxic agents [19]. *KLK6* has recently been found to be involved in the control of metastasis formation in colon cancer [66].

PTAFR located on the 1p arm encodes for a G-protein coupled receptor involved in the regulation of cell proliferation and angiogenesis. *PTAFR* was strongly underexpressed in oligodendrogliomas compared to normal brain. *PTAFR* is a putative oncogene and has been reported to

play a role in different types of cancer including the activation of PI3K-Akt signaling in esophageal cancer [12] or the support of prostate cancer development via *ERK1/ERK2* signaling [30].

ZBTB17 located on the 1p arm was underexpressed in oligodendrogliomas compared to normal brain tissue. *ZBTB17* encodes a transcriptional regulator interacting with *MYC*-genes. Reduced expression of *ZBTB17* due to heterozygous loss of 1p36 has been reported to increase the aggressiveness of neuroblastomas [25]. This suggests that *ZBTB17* is a putative tumor suppressor gene.

CAP1 located on the 1p arm was underexpressed in oligodendrogliomas compared to normal brain tissue. *CAP1* is involved in the cyclic AMP pathway and interacts with the actin cytoskeleton influencing cell adhesion [82]. *CAP1* expression has been reported to be positively correlated with proliferation, migration, invasion, and WHO grade of gliomas [3, 21].

So far, no roles in cancer have been reported for the two overexpressed high-impact genes *THAP3* and *ZMYM1* located on the 1p arm. Both genes are likely to encode transcription factors. *THAP3* is involved in the regulation of cell proliferation [51]. *ZMYM1* could be involved in the regulation of the cytoskeletal organization and cell morphology.

Further, we used our network propagation algorithm to predict potential regulatory downstream effects of high-impact genes on individual cancer-relevant signaling pathways (Additional file 3: Figure S6a). Especially the overexpression of *ELTD1* and the underexpression of *PTAFR* in oligodendrogliomas tend to influence the expression of several signaling pathways suggesting complex regulatory dependencies that support or counteract oligodendroglioma growth. Specific impacts of individual high-impact genes are summarized in Additional file 3: Texts S1.

In summary, depending on the expression states in combination with reported roles in cancer, genes like *ELTD1*, *ZBTB17*, or *THAP3* are likely to support oligodendroglioma growth, whereas other genes like *KLK6*, *PTAFR*, or *CAP1* may restrict the speed of tumor growth. This might contribute to the overall better prognosis of oligodendroglioma patients in comparison to patients with other gliomas [69].

Genes directly affected by the 1p/19q co-deletion strongly impact on metabolic pathways

Similar to the analysis of signaling pathways, we used network propagation to identify those differentially expressed genes within the region of the 1p/19q co-deletion that had strong impacts on the expression of metabolic pathway genes. We predicted 14 high-impact genes widespread across the 1p/19q region (q-value < 0.05, Fig. 4b, Additional file 6: Table S5). All genes were underexpressed in oligodendrogliomas compared to normal brain, except for overexpressed *DPH5*. Two genes with strong impact on signaling pathways (*KLK6*, *PTAFR*) were also among these high-impact genes. In contrast to our previous impact quantification for signaling pathways that revealed only one high-impact gene on the 19q arm (Fig. 4a), we now found six genes (*MAG*, *COX6B1*, *EIF3K*, *SEPW1*, *SLC17A7*, *KLK6*) with strong impact on the expression of metabolic pathway genes on 19q (Fig. 4b). We again performed in-depth gene annotation analyses and literature searches to summarize known functions and roles in cancer.

SLC17A7, the gene with the greatest impact on the expression of metabolic pathways, has been observed as tumor suppressor in a glioblastoma stem cell line [44]. *SLC17A7* is located on the 19q arm, encodes for a vesicle-bound, sodium-dependent phosphate transporter expressed in neuron-rich regions, and was strongly underexpressed in oligodendrogliomas compared to normal brain.

SDHB is located on the 1p arm, encodes for the succinate dehydrogenase complex subunit B, and was underexpressed in oligodendrogliomas in comparison to normal brain. Germline mutations of *SDHB* have been reported for patients with head and neck paraganglioma [5] and phaeochromocytomas [7]. Succinate accumulated in *SDHB*-mutated cells inhibits alpha-ketoglutarate-dependent enzymes leading to the activation of hypoxia induced genes and hypermethylation of DNA and histones in paraganglioma [4, 40]. Similarly, a knockdown of *SDHB* in mouse ovarian cancer cells enhanced cell proliferation and induced hypermethylation of histones promoting an epithelial-to-mesenchymal transition [2]. All these findings suggest that reduced expression of *SDHB* in oligodendrogliomas may support and possibly enhance the epigenetic reprogramming via the same pathomechanism induced by a heterozygous *IDH*-mutation that is found in each oligodendroglioma [13, 48].

PARK7 located on the 1p arm was underexpressed in oligodendrogliomas compared to normal brain tissue. *PARK7* encodes for a peptidase that protects cells against oxidative stress. Downregulation of *PARK7* has been associated with a reduction of cell proliferation, migration, and invasion of glioma cell lines [31]. Downregulation of *PARK7* in clear renal cell carcinoma cells increased cisplatin-induced apoptosis [73]. *PARK7* has been reported as oncogene in different cancers activating PI3K-Akt, MAPK, and mTOR signaling to protect cells against hypoxic stress [77].

HBXIP located on the 1p arm was underexpressed in oligodendrogliomas compared to normal brain tissue. *HBXIP* functions as a cofactor of survivin in the suppression of apoptosis [49]. *HBXIP* has been reported to promote the proliferation and migration of breast cancer

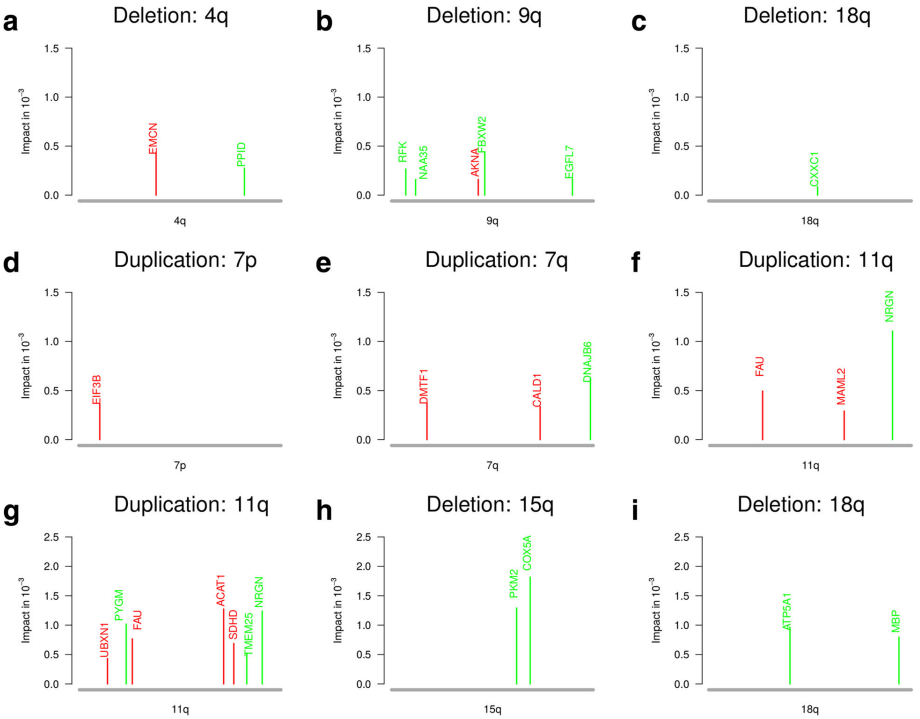

Fig. 5 Genes located on rarely mutated chromosomal arms with strong impact on signaling and metabolic pathways. Impacts of differentially expressed genes of rarely mutated chromosomal arms (Table 1) on the expression of known cancer-relevant signaling pathway genes (**a**–**f**) and metabolic pathway genes (**g**–**i**). All shown genes had significantly greater impacts under the oligodendroglioma-specific gene regulatory networks than under corresponding random networks of same complexity (q-value ≤ 0.1). Genes colored in green were underexpressed and genes colored in red were overexpressed in oligodendrogliomas with the corresponding chromosomal arm mutation in comparison to normal brain tissue

cells [45]. Conversely, suppression of *HBXIP* has been found to reduce cell proliferation, migration and invasion of bladder carcinomas [42]. This suggests that the underexpression of *HBXIP* could counteract oligodendroglioma growth.

SEPW1 is located on 19q, encodes for a selenoprotein that functions as an glutathione antioxidant, and was underexpressed in oligodendrogliomas compared to normal brain. *SEPW1* has been mapped to a putative tumor suppressor region on the 19q arm of gliomas [67]. *SEPW1* has been shown to be involved in the control of cell cycle progression [23] and to regulate expression, activation and degradation of *EGFR* [1].

C1orf144 (*SZRD1*) located on 1p was underexpressed in oligodendrogliomas in comparison to normal brain tissue. *C1orf144* has recently been reported as a potential tumor suppressor in cervical cancer involved in the regulation of cell cycle arrest in G2 and induction of apoptosis [84].

MAG located on 19q was underexpressed in oligodendrogliomas compared to normal brain tissue. *MAG* encodes for a membrane protein involved in myelination of oligodendrocytes, protection of neurons against apoptosis, and inhibition of neurite outgrowth [59].

Further, only *DPH5* located on 1p was overexpressed in oligodendrogliomas compared to normal brain. *DPH5* encodes for a specific methionine-dependent methyltransferase involved in diphthamide synthesis. Diphthamide, a post-transcriptionally modified histidine, is required for eEF-2, which is essential for protein biosynthesis. Further, two underexpressed high-impact genes, *RPL22* and *EIF3K*, known to be important for protein synthesis were found. Strong impacts of genes involved

in protein synthesis might represent a byproduct of increased transcription in tumors. In addition, *COX6B1* located on 19q and *ATP5F1* located on 1p were underexpressed in oligodendrogliomas in comparison to normal brain. Both genes have functions in the respiratory chain.

In addition, we also used our network propagation algorithm to further predict potential regulatory downstream effects of high-impact genes from Fig. 4b on individual metabolic pathways (Additional file 3: Figure S6b). Interestingly, six genes were predicted to contribute to a downregulation of the oxidative phosphorylation. Detailed information to specific impacts of individual high-impact genes are summarized in Additional file 3: Text S1.

Again, genes like *SLC17A7*, *SDHB*, *SEPW1*, or *SZRD1* may support oligodendroglioma growth and other genes like *PARK7* or *HBXIP* may restrict the speed of tumor growth. Such counteracting impacts could contribute to a better prognosis [69].

Impact of rare gene copy number mutations on cancer-relevant signaling and metabolic pathways

We further used our network-based impact quantification strategy to determine if potential candidate genes with high-impact on signaling or metabolic pathways are located on chromosomal arms that were rarely affected by deletions or duplications in oligodendrogliomas with 1p/19q co-deletion (Table 1; deletions: 4q, 9q, 13q, 15q, 18q; duplications: 7p, 7q, 11q; Additional file 3: Figure S1). All these additional mutations have previously been observed in the POLA cohort [33] and several of these mutations were also observed in copy number profiles of single oligodendroglioma cells [71]. These additional copy number mutations occurred more frequently in oligodendrogliomas of WHO grade III than in grade II tumors suggesting that they are associated with tumor progression and may impact on survival [35, 74]. See Additional file 3: Text S2 for further details to subgroups of oligodendrogliomas with additional chromosomal arm mutations. We first determined for each subcohort of oligodendrogliomas with a specific chromosomal arm mutation all differentially expressed genes in comparison to normal brain tissue (q-value $\leq$ 0.05). We next analyzed all differentially expressed genes of a mutated chromosomal arm to identify those genes that were predicted to have a strong impact on the expression of cancer-relevant signaling (Additional file 7: Table S6) and metabolic pathways (Additional file 8: Table S7) utilizing network propagation. We predicted 15 differentially expressed genes with strong impact on signaling pathways on the chromosomal arms 4q, 9q, 7p, 7q, 11q, and 18q (Fig. 5a–f) and 12 genes with strong impact on metabolic pathways on the chromosomal arms 7p, 11q, 15q, and 18q (Fig. 5g–i, 7p not shown) at a q-value cutoff of 0.1 (less stringent than before because of much smaller sample sizes). Functional annotations and literature searches of all predicted high-impact genes are summarized in Additional file 3: Text S3 for signaling pathways and in Additional file 3: Text S4 for metabolic pathways. Next, we only briefly highlight some findings.

Considering genes with high-impact on signaling pathways (Fig. 5a–f), we identified several overexpressed genes in subcohorts of oligodendrogliomas with additional chromosomal arm mutations that were previously found to be involved in tumorigenesis. For example, *EMCN* located on the q-arm of chromosome 4 was overexpressed in oligodendrogliomas with 4q deletion. *EMCN* encodes a glycoprotein that can inhibit adhesion of cells to the extracellular matrix [34]. *EIF3B* located on the p-arm of chromosome 7 was overexpressed in oligodendrogliomas with 7p duplication. *EIF3B* encodes a subunit of the eukaryotic translation initiation factor. A knockdown of *EIF3B* inhibited cell proliferation and increased apoptosis in a glioblastoma cell line [43]. *CALD1* located on the q-arm of chromosome 7 was overexpressed in oligodendrogliomas with 7q duplication. *CALD1* is involved in the regulation of the neovascularization of gliomas [85] and has been associated with tamoxifen resistance of breast cancer [17]. Also *DNAJB6* located on the q-arm of chromosome 7 was overexpressed in oligodendrogliomas with 7q duplication. Overexpression of *DNAJB6* has been reported to promote invasion of colorectal cancer [83]. In addition to these putative oncogenes, we also observed two overexpressed genes with potential tumor suppressor functions that may counteract oligodendroglioma development. *DMTF1* located on 7q encodes a transcription factor with a cyclin D-binding domain that has been shown to inhibit cell growth and cell cycle progression in bladder cancer [57]. Further, *FAU* located on 11q encodes a fusion protein that has been reported to be involved in the regulation of apoptosis of breast cancer [58].

Considering genes with high-impact on metabolic pathways (Fig. 5g–i), we identified four underexpressed genes in subcohorts of oligodendrogliomas with additional chromosomal arm mutations with functions in cellular energy metabolism and impacts on cell migration, apoptosis, or blood vessel development in cancer (deletion of 15q: *COX5A*, *PKM2*; deletion of 18q: *ATP5A1*; duplication of 11q: *PYGM*; Additional file 3: Text S4). In addition, *UBXN1* located on the q-arm of chromosome 11 was overexpressed in oligodendrogliomas. *UBXN1* encodes a ubiquitin-binding protein and has been reported to inhibit the tumor suppressor *BRAC1* [81]. Interestingly, we found that *SDHD* located on 11q was overexpressed in oligodendrogliomas with 11q duplication. This might represent a response to the reduced expression of *SDHB* discussed before. Further, activation of the expression of the tumor suppressor *CDKN1A* in response to the loss of *SDHD* expression has been reported [52]. Thus, overexpressed

SDHD might counteract the expression of *CDKN1A* to support cell proliferation. Moreover, also *ACAT1* located on 11q was overexpressed. *ACAT1* encodes a mitochondrially localized acetyl-CoA acetyltransferase. Inhibition of *ACAT1* by Avasimibe inhibited cell growth by inducing cell cycle arrest and apoptosis in glioblastoma cell lines [6, 55]. Further, inhibition of *ACAT1* has also been shown to suppress growth and metastasis of pancreatic cancer [41].

In-depth analysis of known potential tumor suppressor genes *FUBP1* and *CIC*

We also performed a detailed analysis of the expression behavior and corresponding network-based impacts of the potential tumor suppressors *FUBP1* and *CIC* reported for oligodendrogliomas [8]. *FUBP1* located on 1p and *CIC* located on 19q were both moderately underexpressed in oligodendrogliomas with 1p/19q co-deletion compared to normal brain references (Additional file 4: Table S3). Further, oligodendrogliomas with additional small deletions, insertions or point mutations within *FUBP1* or *CIC* showed moderately reduced expression of these genes compared to oligodendrogliomas without mutations. This trend was much stronger for tumors with *FUBP1* mutations (average expression 4.58 vs. 5.25 comparing 38 tumors with to 95 tumors without mutation, t-test: $P = 0.0001$) than for tumors with *CIC* mutations (average expression 6.42 vs. 6.60 comparing 85 tumors with to 48 tumors without mutation, t-test: $P = 0.01$).

Further, *FUBP1* has been reported to negatively regulate the expression of *MYC* [8]. This relationship was also predicted by our network propagation approach. *FUBP1* had a stronger impact on *MYC* comparing our networks to corresponding random networks (paired Wilcoxon rank test: $P < 0.001$, see "Materials and methods" for details). Globally, *FUBP1* and *CIC* underexpression had moderate impacts on different signaling and metabolic pathways (Additional file 3: Figure S7). Thus, reduced expression of both genes due to the 1p/19q co-deletion could contribute to tumor development, but both genes were not among the predicted putative high impact genes with altered gene expression levels. Still, other pathomechanisms triggered by small deletions, insertions or point mutations within *FUBP1* or *CIC* could play an important role in affected tumors.

Conclusions

The clinical relevance of the 1p/19q co-deletion has been known for many years, but detailed insights to underlying pathomechanisms are not known. Our computational approach provides a novel starting point to characterize molecular changes induced by the 1p/19q co-deletion. We predicted several interesting cancer candidate genes widespread across the region of the 1p/19q co-deletion with strong impact on signaling and metabolic pathways. These candidate genes are possibly involved in the development of oligodendrogliomas. Interestingly, several of these genes (e.g. *ELTD1*, *SDHB*, *SEPW1*, *SLC17A7*, *SZRD1*, *THAP3*, *ZBTB17*) are likely to push, whereas other genes (e.g. *CAP1*, *HBXIP*, *KLK6*, *PARK7*, *PTAFR*) might restrict oligodendroglioma development. This observation could contribute to the fact that oligodendrogliomas have an improved prognosis in comparison to other types of gliomas. Importantly, the overexpression of *ELTD1* in oligodendrogliomas despite the loss of 1p indicates that this gene may act as oncogene as reported for closely related glioblastomas. Similarly, the underexpression of *SLC17A7* in oligodendrogliomas may counteract its known function as tumor suppressor reported for glioblastomas. Moreover, the underexpression of *SDHB* may contribute to the epigenetic reprogramming of oligodendrogliomas via the same pathomechanism as triggered by the *IDH*-mutation. All these findings indicate that several genes located on 1p/19q may simultaneously influence tumor development. Further, we also predicted cancer candidate genes on rarely mutated chromosomal arms that are likely to contribute to oligodendroglioma development and tumor progression in subcohorts of patients. In sum, our computational predictions contribute to a better understanding of the pathology of the 1p/19q co-deletion, might open opportunities for novel experimental studies, and possibly trigger ideas for the development of targeted treatment strategies.

Additional files

Additional file 1: Table S1. Gene copy number data. (XLS 68300 kb)

Additional file 2: Table S2. Gene expression data. (XLS 48300 kb)

Additional file 3: Texts S1–S4 and **Figures S1–S7**. (PDF 1800 kb)

Additional file 4: Table S3. Differentially expressed genes between oligodendrogliomas with 1p/19q co-deletion and normal brain tissue. (XLS 2240 kb)

Additional file 5: Table S4. Genes located on 1p/19q with strong impact on signaling pathways. (XLS 115 kb)

Additional file 6: Table S5. Genes located on 1p/19q with strong impact on metabolic pathways. (XLS 115 kb)

Additional file 7: Table S6. Genes located on rarely mutated chromosomal arms with strong impact on signaling pathways. (XLS 187 kb)

Additional file 8: Table S7. Genes located on rarely mutated chromosomal arms with strong impact on metabolic pathways. (XLS 187 kb)

Abbreviations

1p: p-arm of chromosome 1; 19q: q-arm of chromosome 19; OD: oligodendroglioma; TCGA: The cancer genome atlas; WHO: World health organization

Acknowledgements

This study would have been impossible without the comprehensive data sets made publicly available by the TCGA Research Network. We thank the reviewers for their valuable comments. We thank Chris Lauber (IMB TU Dresden) for support in data acquisition.

Funding
We thank the Else Kröner-Fresenius-Stiftung for financial support of the study within the Else Kröner Promotionskolleg Dresden. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the SLUB/TU Dresden to cover the article processing charge.

Availability of data and materials
Data of all considered TCGA tumors are publicly available from the The Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Processed gene expression and copy number data are contained in the Additional files 1 and 2. Utilized algorithms for network inference and network propagation are publicly available from GitHub (https://github.com/seifemi/regNet).

Authors' contributions
MS designed the study. JG performed the analysis. BK supported the biological interpretation. MS and JG wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
No ethical approval was required for this study. All utilized public omics data sets were generated by others who obtained ethical approval.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [2]Institute for Clinical Genetics, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany. [3]National Center for Tumor Diseases, Dresden, Germany.

Received: 19 April 2018 Accepted: 8 May 2018
Published online: 11 June 2018

References

1. Alkan Z, Duong FL, Hawkes WC (2015) Selenoprotein W controls epidermal growth factor receptor surface expression, activation and degradation via receptor ubiquitination. Biochim Biophys Acta 1853(5):1087–95. https://doi.org/10.1016/j.bbamcr.2015.02.016
2. Aspuria PP, Lunt SY, Väremo L, Vergnes L, Gozo M, Beach JA, et al (2014) Succinate dehydrogenase inhibition leads to epithelial-mesenchymal transition and reprogrammed carbon metabolism. Cancer Metab 2:21. https://doi.org/10.1186/2049-3002-2-21
3. Bao Z, Qiu X, Wang D, Ban N, Fan S, Chen W, et al (2016) High expression of adenylate cyclase-associated protein 1 accelerates the proliferation, migration and invasion of neural glioma cells. Pathol - Res Pract 212(4):264–273. https://doi.org/10.1016/j.prp.2015.12.017
4. Baysal BE, Maher ER (2015) 15 YEARS OF PARAGANGLIOMA: genetics and mechanism of pheochromocytoma-paraganglioma syndromes characterized by germline SDHB and SDHD mutations. Endocr Relat Cancer 22(4):71–82. https://doi.org/10.1530/ERC-15-0226
5. Baysal BE, Willett-Brozick JE, Lawrence EC, Drovdlic CM, Savul SA, McLeod DR, et al (2002) Prevalence of SDHB, SDHC, and SDHD germline mutations in clinic patients with head and neck paragangliomas. J Med Genet 39(3):178–83
6. Bemlih S, Poirier M-D, Andaloussi AE (2010) Acyl-coenzyme A: cholesterol acyltransferase inhibitor Avasimibe affect survival and proliferation of glioma tumor cell lines. Cancer Biol Ther 9(12):1025–1032. https://doi.org/10.4161/cbt.9.12.11875
7. Benn DE, Croxson MS, Tucker K, Bambach CP, Richardson AL, Delbridge L, et al (2003) Novel succinate dehydrogenase subunit B (SDHB) mutations in familial phaeochromocytomas and paragangliomas, but an absence of somatic SDHB mutations in sporadic phaeochromocytomas. Oncogene 22(9):1358–64. https://doi.org/10.1038/sj.onc.1206300
8. Bettegowda C, Agrawal N, Jiao Y, Sausen M, Wood LD, Hruban RH, et al (2011) Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science 333(6048):1453–1455. https://doi.org/10.1126/science.1210557
9. Cairncross JG, Ueki K, Zlatescu MC, Lisle DK, Finkelstein DM, Hammond RR, et al (1998) Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. J Natl Cancer Inst 90(19):1473–9
10. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 164(3):550–563. https://doi.org/10.1016/j.cell.2015.12.028
11. Chakrabarti M, Klionsky DJ, Ray SK (2016) mir-30e blocks autophagy and acts synergistically with proanthocyanidin for inhibition of AVEN and BIRC6 to increase apoptosis in glioblastoma stem cells and glioblastoma SNB19 cells. PLoS ONE 11(7):0158537. https://doi.org/10.1371/journal.pone.0158537
12. Chen J, Lan T, Zhang W, Dong L, Kang N, Zhang S, et al (2015) Platelet-activating factor receptor-mediated PI3k/AKT activation contributes to the malignant development of esophageal squamous cell carcinoma. Oncogene 34(40):5114–5127. https://doi.org/10.1038/onc.2014.434
13. Cohen A, Holmen S, Colman H (2013) IDH1 and IDH2 mutations in gliomas. Curr Neurol Neurosci Rep 13(5):345. https://doi.org/10.1007/s11910-013-0345-4
14. Coons SW, Johnson PC, Scheithauer BW, Yates AJ, Pearl DK (1997) Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. Cancer 79:1381–1393
15. Dai S, Wang X, Li X, Cao Y (2015) MicroRNA-139-5p acts as a tumor suppressor by targeting ELTD1 and regulating cell cycle in glioblastoma multiforme. Biochem Biophys Res Commun 467(2):204–210. https://doi.org/10.1016/j.bbrc.2015.10.006
16. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell 155(4):948–962. https://doi.org/10.1016/j.cell.2013.10.011
17. De Marchi T, Timmermans AM, Smid M, Look MP, Stingl C, Opdam M (2016) Annexin-A1 and caldesmon are associated with resistance to tamoxifen in estrogen receptor positive recurrent breast cancer. Oncotarget 7(3):3098–110. https://doi.org/10.18632/oncotarget.6521
18. Drucker KL, Gianinni C, Decker PA, Diamandis EP, Scarisbrick IA (2015) Prognostic significance of multiple kallikreins in high-grade astrocytoma. BMC Cancer 15:565. https://doi.org/10.1186/s12885-015-1566-5
19. Drucker K, Paulsen AR, Giannini C, Decker PA, Blaber SI, Blaber M, et al (2013) Clinical significance and novel mechanism of action of kallikrein 6 in glioblastoma. Neuro Oncol 15(3):305–318. http://doi.org/10.1093/neuonc/nos313
20. Eisenreich S, Abou-El-Ardat K, Szafranski K, Campos Valenzuela JA, Rump A, Nigro JM, et al (2013) Novel CIC point mutations and an exon-spanning, homozygous deletion identified in oligodendroglial tumors by a comprehensive genomic approach including transcriptome sequencing. PLoS ONE 8(9):e76623. https://doi.org/10.1371/journal.pone.0076623
21. Fan YC, Cui CC, Zhu YS, Zhang L, Shi M, Yu JS (2016) Overexpression of CAP1 and its significance in tumor cell proliferation, migration and invasion in glioma. Oncol Rep 36(3):1619–1625
22. Giachino C, Boulay JL, Ivanek R, Alvarado A, Tostado C, Lugert S (2015) A tumor suppressor function for notch signaling in forebrain tumor subtypes. Cancer Cell 28(6):730–742. https://doi.org/10.1016/j.ccell.2015.10.008
23. Hawkes WC, Alkan Z (2011) Delayed cell cycle progression from SEPW1 depletion is p53- and p21-dependent in MCF-7 breast cancer cells. Biochem Biophys Res Commun 413(1):36–40. https://doi.org/10.1016/j.bbrc.2011.08.032
24. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. Nat Methods 10(11):1108–1115. https://doi.org/10.1038/nmeth.2651
25. Ikegaki N, Gotoh T, Kung B, Riceberg JS, Kim DY, Zhao H, et al (2007) De novo identification of MIZ-1 (ZBTB17) encoding a MYC-interacting zinc-finger protein as a new favorable neuroblastoma gene. Clin Cancer Res 13(20):6001–6009. https://doi.org/10.1158/1078-0432.CCR-07-0071
26. Iwai A, Hijikata M, Hishiki T, Isono O, Chiba T, Shimotohno K (2008) Coiled-coil domain containing 85B suppresses the beta-catenin activity in

a p53-dependent manner. Oncogene 27(11):1520–1526. https://doi.org/10.1038/sj.onc.1210801
27. Jacob A, Linklater E, Bayless BA, Lyons T, Prekeris R (2016) The role and regulation of Rab40b-Tks5 complex during invadopodia formation and cancer cell invasion. J Cell Sci 129(23):4341–4353. https://doi.org/10.1242/jcs.193904
28. Jansen M, Yip S, Louis DN (2010) Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. Lancet Neurol 9(7):717–726. http://doi.org/10.1016/S1474-4422(10)70105-8
29. Jenkins RB, Blair H, Ballman KV, Giannini C, Arusell RM, Law M (2006) A t(1;19)(q10;p10) mediates the combined deletions of 1p and 19q and predicts a better prognosis of patients with oligodendroglioma. Cancer Res 66(20):9852–61. https://doi.org/10.1158/0008-5472.CAN-06-1796
30. Ji W, Chen J, Mi Y, Wang G, Xu X, Wang W (2016) Platelet-activating factor receptor activation promotes prostate cancer cell growth, invasion and metastasis via ERK1/2 pathway. Int J Oncol 49(1):181–188. https://doi.org/10.3892/ijo.2016.3519
31. Jin S, Dai Y, Li C, Fang X, Han H, Wang D (2016) MicroRNA-544 inhibits glioma proliferation, invasion and migration but induces cell apoptosis by targeting PARK7. Am J Transl Res 8(4):1826–37
32. Jiang K, Liu M, Lin G, Mao B, Cheng W, Liu H, et al (2016) Tumor suppressor Spred2 interaction with LC3 promotes autophagosome maturation and induces autophagy-dependent cell death. Oncotarget 7(18):25652–25667. https://doi.org/10.18632/oncotarget.8357
33. Kamoun A, Idbaih A, Dehais C, Elarouci N, Carpentier C, Letouzé E, et al (2016) Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. Nat Commun 7:11263. https://doi.org/10.1038/ncomms11263
34. Kinoshita M, Nakamura T, Ihara M, Haraguchi T, Hiraoka Y, Tashiro K, et al (2001) Identification of human endomucin-1 and -2 as membrane-bound O-sialoglycoproteins with anti-adhesive activity. FEBS Lett 499(1-2):121–126
35. Klink B, Schlingelhof B, Klink M, Stout-Weider K, Patt S, Schrock E (2010) Glioblastomas with oligodendroglial component - common origin of the different histological parts and genetic subclassification. Anal Cell Pathol (Amst) 33(1):37–54. https://doi.org/10.3233/ACP-CLO-2010-0530
36. Knudson AG (2001) Two genetic hits (more or less) to cancer. Nat Rev Cancer 1(2):157–62. https://doi.org/10.1038/35101031
37. Labussiere M, Idbaih A, Wang X-W, Marie Y, Boisselier B, Falet C, et al (2010) All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. Neurology 74(23):1886–1890. https://doi.org/10.1212/WNL.0b013e3181e1cf3a
38. Lauber C, Klink B, Seifert M (2018) Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups. BMC Cancer 18:399. https://doi.org/10.1186/s12885-018-4251-7
39. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 47(2):106–114
40. Letouzé E, Martinelli C, Loriot C, Burnichon N, Abermil N, Ottolenghi C (2013) SDH mutations establish a hypermethylator phenotype in paraganglioma. Cancer Cell 23(6):739–52. https://doi.org/10.1016/j.ccr.2013.04.018
41. Li J, Gu D, Lee SS, Song B, Bandyopadhyay S, Chen S, et al (2016) Abrogating cholesterol esterification suppresses growth and metastasis of pancreatic cancer. Oncogene 35(50):6378–6388. https://doi.org/10.1038/onc.2016.168
42. Li X, Liu S (2016) Suppression of HBXIP reduces cell proliferation, migration and invasion in vitro, and tumorigenesis in vivo in human urothelial carcinoma of the bladder. Cancer Biother Radiopharm 31(9):311–316. https://doi.org/10.1089/cbr.2016.2038
43. Liang H, Ding X, Zhou C, Zhang Y, Xu M, Zhang C, et al (2012) Knockdown of eukaryotic translation initiation factors 3B (EIF3B) inhibits proliferation and promotes apoptosis in glioblastoma cells. Neurol Sci 33(5):1057–62. https://doi.org/10.1007/s10072-011-0894-8
44. Lin B, Lee H, Yoon J-G, Madan A, Wayner E, Tonning S, et al (2015) Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene. Oncotarget 6(7):5369–5381
45. Liu S, Li L, Zhang Y, Zhang Y, Zhao Y, You X (2012) The oncoprotein HBXIP uses two pathways to up-regulate S100A4 in promotion of growth and migration of breast cancer cells. J Biol Chem 287:30228–39. https://doi.org/10.1074/jbc.M112.343947
46. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the lasso. Ann Stat 42(2):413–468. https://doi.org/10.1214/13-AOS1175
47. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al (2007) The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol 114(2):97–109. https://doi.org/10.1007/s00401-007-0243-4
48. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al (2016) The 2016 World health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol 131(6):803–820. https://doi.org/10.1007/s00401-016-1545-1
49. Marusawa H, Matsuzawa S, Welsh K, Zou H, Armstrong R, Tamm I, et al (2003) HBXIP functions as a cofactor of survivin in apoptosis suppression. EMBO J 22:2729–40. http://doi.org/10.1093/emboj/cdg263
50. Masiero M, Simoes FC, Han HD, Snell C, Peterkin T, Bridges E, et al (2013) A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis. Cancer Cell 24(2):229–241. https://doi.org/10.1016/j.ccr.2013.06.004
51. Mazars R, Gonzales-de-Peredo A, Cayrol C, Lavigne AC, Vogel JL, Ortega N (2010) The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between DYT6 and DYT3 dystonias. J Biol Chem 285(18):13364–71. https://doi.org/10.1074/jbc.M109.072579
52. Millán-Uclés A, Diaz-Castro B, Garcia-Flores P, Báez A, Pérez-Simón JA, López-Barneo J, et al (2014) A conditional mouse mutant in the tumor suppressor SdhD gene unveils a link between p21(WAF1/Cip1) induction and mitochondrial dysfunction. PLoS ONE 9(1):85528. https://doi.org/10.1371/journal.pone.0085528
53. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17(5):510–522. https://doi.org/10.1016/j.ccr.2010.03.017
54. Ohgaki H, Kleihues P (2013) The definition of primary and secondary glioblastoma. Clin Cancer Res 19(4):764–772. https://doi.org/10.1158/1078-0432.CCR-12-3002
55. Ohmoto T, Nishitsuji K, Yoshitani N, Mizuguchi M, Yanagisawa Y, Saito H, et al (2015) K604, a specific acyl-CoA:cholesterol acyltransferase 1 inhibitor, suppresses proliferation of U251-MG glioblastoma cells. Mol Med Rep 12(4):6037–42. https://doi.org/10.3892/mmr.2015.4200
56. Park HC, Appel B (2003) Delta-Notch signaling regulates oligodendrocyte specification. Development 130(16):3747–3755
57. Peng Y, Dong W, Lin TX, Zhong GZ, Liao B, Wang B (2015) MicroRNA-155 promotes bladder cancer growth by repressing the tumor suppressor DMTF1. Oncotarget 6(18):16043–58. https://doi.org/10.18632/oncotarget.3755
58. Pickard MR, Green AR, Ellis IO, Caldas C, Hedge VL, Mourtada-Maarabouni M (2009) Dysregulated expression of Fau and MELK is associated with poor prognosis in breast cancer. Breast Cancer Res 11(4):60. https://doi.org/10.1186/bcr2350
59. Quarles RH (2007) Myelin-associated glycoprotein (MAG): past, present and beyond. J Neurochem 100(6):1431–1448. https://doi.org/10.1111/j.1471-4159.2006.04319.x
60. Reifenberger J, Reifenberger G, Liu L, James CD, Wechsler W, Collins VP (1994) Molecular genetic analysis of oligodendroglial tumors shows preferential allelic deletions on 19q and 1p. Am J Pathol 145(5):1175–90
61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47. http://doi.org/10.1093/nar/gkv007
62. Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, et al (2010) GeneCards version 3: the human gene integrator. Database 2010:baq020. http://doi.org/10.1093/database/baq020
63. Santarosa M, Ashworth A (2004) Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. Biochim Biophys Acta 1654(2):105–22. https://doi.org/10.1016/j.bbcan.2004.01.001
64. Seifert M, Beyer A (2018) regNet: An R package for network-based propagation of gene expression alterations. Bioinformatics 34(2):308–311. http://doi.org/10.1093/bioinformatics/btx544
65. Seifert M, Friedrich B, Beyer A (2016) Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. Genome Biol 17:204. https://doi.org/10.1186/s13059-016-1058-1

66. Sells E, Pandey R, Chen H, Skovan BA, Cui H, Ignatenko NA (2017) Specific microRNA–mRNA regulatory network of colon cancer invasion mediated by tissue kallikrein–related peptidase 6. J R Stat Soc Ser B 19(5):396–411. https://doi.org/10.1016/j.neo.2017.02.003
67. Smith JS, Tachibana I, Pohl U, Lee HK, Thanarajasingam U, Portier BP, et al (2000) A transcript map of the chromosome 19q-arm glioma tumor suppressor region. Genomics 64(1):44–50. https://doi.org/10.1006/geno.1999.6101
68. Storey JD (2002) A direct approach to false discovery rates. J R Statist Soc B 64(3):479–498. https://doi.org/10.1186/s13059-016-1058-1
69. (2015) The Cancer Genome Atlas Research Network: Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 372(26):2481–2498. https://doi.org/10.1056/NEJMoa1402121
70. Tibshirani R (1994) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
71. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al (2016) Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 539(539):309–313. https://doi.org/10.1038/nature20123
72. Towner RA, Jensen RL, Colman H, Vaillant B, Smith N, Casteel R, et al (2013) ELTD1, a potential new biomarker for gliomas. Neurosurgery 72(1):77–91. https://doi.org/10.1227/NEU.0b013e318276b29d
73. Trivedi R, Dihazi GH, Eltoweissy M, Mishra DP, Mueller GA, Dihazi H (2016) The antioxidant protein PARK7 plays an important role in cell resistance to cisplatin-induced apoptosis in case of clear cell renal cell carcinoma. Eur J Pharmacol 784:99–110. https://doi.org/10.1016/j.ejphar.2016.04.014
74. Trost D, Ehrle rM, Fimmers R, Felsberg J, Sabel MC, Kirsch L, et al (2007) Identification of genomic aberrations associated with shorter overall survival in patients with oligodendroglial tumors. Int J Cancer 120(11):2368–76. https://doi.org/10.1002/ijc.22574
75. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al (2012) IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature 483(7390):479–483. https://doi.org/10.1038/nature10866
76. van den Bent MJ (2010) Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathol 120(3):297–304. https://doi.org/10.1007/s00401-010-0725-7
77. Vasseur S, Afzal S, Tardivel-Lacombe J, Park DS, Iovanna JL, Mak TW (2009) DJ-1/PARK7 is an important mediator of hypoxia-induced cellular responses. Proc Natl Acad Sci 106:1111–6. https://doi.org/10.1073/pnas.0812745106
78. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al (2017) Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science 355(6332):8478. https://doi.org/10.1126/science.aai8478
79. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al (2016) gplots: Various R programming tools for plotting data. R package gplots version 3.0.1
80. Wesseling P, van den Bent M, Perry A (2015) Oligodendroglioma: pathology, molecular mechanisms and markers. Acta Neuropathol 129(6):809–27. https://doi.org/10.1007/s00401-015-1424-1
81. Wu-Baer F, Ludwig T, Baer R (2010) The UBXN1 protein associates with autoubiquitinated forms of the BRCA1 tumor suppressor and inhibits its enzymatic function. Mol Cell Biol 30(11):2787–2798. https://doi.org/10.1128/MCB.01056-09
82. Zhang H, Ghai P, Wu H, Wang C, Field J, Zhou G-L (2013) Mammalian adenylyl cyclase-associated protein 1 (CAP1) regulates cofilin function, the actin cytoskeleton, and cell adhesion. J Biol Chem 288(29):20966–20977. https://doi.org/10.1074/jbc.M113.484535
83. Zhang TT, Jiang YY, Shang L, Shi ZZ, Liang JW, Wang Z (2015) Overexpression of DNAJB6 promotes colorectal cancer cell invasion through an IQGAP1/ERK-dependent signaling pathway. Mol Carcinog 54:1205–13. https://doi.org/10.1002/mc.22194
84. Zhao N, Zhang G, He M, Huang H, Cao L, Yin A (2017) SZRD1 is a novel protein that functions as a potential tumor suppressor in cervical cancer. J Cancer 8(11):2132–2141. https://doi.org/10.7150/jca.18806
85. Zheng PP, Sieuwerts AM, Luider TM, van der Weiden M, Sillevis-Smitt PA, Kros JM (2004) Differential expression of splicing variants of the human caldesmon gene (CALD1) in glioma neovascularization versus normal brain microvasculature. Am J Pathol 164(6):2217–28. http://doi.org/10.1016/S0002-9440(10)63778-9
86. Ziegler J, Pody R, Coutinho de Souza P, Evans B, Saunders D, Smith N, et al (2017) ELTD1, an effective anti-angiogenic target for gliomas: preclinical assessment in mouse GL261 and human G55 xenograft glioma models. Neuro-Oncology 19(2):175–185. http://doi.org/10.1093/neuonc/now147

4.7 Publication:

Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse

Journal: PLoS Computational Biology

Received: 13 March 2019; **Accepted**: 5 October 2019; **Published**: 4 November 2019

Citation: Michael Seifert, Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner, Barbara Klink and Anna Dubrovska (2019): Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse, PLoS Comput Biol, 15(11):e1007460.

Placement and summary of the publication

Radiotherapy is an important and effective treatment option for prostate cancer. Up to 90% of prostate cancer patients can be cured by irradiation (Johansson et al. (2012); Pahlajani et al. (2012); Zietman et al. (2010)), but the delivery of a tumor curative radiation dose is limited by radiation-induced normal tissue toxicity (Bonkhoff (2012)). Therefore, local recurrence of prostate cancer after radiotherapy can be attributed to radioresistance of cancer cells (Chang et al. (2014)). Molecular mechanisms that contribute to radioresistance of prostate cancer are only partly understood (Di Lorenzo et al. (2011); Chang et al. (2014); Barker et al. (2015); McAllister et al. (2019)). Thus, the occurrence of radioresistance is highly unpredictable leading to less effective treatments for many patients supporting local recurrence and metastasis of prostate cancer (Chaiswing et al. (2018)).

Prostate cancer cell lines are frequently considered as model system to compare radioresistant to radiosensitive cells with the goal to identify genes and molecular mechanisms involved in radioresistance of prostate cancer (e.g. Cojoc et al. (2015); Peitzsch et al. (2016)). Only very few radioresistant prostate cancer cell lines are usually analyzed in such studies, but their genomes are typically characterized by large chromosomal deletions and amplifications as a consequence of error-prone DNA repair of double strand breaks induced by irradiation (Mateo et al. (2017)). Thus, hundreds or thousands of genes can be located in these altered DNA regions that further differ between different radioresistant cells. This complex situation does not

allow a straightforward prediction of radioresistance driver genes by standard approaches for gene copy number and expression analysis without prior knowledge about involved molecular processes (Seifert et al. (2016)).

This motivated us to developed a network-based approach for the analysis of prostate cancer cell lines with acquired radioresistance to identify clinically relevant marker genes associated with radioresistance of prostate cancer patients. We utilized my R package regNet (Seifert and Beyer (2018)) with the underlying mathematical theory for network inference and network propagation developed in Seifert et al. (2016) to realize this.

In our study, we first compared gene copy number and gene expression profiles of radioresistant and radiosensitive DU145 and LNCaP prostate cancer cell lines that have previously been established in the Dubrovska laboratory (Cojoc et al. (2015); Peitzsch et al. (2016)). We observed that radioresistant DU145 showed much more gene copy number alterations than LNCaP and that their gene expression profiles were highly cell line specific. Next, we learned a genome-wide prostate cancer-specific gene regulatory network based on publicly available gene expression and gene copy number profiles of prostate cancer patients from TCGA (Cancer Genome Atlas Research Network (2015)). We further used this network to quantify impacts of differentially expressed genes with directly underlying copy number alterations in radioresistant DU145 and LNCaP on known radioresistance marker genes. This enabled us to reveal ten potential driver candidates from DU145 (ADAMTS9, AKR1B10, CXXC5, FST, FOXL1, GRPR, ITGA2, SOX17, STARD4, VGF) and four from LNCaP (FHL5, LYPLAL1, PAK7, TDRD6) that were able to distinguish irradiated prostate cancer patients into early and late relapse groups. Moreover, in-depth *in vitro* validations for VGF showed that siRNA-mediated gene silencing increased the radiosensitivity of DU145 and LNCaP cells.

Overall, our computational approach enabled to predict novel radioresistance driver gene candidates for prostate cancer. Additional studies are necessary to further validate the role of VGF and other candidate genes as potential biomarkers for the prediction of radiotherapy responses and as potential targets for radiosensitization of prostate cancer.

Author contribution

I developed the concept of the study together with Anna Dubrovska. I realized all computational analyses, wrote the manuscript, created all figures and revised the manuscript. I discussed the findings with Claudia Peitzsch and Anna Dubrovska, which supported the biological interpretation of the results. Barbara Klink measured the gene copy number profiles. Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner and Anna Dubrovska performed the VGF validation experiments and provided methodological details for the methods section of the manuscript. Anna Dubrovska further contributed parts to the introduction and discussion.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse

Michael Seifert[1,2]*, **Claudia Peitzsch**[2,3], **Ielizaveta Gorodetska**[3], **Caroline Börner**[3], **Barbara Klink**[4], **Anna Dubrovska**[3,5,6]

1 Institute for Medical Informatics and Biometry (IMB), Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany, **2** National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany, **3** OncoRay - National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, **4** Institute for Clinical Genetics, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany, **5** Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Institute of Radiooncology-OncoRay, Dresden, Germany, **6** German Cancer Consortium (DKTK) Partner Site Dresden, Germany, and German Cancer Research Center (DKFZ), Heidelberg, Germany

* michael.seifert@tu-dresden.de

OPEN ACCESS

Citation: Seifert M, Peitzsch C, Gorodetska I, Börner C, Klink B, Dubrovska A (2019) Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse. PLoS Comput Biol 15(11): e1007460. https://doi.org/10.1371/journal.pcbi.1007460

Editor: Matthew J. Lazzara, University of Virginia, UNITED STATES

Received: March 13, 2019

Accepted: October 5, 2019

Published: November 4, 2019

Data Availability Statement: All used data sets and algorithms are publicly available. Gene copy number and gene expression data of DU145 and LNCaP are contained in S1 Table and in S2 Table, respectively. Raw aCGH and gene expression data have been deposited in the Gene Expression Omnibus (GEO) database, accession no GSE134500. TCGA prostate cancer data are available from https://portal.gdc.cancer.gov. Network-based computations were done using the

Abstract

Radiation therapy is an important and effective treatment option for prostate cancer, but high-risk patients are prone to relapse due to radioresistance of cancer cells. Molecular mechanisms that contribute to radioresistance are not fully understood. Novel computational strategies are needed to identify radioresistance driver genes from hundreds of gene copy number alterations. We developed a network-based approach based on lasso regression in combination with network propagation for the analysis of prostate cancer cell lines with acquired radioresistance to identify clinically relevant marker genes associated with radioresistance in prostate cancer patients. We analyzed established radioresistant cell lines of the prostate cancer cell lines DU145 and LNCaP and compared their gene copy number and expression profiles to their radiosensitive parental cells. We found that radioresistant DU145 showed much more gene copy number alterations than LNCaP and their gene expression profiles were highly cell line specific. We learned a genome-wide prostate cancer-specific gene regulatory network and quantified impacts of differentially expressed genes with directly underlying copy number alterations on known radioresistance marker genes. This revealed several potential driver candidates involved in the regulation of cancer-relevant processes. Importantly, we found that ten driver candidates from DU145 (*ADAMTS9*, *AKR1B10*, *CXXC5*, *FST*, *FOXL1*, *GRPR*, *ITGA2*, *SOX17*, *STARD4*, *VGF*) and four from LNCaP (*FHL5*, *LYPLAL1*, *PAK7*, *TDRD6*) were able to distinguish irradiated prostate cancer patients into early and late relapse groups. Moreover, in-depth *in vitro* validations for *VGF* (Neurosecretory protein VGF) showed that siRNA-mediated gene silencing increased the radiosensitivity of DU145 and LNCaP cells. Our computational approach enabled to predict novel radioresistance driver gene candidates. Additional preclinical and clinical studies are required to further validate the role of *VGF* and other candidate genes as

PLOS COMPUTATIONAL BIOLOGY

R package regNet available at https://github.com/seifemi/regNet under GNU GPL-3.

Funding: Studies in the Dubrovska lab were partially supported by grants from the German Research Foundation (DFG) (273676790, 401326337 and 416001651), the Wilhelm Sander-Stiftung (2017.106.1), the DLR Project Management Agency (01DK17047), the German Federal Ministry of Education and Science (BMBF) (03Z1NN11), and the German Cancer Consortium (DKTK) partner site Dresden. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

potential biomarkers for the prediction of radiotherapy responses and as potential targets for radiosensitization of prostate cancer.

Author summary

Prostate cancer cell lines represent an important model system to characterize molecular alterations that contribute to radioresistance, but irradiation can cause deletions and amplifications of DNA segments that affect hundreds of genes. This in combination with the small number of cell lines that are usually considered does not allow a straight-forward identification of driver genes by standard statistical methods. Therefore, we developed a network-based approach to analyze gene copy number and expression profiles of such cell lines enabling to identify potential driver genes associated with radioresistance of prostate cancer. We used lasso regression in combination with a significance test for lasso to learn a genome-wide prostate cancer-specific gene regulatory network. We used this network for network flow computations to determine impacts of gene copy number alterations on known radioresistance marker genes. Mapping to prostate cancer samples and additional filtering allowed us to identify 14 driver gene candidates that distinguished irradiated prostate cancer patients into early and late relapse groups. In-depth literature analysis and wet-lab validations suggest that our method can predict novel radioresistance driver genes. Additional preclinical and clinical studies are required to further validate these genes for the prediction of radiotherapy responses and as potential targets to radiosensitize prostate cancer.

Introduction

Radiation therapy and surgery with or without anti-androgen treatment are key therapies for prostate carcinoma. Depending on the stage of tumor and type of applied irradiation, up to 90% of prostate cancer patients can be permanently cured by radiotherapy [1–3]. Nevertheless, normal tissue toxicity limits the delivery of a tumor curative radiation dose and is therefore one of the major obstacles to effective external beam radiotherapy [4]. Local recurrence of prostate cancer after radiotherapy can be attributed to radioresistance of cancer cells [5]. Molecular mechanisms and cellular properties that contribute to radioresistance of prostate cancer are only partly understood involving activations of signaling pathways such as PI3K/Akt and mTOR, alterations of DNA repair pathways, autophagy, and epithelial-mesenchymal transition, and the potential existence of cancer stem cells [5]. Another important factor involved in radioresistance of prostate cancer is the tumor microenvironment [6, 7]. Tumor progression and therapy response can be influenced by changes of the tumor microenvironment as a consequence of a radiation therapy [8, 9]. Closely related to this are immunomodulatory alterations triggered by radiation therapies that offer possibilities for new treatment options [10–12]. Also changes of the metabolism of cancer cells after a radiotherapy can alter the radiosensitivity of cells [13]. Still, the occurrence of radioresistance is highly unpredictable leading to less effective treatments for many patients supporting local recurrence and metastasis of prostate cancer [14]. Adjuvant therapies to further improve the efficiency of radiation therapies are urgently needed. Different molecular mechanisms and various agents have already been identified to improve the radiosensitization of prostate cancer. This includes androgen deprivation therapy, vascular endothelial growth factor (VEGF) inhibition, mTOR

inhibition, and cytochrome P450 enzyme CYP17A1 inhibition [15]. Several other potential adjuvant strategies have also been suggested including the application of a Bcl-2 inhibitor [16], cytolethal distending toxin [17], PARP inhibition [18], resveratrol [19], and an ATM kinase inhibitor [20] to improve radiosensitization. However, additional molecular characterizations and studies are necessary to enable a targeted transfer into the clinics to further improve the efficiency of radiation therapies.

Still, clinical, pathological and biological factors for the prediction of treatment outcomes are of great importance for the personalization of prostate cancer treatment. The current pre-treatment risk stratification system for prostate cancer is based on the analysis of prostate-specific antigen, clinical T-stage and Gleason scores to guide medical decision making [21]. This concept for risk assessment of prostate cancer is of a high clinical value, but limited by the heterogeneity of patients within disease-risk groups [22]. Therefore, novel prognostic factors are required to obtain more accurate risk estimations for radioresistance.

Over the last years, different large-scale studies were performed to obtain a better general characterization of prostate cancer at the molecular level. This has contributed to the identification of molecular subtypes, recurrent gene mutations and DNA copy number alterations, and the characterization of signaling and DNA repair pathways involved in the development of prostate cancer (e.g. [23–26]). Especially the multi-omics study by The Cancer Genome Atlas (TCGA) [23] provides omics profiles of different molecular layers along with clinical information for hundreds of prostate cancer patients. Such data sets represent an important basis to gain novel insights into genes and molecular mechanisms driving radioresistance, but this search for novel candidate genes is very challenging comparable to the search for the needle in the haystack.

Irradiation of prostate cancer cells causes DNA double strand breaks and cells that survive this highly toxic damage can show complex genomic alterations such as large deletions or amplifications of chromosomal regions due to error-prone DNA repair [27]. Many genes are located in such altered regions and these altered regions differ substantially between radioresistant cells. Therefore, an identification of radioresistance drivers by standard statistical approaches is nearly impossible. Innovative computational concepts are required to separate potential drivers from passengers. A promising strategy is the analysis of gene dosage effects triggered by underlying deletions and amplifications with the help of gene regulatory networks [28–30]. This strategy is related to network-based stratification of gene mutations [31, 32]. We recently demonstrated that gene regulatory networks learned from gene expression and copy number profiles of cancer cell lines or cancer patients are capable to predict impacts of gene copy number alterations on cancer-relevant target genes, signaling pathways and patient survival [28–30]. The key principle behind this approach is the usage of a specifically designed network propagation algorithm to propagate gene expression alterations along the edges of a gene regulatory network to quantify how individual gene copy number alterations influence the expression of other genes. This concept can be adapted to the analysis of radioresistant prostate cancer cell lines offering the great opportunity to identify novel candidate genes involved in radioresistance.

Here, we developed an approach for the network-based analysis of prostate cancer cell lines with acquired radioresistance to identify clinically relevant marker genes associated with radioresistance in prostate cancer patients (Fig 1). We considered the existing prostate cancer cell lines DU145 (androgen-independent with high metastatic potential derived from a brain metastasis) and LNCaP (androgen-dependent with low metastatic potential derived from a lymph node metastasis) and analyzed molecular data of radiosensitive parental cells and corresponding radioresistant cells that we had established in [33] and which we had further analyzed in [34]. We compared gene copy number and expression profiles of the radioresistant

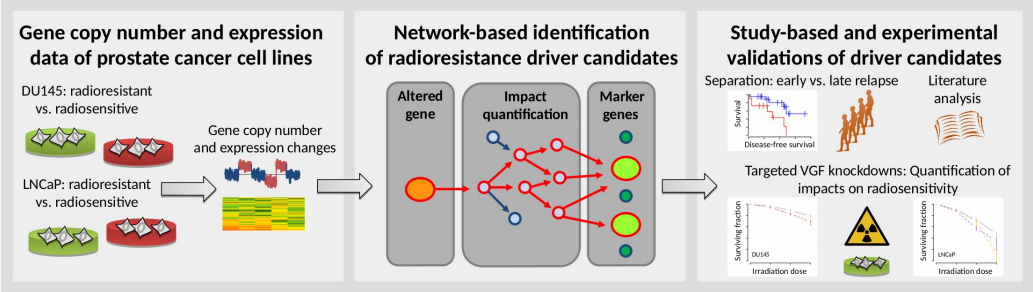

Fig 1. Methodological overview. Left box, Prostate cancer cell lines DU145 and LNCaP were purchased from the American Type Culture Collection and used to establish radioresistant cell lines. Gene copy number and expression profiles of radioresistant and corresponding age-matched non-irradiated radiosensitive parental cell lines were measured. Middle box, A prostate cancer-specific gene regulatory network was learned from gene expression and copy number data from 541 prostate cancer patients from The Cancer Genome Atlas (TCGA) and validated on 768 cancer cell lines of the Cancer Cell Line Encyclopedia (CCLE). This network was used to quantify putative impacts of genes with differential expression and directly underlying copy number alterations between radioresistant and radiosensitive cell lines (orange circle) on known marker genes of radioresistance (green circles) utilizing network propagation (red arrows). Right box, Identified potential radioresistance driver genes were evaluated for their potential to separate irradiated prostate cancer patients from TCGA into early and late relapse groups. In-depth literature analysis was done for all cell line-based candidate genes that were predictive for the relapse behavior of irradiated prostate cancer patients. Sophisticated experimental validations were done for the candidate gene *VGF* by analyzing the impact of siRNA-based *VGF* knockdowns on radiosensitivity. A detailed technical flow chart is shown in S1 Fig.

https://doi.org/10.1371/journal.pcbi.1007460.g001

cell lines to their radiosensitive parental cells and further utilized our network-based approach to quantify the impact of differentially expressed genes with directly underlying copy number alterations on known marker genes of radioresistance. We identified several novel gene candidates that are potentially involved in the manifestation of radioresistance enabling to separate prostate cancer patients treated with radiotherapy into early and late relapse groups. We performed in-depth wet-lab validations of a selected candidate gene (*VGF*: Neurosecretory protein VGF) providing further evidence that our computational approach can contribute to the identification of genes involved in radioresistance.

Results

DU145 shows more gene copy number alterations than LNCaP

We considered radioresistant cell lines of the prostate cancer cell lines DU145 and LNCaP that were established in [33] and further characterized in [34]. We analyzed corresponding array-based comparative genomic hybridization (aCGH) experiments to identify gene copy number alterations distinguishing radioresistant DU145 and LNCaP from their radiosensitive parental cell line (Fig 1, S1 Table). Generally, radioresistant DU145 showed more copy number alterations than radioresistant LNCaP (Fig 2). In more detail, comparing radioresistant to radiosensitive DU145, 24.8% of genes (6,109 of 24,625) had reduced and 38.6% (9,498 of 24,625) had increased copy numbers (Fig 2a, S2 Table), whereas only 3.1% (765 of 24,625) of genes had reduced and 1.5% (377 of 24,625) had increased copy numbers comparing radioresistant to radiosensitive LNCaP (Fig 2b, S2 Table). For DU145, broad segments of gene copy number alterations across all chromosomes and few focal gene copy number alterations were observed (Fig 2a). In contrast, LNCaP only showed some broad regions of reduced gene copy numbers on chromosomes 1, 6, and 20, greater gene copy numbers for a broad region on chromosome 12, and some focal gene copy number alterations on different chromosomes (Fig 2b). Both cell lines further showed a significant overlap of 389 genes with reduced gene copy numbers

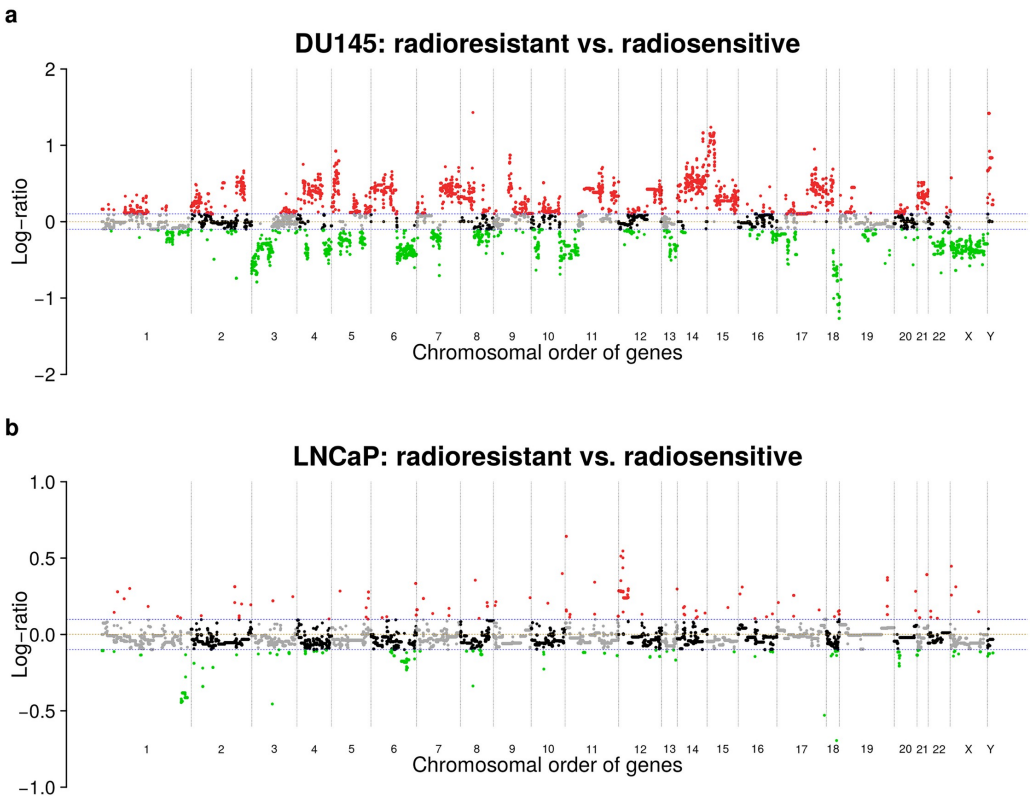

Fig 2. Gene copy number alterations of DU145 and LNCaP. Gene copy number profiles of DU145 (a) and LNCaP (b) comparing radioresistant to radiosensitive cell lines. Gene copy number alterations are quantified by $\log_2$-ratios of radioresistant versus radiosensitive and plotted in the chromosomal order of genes. Deviations of $\log_2$-ratios from zero (brown dashed line) indicate the presence of gene copy number alterations. Considered reduced (green dots below blue dashed line: $\log_2$-ratios $<$ -0.1) or increased (red dots above blue dashed line: $\log_2$-ratios $>$ 0.1) gene copy numbers in the corresponding radioresistant cell lines of DU145 and LNCaP are highlighted. Ends of chromosomes are marked by black dotted vertical lines. Unchanged genes on a chromosome are shown by alternating grey and black dots to further support the visual separation between chromosomes. An additional heatmap representation including comparisons of radioresistant and radiosensitive DU145 and LNCaP to normal reference DNA is shown in S2 Fig.

https://doi.org/10.1371/journal.pcbi.1007460.g002

(mainly on chromosomes 1 and 6, Fisher's exact test: $P = 7.16 \cdot 10^{-56}$) and a non-significant overlap of 47 genes with increased copy numbers widespread across the genome.

We also compared the gene copy number alterations of radioresistant and radiosensitive DU145 and LNCaP to normal male reference DNA to better understand the observed differences between both cell lines (S2 Fig). We found that radiosensitive DU145 had much more gene copy number alterations than radiosensitive LNCaP, radioresistant DU145 and LNCaP were clearly more similar to their corresponding radiosensitive counterpart than to each other, and radioresistant DU145 had much more gene copy number alterations than radioresistant LNCaP. These findings indicate that DU145 is generally more prone to DNA copy number alterations than LNCaP, which could explain the strong differences observed between both cell

lines. An increased radioresistance of DU145 in comparison to LNCaP can also contribute to these observations [33].

DU145 and LNCaP mainly show cell line specific expression patterns

We analyzed gene expression for DU145 and LNCaP to identify differentially expressed genes between established radioresistant DU145 and LNCaP and their radiosensitive parental cell line (Fig 1). We used a Hidden Markov Model to determine differentially expressed genes [35] (see Methods). We found 857 under- and 835 overexpressed genes in radioresistant DU145 and 855 under- and 670 overexpressed genes in radioresistant LNCaP compared to their radiosensitive parental cell lines (S3 Table). The overlap of differentially expressed genes between both cell lines was small but significant (Fig 3a: 81 under- and 51 overexpressed genes, $P < 7.46 \cdot 10^{-8}$, Fisher's exact test, S3 Table). The majority of these genes was part of signaling pathways and/or encode for transcription factors/co-factors (Fig 3b). Commonly underexpressed genes in radioresistant DU145 and LNCaP included e.g. known tumor suppressors (e.g. *BCL10*, *EPB41L4A*, *SPRED1*, *SERPINB5*) and commonly overexpressed genes included e.g. *SEMA4A* involved in cell-cell signaling and migration, *NROB1* associated with stem cell pluripotency, and genes involved in cytokine signaling (e.g. *IL19*, *IL3RA*) [36].

We found similar patterns of differential expression among known cancer-relevant signaling pathways for both cell lines (Fig 3). Further, radioresistant DU145 and LNCaP showed an enrichment of overexpressed PI3K-Akt pathway genes (Fig 3d). DU145 also showed an enrichment of underexpressed genes for the cytokine pathway, the p53 pathway, the PI3K-Akt pathway, and the Jak-STAT pathway (Fig 3c) and an enrichment of overexpressed genes for the cytokine pathway, the ECM receptor pathway, the focal adhesion pathway, and the hedgehog pathway (Fig 3d). LNCaP showed an enrichment of underexpressed TGF-Beta signaling genes (Fig 3c). Most of these pathways have already been associated with radioresistance of prostate cancer and other types of cancers (e.g. [5, 14, 37–39]).

Direct impact of copy number alterations on expression of affected genes

We analyzed which genes with copy number alterations showed altered expression. LNCaP showed more gene expression alterations than gene copy number alterations (1,525 vs. 1,142) and only 8.9% (102 of 1,142) of genes with copy number alterations showed altered expression. 66 of these 102 genes showed putative direct impacts of the underlying copy number alteration on the expression level (S4 Table; 49 genes with reduced copy number and decreased expression; 17 genes with increased copy number and expression). These findings are similar to a related analysis of radiosensitive and radioresistant subclones of a head and neck squamous cell carcinoma cell line that only found few differentially expressed genes with directly underlying copy number alterations [40]. Further, tumor suppressor genes such as *PRDM1* and *RNF217* had a reduced copy number and showed reduced expression in radioresistant compared to radiosensitive LNCaP.

In contrast, we found substantially more gene copy number alterations than gene expression alterations for DU145 (15,607 vs. 1,692), but only 7.3% (1,144 of 15,607) of genes with altered copy numbers also showed altered expression. 447 of these 1,144 genes showed expression changes in the same direction (S4 Table; 191 genes with reduced copy number and reduced expression; 256 genes with increased copy number and increased expression), whereas the other genes had expression differences in the opposite direction possibly due to the complex genomic rearrangements observed for DU145 affecting many transcriptional regulators (Figs 2a and 3b). These findings are supported by our previous analysis of DU145 [34].

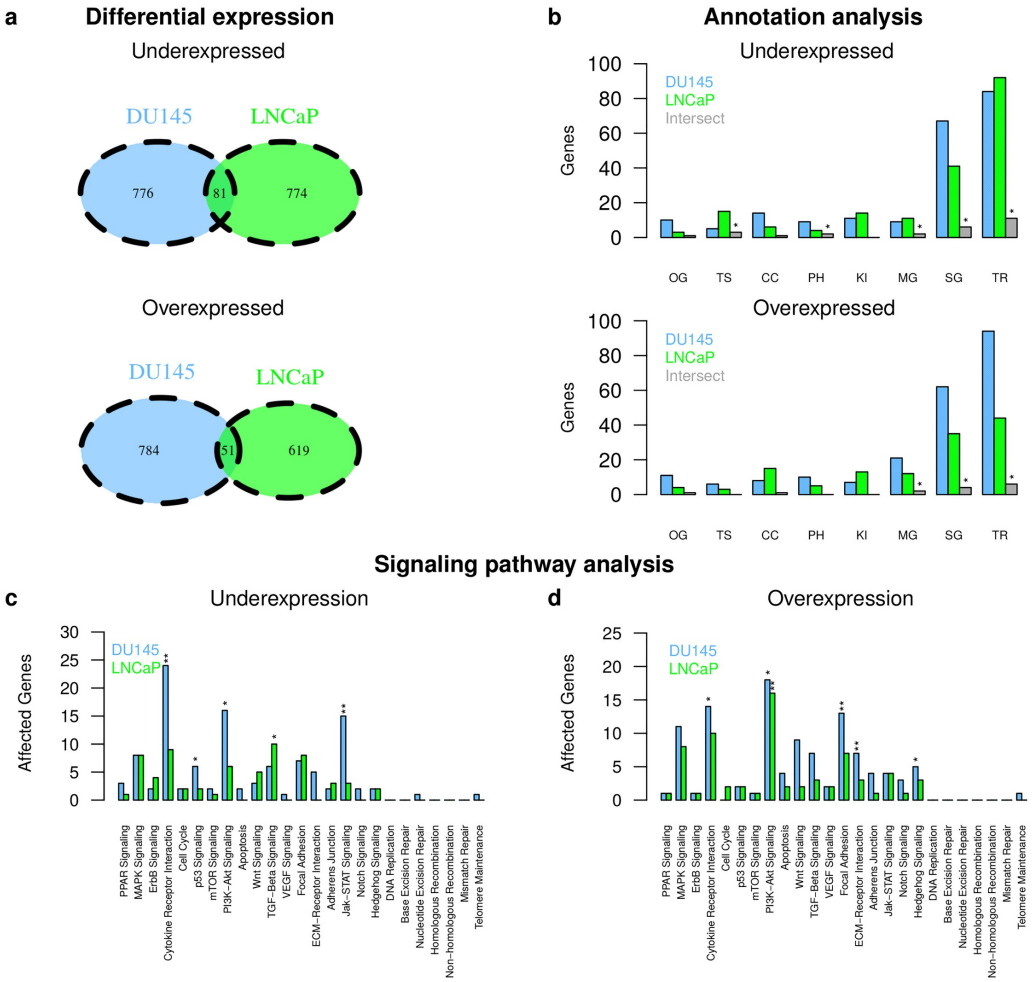

Fig 3. Gene expression differences between DU145 and LNCaP. Differentially expressed genes between radioresistant and radiosensitive cell lines were determined for DU145 and LNCaP. Identified under- (top panels) and overexpressed genes (bottom panels) in the radioresistant cell lines of DU145 and LNCaP were compared to each other at the single gene level (a) and at the level of cancer-relevant gene annotation categories (b; categories: oncogenes (OG), tumor suppressor genes (TS), cancer census genes (CC), phosphatases (PH), kinases (KI), metabolic pathway gene (MG), signaling pathway gene (SG), transcriptional regulator (TR)). Significant overlaps between categories are denoted by '*' (b; grey columns, $P < 0.001$, Fisher's exact test). Identified under- (c) and overexpressed genes (d) were further mapped to known cancer-relevant signaling pathways. Overrepresented pathways were highlighted by '*' ($P < 0.05$, Fisher's exact test) and '**' ($P < 0.01$).

https://doi.org/10.1371/journal.pcbi.1007460.g003

Further, tumor suppressor genes such as *EPB41L4A* and *TNFAIP3* had a reduced copy number and showed reduced expression, whereas oncogenes such as *ALDH1L2* and *WNT11* had an increased copy number and showed increased expression in radioresistant compared to radiosensitive DU145.

Generally, all genes with copy number alterations and consistent expression responses in the same direction represent putative direct driver candidates that could be involved in the manifestation of radioresistance.

Gene copy number alterations impact on expression of known radioresistance markers

To determine which of the radioresistance driver candidates with altered expression and underlying copy number alteration putatively contribute to the manifestation of radioresistance, we computed direct and indirect impacts of these candidates on the expression of known radioresistance marker genes (Fig 1). To realize this, we first used expression and copy number data of 14,780 genes of 541 prostate cancer patients from TCGA [23] to learn a prostate cancer-specific gene regulatory network (see Methods for details). This network was able to predict expression levels of individual genes across 768 independent cancer cell lines [41] with comparable power as in a previous study with other cancer types [28] (S3 Fig). Next, we used this network to compute for each putative radioresistance driver candidate (S4 Table) its potential impact on the expression of known altered radioresistance marker genes utilizing network propagation [28, 29] (see Methods for details, Fig 1 for an illustration, and S1 Fig for a detailed work flow illustration). Putative impacts of the DU145 and LNCaP driver candidates on the expression of differentially expressed cell line specific radioresistance marker genes are shown in Fig 4.

We found 162 driver candidates for DU145 (Fig 4a) and 27 for LNCaP (Fig 4b) that strongly impact on the expression of cell line specific radioresistance markers (S5 Table, $q < 0.01$). These driver candidates comprise overexpressed genes with increased copy number and underexpressed genes with decreased copy number. Potential driver candidates were distributed across the whole DU145 genome, whereas they were more focally distributed in LNCaP (Fig 4), which is expected because of the strong differences in DNA copy number alterations between both cell lines (Fig 2).

Considering the 162 driver candidates identified from DU145 (S5 Table, Fig 4a), several overexpressed genes with increased copy numbers encode membrane proteins (e.g. *RHBDL2*, *FZD7*, *SEMA5A*, *IL7R*, *STAB2*, *GPR124*, *NGFR*, *CAV1*) and transcriptional regulators (e.g. *ETV7*, *FOS*, *ATXN1*, *LEF1*) [36]. Further, *SOX17*, a transcription factor important for embryonic development and cell fate determination, and the tumor suppressors *SEPINB5* and *PTRO* were underexpressed with underlying reduced copy number [36]. Generally, these and other driver candidates were involved in the regulation of diverse cellular processes such as cytoskeletal remodeling, cell growth, proliferation, adhesion, or migration.

Considering the 27 driver candidates identified from LNCaP (S5 Table, Fig 4b), most genes were involved in cell adhesion (underexpressed with reduced copy number: *CDH19*, *DCC*, *FERMT1*, *FYN*, *VNN2* except *CLEC1A* and *KAL1*) [36]. Again genes involved in other cancer-relevant processes such as cell proliferation, migration, differentiation, apoptosis, or cytoskeletal remodeling were among the driver candidates (all underexpressed with reduced copy number: *ARHGAP18*, *DUSP10*, *PAK7*, *PDGFC*, *RNF217*) [36]. Further, the known tumor suppressors *DCC* and *GPRC5A* were underexpressed with underlying reduced copy number.

Only *KAL1* located on chromosome X was found as common high impact gene in DU145 and LNCaP (S5 Table, Fig 4), but *KAL1* was underexpressed with reduced copy number in DU145 and overexpressed with increased copy number in LNCaP comparing radioresistant to radiosensitive cell lines. *KAL1* encodes an extracellular matrix protein involved in cell migration [36]. Downregulation of *KAL1* has been associated with increased tumor size and vascular

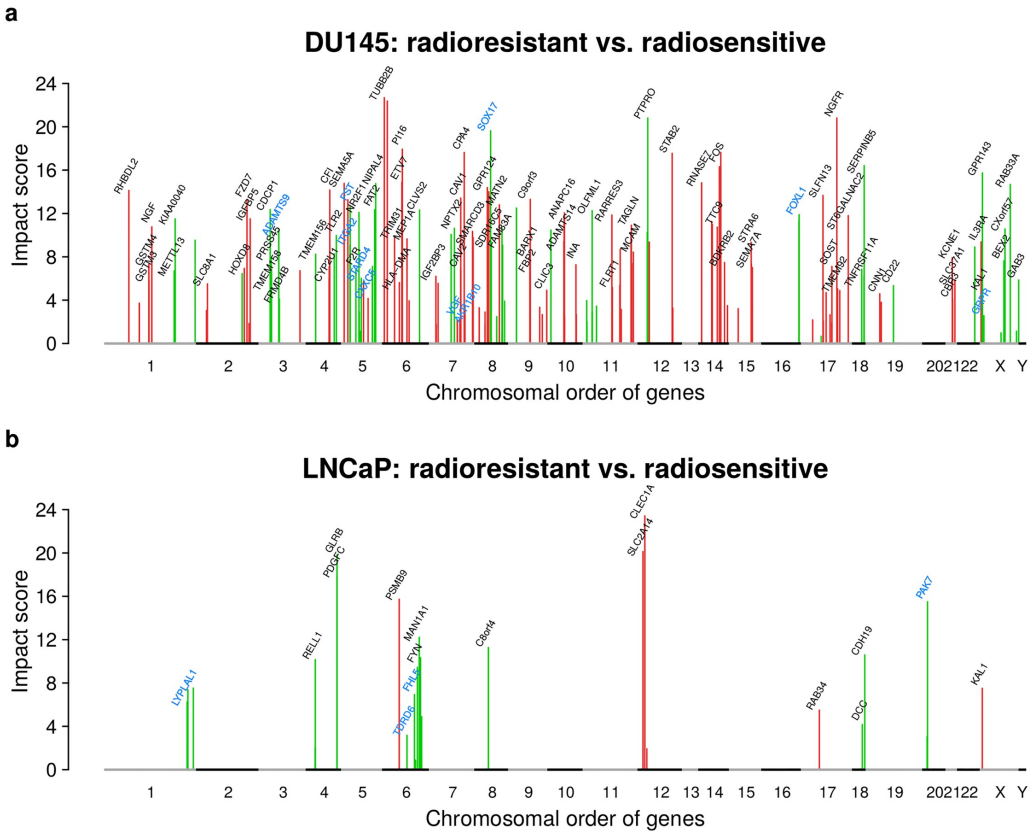

Fig 4. Impacts of potential radioresistance driver genes on known radioresistance markers. Impacts of differentially expressed genes with directly underlying copy number alterations in radioresistant DU145 (a) and radioresistant LNCaP (b) on known markers of radioresistance. The impact score represents the $\log_{10}$-ratio of the gene-specific impact on known radioresistance marker genes comparing the impact score reached for the prostate cancer specific network to the average impact score obtained under 10 random networks of same complexity (degree-preserving network permutations). Impact scores of genes with significantly greater impacts under the original network ($q < 0.01$) are shown by colored peaks (green: deleted and underexpressed; red: amplified and overexpressed for radioresistant vs. radiosensitive). The majority of gene names are shown. See S5 Table for names of all putative high impact genes. High impact genes that enabled a separation of TCGA prostate cancer patients into early and late relapse groups (Fig 5, S5 Fig) are highlighted in blue.

https://doi.org/10.1371/journal.pcbi.1007460.g004

invasion of hepatocellular carcinoma [42]. Similarly, silencing of KAL1 squamous cell carcinoma accelerated the G1 to M phase transition promoting cell proliferation and colony formation [43].

Generally, the small overlap between DU145 and LNCaP was expected due to strongly different copy number alteration profiles (Fig 2). Still, both sets of cell line specific driver candidates tend to act on the same cellular processes that could contribute to the manifestation of radioresistance.

Potential radioresistance drivers separate irradiated patients into early and late relapse groups

Next, we tested which of the identified cell line specific radioresistance driver gene candidates could potentially be relevant to predict the relapse behavior of prostate cancer patients treated by radiation therapy. To realize this, we analyzed the expression behavior of the driver candidates within a subgroup of 32 irradiated prostate cancer patients available from TCGA [23] (S6 Table). To enable relapse predictions for patients, only driver candidates with consistent expression behavior between radioresistant cell lines and irradiated patients were considered. Thus, a driver candidate that was underexpressed in radioresistant DU145 or LNCaP shows consistent behavior when irradiated patients with low expression of this gene tend to show faster relapses than patients with higher expression. In analogy, a driver candidate that was overexpressed in radioresistant DU145 or LNCaP shows consistent behavior if irradiated patients with high candidate gene expression tend to show faster relapses than patients with lower expression. We applied this consistency filtering to all driver candidates by comparing the cell line specific candidate gene expression behavior to the corresponding correlation between candidate expression and time until relapse in patients (see Methods for details). We found that 61 of 162 candidates from DU145 and 14 of 27 from LNCaP showed consistent expression behavior between cell lines and irradiated patients (S5 Table).

Next, we analyzed each of these candidate genes for its potential to distinguish between early and late relapse of irradiated prostate cancer patients by performing a Kaplan-Meier analysis. Under consideration that the early or late relapse group must contain at least eight patients, we found that 10 of 61 driver candidates from DU145 and 4 of 14 from LNCaP have the potential to distinguish between early and late relapse (S5 Table, Log-rank tests: $P < 0.05$ and corresponding conservative false discovery rates estimated between 14% and 22% [44] and more liberal estimates between 3% and 5% [45]). We also analyzed if the standard log-rank p-value computation for our small cohort of 32 patients with its determined different-sized early and late relapse subgroups had led to biased p-values [46]. We therefore computed the exact permutational log-rank p-values with the ExaLT method [46] for all DU145 and LNCaP driver candidates and compared them to the corresponding approximate log-rank p-values of our initial standard analysis. We found that the approximate log-rank p-values mostly overestimated the significance of the marker candidates, but this only marginally affected the ten driver candidates from DU145 (except for *FOXL1*: log-rank p-value increased from 0.014 to 0.076) and the four driver candidates from LNCaP and was clearly more pronounced for larger insignificant p-values (S4 Fig). The selected driver candidates are shown in Fig 5 and S5 Fig and listed in Table 1. Corresponding copy number alteration levels are shown in S6 Fig.

We found that high expression of *AKR1B10* or *VGF* was associated with patients that had a faster relapse than patients with lower expression of these genes (Fig 5). Further, low expression of *ADAMTS9*, *FOXL1*, *FST*, *GRPR*, *SOX17*, *STARD4*, *FHL5*, *LYPLAL1*, *PAK7*, *TDRD6*, *CXXC5*, or *ITGA2* was associated with patients that showed a faster relapse than patients with corresponding higher expression levels (Fig 5, S5 Fig). A detailed discussion of the identified driver candidates in the context of the existing literature is provided in S1 Text. Since patient-specific expression profiles were measured before radiation, theses driver candidates potentially represent markers whose expression behavior may allow to decide if a prostate cancer patient would profit from a radiation therapy or not.

Further, we investigated if the disease status after initial treatment of irradiated patients had biased the observed separations into early and late relapse groups, but we did not find any significant difference with respect to the distribution of patients with complete and non-complete

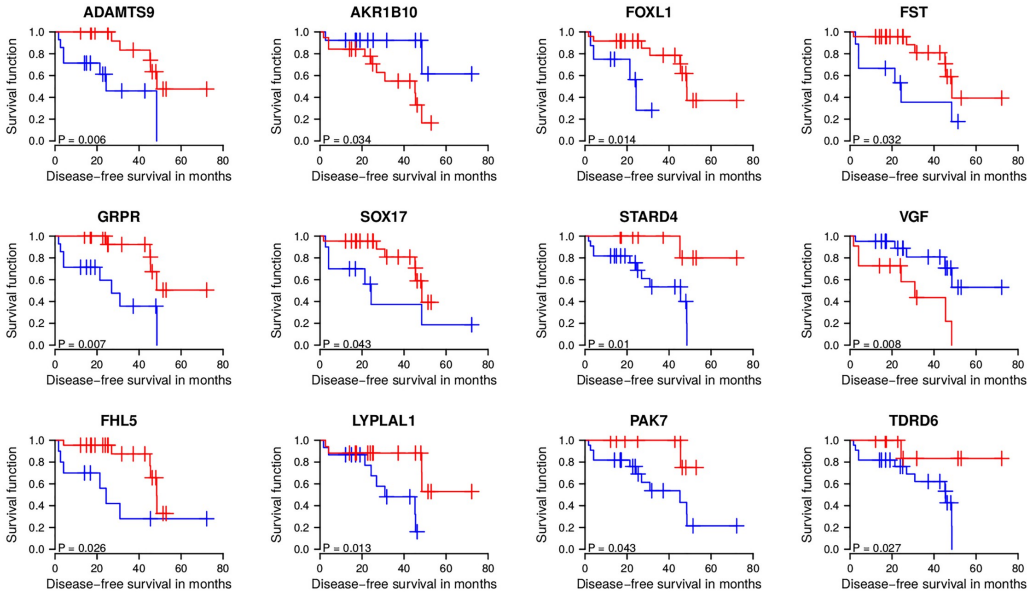

Fig 5. Marker gene-based separation of irradiated prostate cancer patients into early and late relapse groups. Potential radioresistance driver genes revealed from DU145 (top and middle row) and LNCaP (bottom row) were analyzed for their expression behavior in 32 irradiated prostate cancer patients from TCGA. Expression levels of each marker gene across the 32 patients were used to determine a marker gene-specific optimal cutoff for disease-free survival risk curves separating patients with low (blue curve) and high (red curve) marker gene expression with respect to the constraint that at least 8 patients must be assigned to each curve. Log-rank test p-values indicate that these selected marker genes enable a separation of irradiated prostate cancer patients into early and late relapse groups. Shown are standard approximate log-rank test p-values that only marginally deviated from exact log-rank p-values determined by exhaustive computations, except for FOXL1 that had a clearly less significant exact log-rank p-value of 0.076 (see Methods for details and S4 Fig). See S1 Text for a detailed discussion of the driver candidates in the context of the existing literature.

https://doi.org/10.1371/journal.pcbi.1007460.g005

remission after initial treatment between both groups (Fisher's exact tests: *P* ranged from 0.69 to 1).

Finally, we analyzed how our predicted driver candidates contribute to the modeling of the disease-free survival in the presence of additional covariates. Therefore, we used Cox regression [47, 48] to determine the contribution of prognostic factors (age, clinical T-stage, Gleason score, psa) with and without considering group assignments based on each driver candidate. We found that the prognostic factors alone were not well-suited to model the disease-free survival, whereas the driver candidates provided important information for the modeling of the disease-free survival in the presence of the other factors (S7 Fig).

VGF and *FHL5* also tend to predict relapse behavior of non-irradiated patients

Some of these marker candidates might also be associated with relapse of prostate cancer independent of radiation therapy. We therefore further analyzed the expression behavior of the marker candidates for 182 prostate cancer patients from TCGA that did not receive an adjuvant radiation therapy (S6 Table) [23]. We again tried to group the patients into early and late

Table 1. Summary of potential radioresistance drivers.

Gene	Faster Relapse	Annotations
ADAMTS9	low expression	protease function, renal tumors
AKR1B10	high expression	all-trans-retinaldehyde reductase, detoxification
FOXL1	low expression	transcription factor, proliferation, differentiation, metabolism
FST	low expression	follistatin, sexual hormone
GRPR	low expression	receptor for gastrin releasing peptide, associated with activation of phosphatidylinositol messenger system
SOX17	low expression	transcription factor, inhibits Wnt signaling, key regulator of embryonic development
STARD4	low expression	putative role in intracellular transport of sterols and other lipids
VGF	high expression	nerve growth factor inducible protein, regulation of cell-cell interactions
FHL5	low expression	putative role in spermatogenesis, stimulates CREM activity
LYPLAL1	low expression	lysophospholipase like 1, no phopholipase activity, able to hydrolyze short chain substrates
PAK7	low expression	protein kinase, involved in cytoskeleton regulation, cell migration, cell proliferation, and cell survival
TDRD6	low expression	involed in spermatogenesis, chromatin body formation, miRNA expression
CXXC5	high expression	required for DNA-damage induced phosphorylation, p53 activation and cell cycle arrest
ITGA2	low expression	trans-membrane receptor subunit, cell adhesion

Potential driver genes of radioresistance dividing irradiated prostate cancer patients from TCGA into early and late relapse groups. The column 'Faster Relapse' reports if patients with low or high gene-specific expression levels showed a faster relapse in the corresponding Kaplan-Meier curves shown in Fig 5 and S5 Fig. See S1 Text for a detailed discussion of the driver candidates in the context of the existing literature.

https://doi.org/10.1371/journal.pcbi.1007460.t001

relapse groups by performing a Kaplan-Meier analysis using the same driver gene-specific expression cutoffs determined in the prior analysis. We found that only *VGF* and *FHL5* enabled a similar separation of non-irradiated patients as observed for irradiated patients (S5 Fig, Log-rank test: $P < 0.1$). As for irradiated patients (Fig 5), high expression of *VGF* was associated with early relapse, whereas high expression of *FHL5* was associated with late relapse of non-irradiated patients (S5 Fig). Thus, both marker genes may also have at least some general prognostic potential for relapse, but only the increased expression of *VGF* in early relapse patients is of greater potential therapeutic relevance, because knockdowns are potentially better to realize than knockins.

Validation of *VGF* by *in vitro* radiobiological assays

We selected the neuroendocrine factor *VGF* for in-depth validation studies. This was motivated by our observation that *VGF* showed increased expression in prostate cancer patients with early relapse (Fig 5) and further triggered by recent studies that highlighted the importance of *VGF* in different types of cancer [49–52].

We found that *VGF* was significantly upregulated in DU145 and LNCaP prostate cancer radioresistant cell lines in our genome-wide gene expression analysis (Fig 6a; average expression difference of 2.85 in DU145 and 1.37 in LNCaP, t-tests: $P < 0.01$, S7 Table). We further analyzed the expression of *VGF* in independent radioresistant clones of DU145 and LNCaP in comparison to their radiosensitive parental cell lines and found that *VGF* was also significantly overexpressed in these clones (Fig 6b, t-test: $P < 0.05$ for DU145 and $P < 0.03$ for LNCaP, S7 Table). Interestingly, two of the four radioresistant DU145 clones had *VGF* expression levels that were comparable to those of the radioresistant LNCaP clones. These two radioresistant DU145 clones may not have an increased *VGF* copy number, but they still showed significantly increased *VGF* expression in comparison to the parental radiosensitive DU145 cell line (t-test: $P < 0.04$). This is comparable to the overexpression of *VGF* in radioresistant LNCaP without a

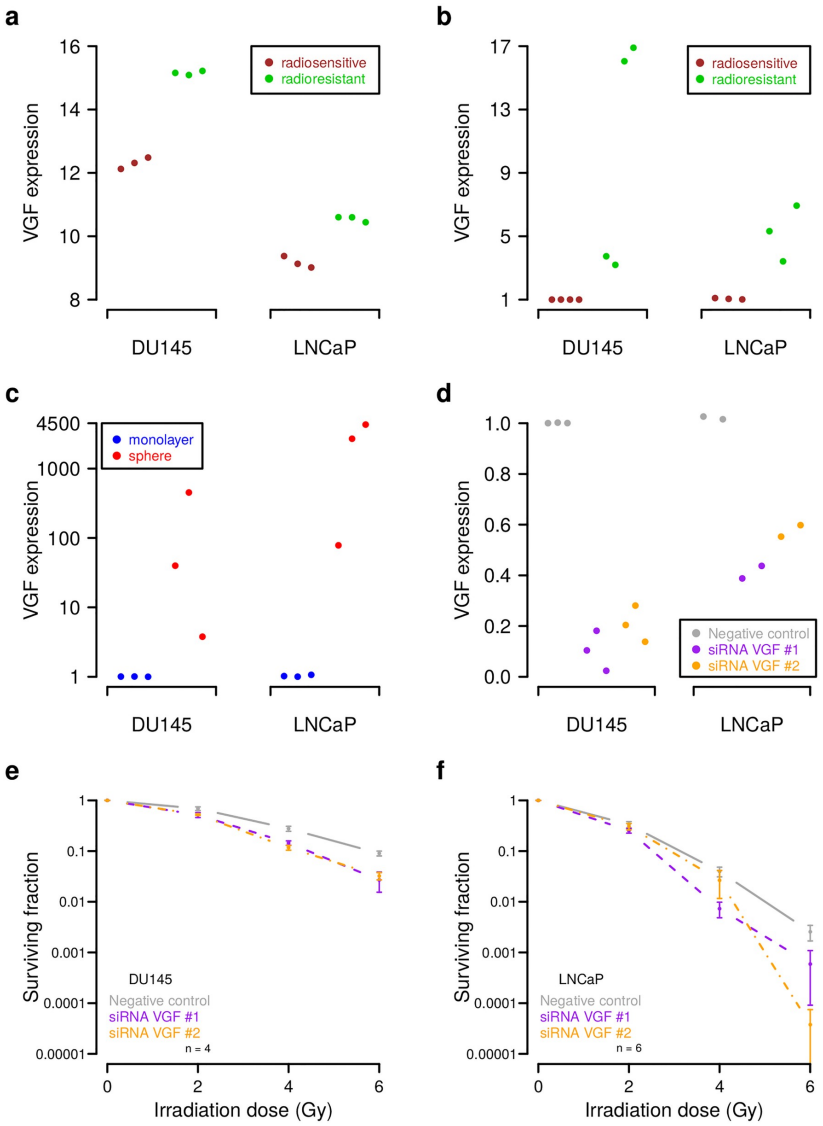

Fig 6. Experimental validation of *VGF* as regulator of cell radioresistance. (a) Increased *VGF* expression in radioresistant DU145 and LNCaP in comparison to their radiosensitive parental cell lines in our microarray data. Three biological replicates were considered for each condition. (b-d) RT-qPCR analysis of *VGF* expression under different conditions. (b) Increased *VGF* expression in four independent radioresistant DU145 and three independent radioresistant LNCaP clones relative to their radiosensitive parental cell lines. (c) Increased *VGF* expression in sphere relative to monolayer cultures of parental DU145 and LNCaP cells.

(d) Reduced *VGF* expression in parental DU145 and LNCaP cells induced by siRNA-mediated gene silencing relative to negative controls. (e-f) Increased radiosensitivity of parental DU145 and LNCaP cells induced by siRNA mediated reduction of *VGF* expression. Shown are average fractions of surviving cells in $\log_{10}$-scale for increasing radiation dose. Error bars represent the standard error of the mean and 'n' specifies the number of biological replicates. Corresponding linear-quadratic (LQ) model curves are shown in S11 Fig.

https://doi.org/10.1371/journal.pcbi.1007460.g006

directly underlying *VGF* copy number alteration and supports that increased *VGF* expression could contribute to increased radioresistance.

We further analyzed the expression behavior of *VGF* in parental DU145 and LNCaP cells grown under sphere-forming conditions (see S8 Fig for microscope images) that enrich cancer stem cell populations [53]. This was motivated by a recent study that showed that *VGF* is an important regulator of glioma stem cells [52]. We found that *VGF* expression was strongly increased under sphere-forming compared to monolayer conditions (Fig 6c, t-tests: $P < 0.06$ for DU145 and $P < 0.02$ for LNCaP, S7 Table). This observation was also supported by an additional analysis of the prostate cancer cell line PC3 that showed a moderately increased *VGF* expression under sphere-forming conditions (S8 and S9 Figs, t-test: $P < 0.02$, S7 Table).

Next, we considered the parental DU145 and LNCaP cells to determine changes in their radiosensitivity in response to reduced *VGF* expression by siRNA-mediated gene silencing. We first validated the knockdown efficiency by RT-qPCR and found clearly reduced *VGF* expression in *VGF* knockdowns compared to negative controls in both cell lines, where the efficiency was greater for DU145 than for LNCaP (Fig 6d, t-tests: $P < 0.002$ for DU145 and $P < 0.02$ for LNCaP, S7 Table). We also tried to validate the *VGF* knockdown by Western blots, but the two tested *VGF* antibodies (anti-VGF Santa Cruz sc-365397, B-8 mouse; St. John's Laboratory, STJ96661, rabbit, polyclonal) gave unspecific bands that were not consistent with the corresponding RT-qPCR data (S10 Fig). Since we had confirmed *VGF* knockdowns by RT-qPCR (Fig 6d), we next performed clonogenic assays to analyze the impact of *VGF* knockdowns on radiosensitivity. We found that an inhibition of *VGF* significantly increased the radiosensitivity of DU145 and LNCaP in comparison to controls transfected with scrambled siRNAs (Fig 6e and 6f; e.g. t-tests: $P < 0.02$ for siRNA VGF #2 vs. negative control at 4 Gy for DU145 and LNCaP, S7 Table). In addition, clearly lower surviving fractions of LNCaP cells further suggest that DU145 is more radioresistant than LNCaP, which is in accordance with our prior findings [33, 34].

Finally, we also considered the prostate cancer cell line PC3 and could confirm the efficiency of *VGF* knockdowns and we also observed a moderately increased radiosensitivity in response to *VGF* knockdowns (S9 Fig, e.g. t-test: siRNA VGF #2 vs. negative control: $P < 0.03$ at 4 Gy, S7 Table). We further estimated linear-quadratic models [54] of the clonogenic survival data of DU145, LNCaP, and PC3 to obtain functional representations of the individual survival curves (S11 Fig).

Discussion

Radioresistance of prostate cancer is driven by different cellular processes enabling cancer cells to survive radiation doses that can safely be delivered to the tumor [4, 5]. Molecular markers are urgently needed to better predict the clinical outcome of radiotherapies and to develop targeted adjuvant strategies to sensitize radioresistant cells. Radioresistant prostate cancer cell lines represent an important model system for the identification of novel candidate genes and the analysis of molecular mechanisms involved in radioresistance, but they typically show large chromosomal deletions and amplifications that affect many genes. This in combination with the small number of cell lines that are usually profiled and their cell line specific gene

copy number and expression profiles does not allow a straightforward identification of radioresistance drivers by standard statistical approaches for gene copy number and expression analysis. In this situation, it is almost impossible to derive promising candidates from hundreds or thousands of differentially expressed genes with directly underlying gene copy number alterations without prior knowledge about genes involved in altered cellular processes that contribute to radioresistance.

Therefore, we developed a network-based method to jointly analyze the gene copy number and expression profiles of an individual cell line to distinguish potential drivers from passengers. The essential basis of this approach was the prostate cancer-specific gene regulatory network that we learned from gene expression and copy number data of 541 prostate cancer patients from TCGA. This network inference was very time and resource consuming requiring 670 hours on a high-performance compute server (Taurus ZIH TUD). Validations on data of 768 cancer cell lines from [28, 41] confirmed that this network can predict the expression behavior of individual genes in cancer cell lines enabling an analysis of the prostate cancer cell lines DU145 and LNCaP. This analysis is limited by the fact that the cancer samples from TCGA and our cell lines were analyzed on different experimental platforms. Both data sets also showed differences in the number of expressed genes, where more genes were expressed in our cell line models than in the cancer samples. Thus, it is clear that not all observations form our *in vitro* prostate cancer cell lines are transferable to the *in vivo* situation in prostate tumors. Nevertheless, we applied network propagation to differentially expressed genes with directly underlying copy number alterations from DU145 and LNCaP to determine their impacts on known markers of radioresistance. Comparisons to random networks of same complexity (degree-preserving network permutations) in combination with further filtering revealed ten candidates from DU145 (*ADAMTS9*, *AKR1B10*, *CXXC5*, *FST*, *FOXL1*, *GRPR*, *ITGA2*, *SOX17*, *STARD4*, *VGF*) and four from LNCaP (*FHL5*, *LYPLAL1*, *PAK7*, *TDRD6*) that were able to distinguish irradiated prostate cancer patients from TCGA into early and late relapse groups. A detailed discussion of these candidate genes is given in S1 Text. These candidate genes may allow to develop biomarkers for the analysis of biopsy samples to predict relapse risk and to adapt treatment for individual prostate cancer patients. Targeted perturbations of these genes may allow to increase the radiosensitivity of prostate cancer cells. Additional preclinical and clinical studies are required to validate these candidates.

We experimentally validated the novel radioresistance marker gene candidate *VGF*, a neuroendocrine factor, that was highly overexpressed in DU145 and LNCaP radioresistant prostate cancer cell lines and whose high expression was associated with shorter disease-free survival of irradiated prostate cancer patients. *VGF* was originally identified in a pheochromocytoma cell line in response to the addition of the nerve growth factor (NGF) [55]. *VGF* is an important regulator of metabolism and endoplasmic reticulum (ER) stress in neurons and endocrine cells [56–58], where it activates pro-survival signaling pathways such as PI3K/AKT/mTOR and MAPK/ERK1/2 [59, 60], but its role in regulation of cancer cells remained unclear for a long time. Experimental evidences from *in vitro* models, mouse xenografts and analysis of patient outcomes showed that *VGF* expression is associated with resistance to EGFR inhibitors and further induces epithelial-mesenchymal transition (EMT) and tumor cell dissemination [50, 51]. In addition, *VGF* has been shown to be preferentially expressed in glioblastoma stem cells promoting glioblastoma stem cell survival and stemness and to further support survival of differentiated glioblastoma cells to promote tumor growth [52]. Our previous studies showed that the emergence of radioresistance also triggers EMT, increases migratory properties, and further results in enrichment of cancer stem cell populations in prostate cancer cells [33]. In accordance with this, our *in vitro* validation experiments confirmed an upregulation of *VGF* expression in additionally analyzed independent radioresistant DU145 and LNCaP

clones, showed that *VGF* is highly expressed under sphere forming conditions, and further demonstrated that *VGF* knockdowns lead to increased radiosensitivity. These results suggest that VGF is involved in radioresistance of prostate cancer. This is also supported by our findings for the prostate cancer cell line PC3.

For Western blotting analysis of *VGF* in response to siRNA-mediated gene silencing, we tried two available antibodies (anti-VGF Santa Cruz sc-365397, B-8 mouse; St. John's Laboratory, STJ96661, rabbit, polyclonal) and we additionally performed RT-qPCR analysis of *VGF* expression as control in parallel. Although we observed a pronounced knockdown of *VGF* by RT-qPCR, we did not observe the specific *VGF* band by Western blotting, which can be explained by the observation of substantial background signals. Therefore, we focused on PCR-based analysis of VGF expression in our validation studies.

We observed that VGF knockdowns were more efficient in DU145 than in LNCaP, but fewer LNCaP cells survived irradiation. The relation between knockdown efficiency and cell survival after irradiation is complex. Different factors can contribute to cell line specific radioresistance. We already know from our prior studies [33, 34] that DU145 is more radioresistant than LNCaP. This is in accordance with our observation that DU145 had substantially more DNA copy number alterations than LNCaP and could explain better survival of DU145 cells in response to irradiation by a greater tolerance of DNA double strand breaks. Further, the knockdown efficiency also depends on the protein turnover rate [61] and highly expressed genes can be more susceptible to siRNA-mediated gene silencing [62]. Thus, the found stronger expression of *VGF* in DU145 than in LNCaP may also have influenced the *VGF* knockdown efficiency observed for both cell lines. Nevertheless, our clonogenic assays clearly indicate that *VGF* could be involved in the regulation of radioresistance.

Further, our *in vitro* characterization of DU145 and LNCaP is limited to the identification of molecular alterations that are associated with intrinsic cellular radioresistance. Additional preclinical and clinical studies are necessary to further analyze the revealed marker genes in *in vivo* studies. Especially the tumor microenvironment and immune signatures of tumors can be altered by radiation therapies influencing tumor progression and therapy response [6–12]. Thus, also microenvironmental and immunomodulatory factors, which we could not cover by our analysis, can strongly influence the response of individual tumors to radiation therapy. Such and other limitations of *in vitro* cancer models have been reported over the last years [63] and special care has to be taken on work with cancer cell lines [64]. For example, in a transgenic breast cancer model tumors with similar growth characteristics but different immune signatures differed in their response to radiation therapy [65]. Therefore, a combination of radiation and immune therapy is important to improve patient outcomes [11, 66]. Another example is the treatment of the prostate cancer cell line PC3 with the HIV protease inhibitor nelfinavir that resulted in a small but significant increase of radiosensitivity *in vitro* which was not observed in corresponding PC3 xenografts [67]. Still, our analysis of revealed markers that distinguished between early and late relapse of irradiated prostate cancer patients provides a first important hint that these markers have the potential to enable predictions for the *in vivo* situation.

In summary, our detailed literature analysis and results of radiobiological assays for the maker gene *VGF* suggest that our network-based approach can predict potentially clinically relevant driver candidates involved in radioresistance of prostate cancer.

Materials and methods

A detailed flow chart of our developed data analysis pipeline is shown in S1 Fig. See Fig 1 for a high-level overview.

Identification of gene copy number alterations

Array-based comparative genomic hybridization (aCGH) was used to compare the genomes of radioresistant to radiosensitive cell lines for DU145 and LNCaP and to compare these genomes to normal reference DNA (Agilent Euro Male). Experiments were done on Agilent's SurePrint G3 Human CGH Microarray Kit 2x400K (Design ID: 028081, Agilent) and performed and standardized as described in [68]. Normalized measurements were used to compute aCGH profiles. An aCGH profile represents for each of the 294,371 genomic probes a $\log_2$-ratio that compares the probe-specific DNA copy number in a radioresistant cell line relative to its radiosensitive counterpart (or to compare DNA copy numbers of a radioresistant or radiosensitive cell line to normal DNA). aCGH profiles were sorted according to chromosomal locations of probes and further segmented into chromosomal regions of constant copy number using DNAcopy [69]. Corresponding DNA segmentation profiles are provided in S1 Table. Copy number values of 24,625 genes (focusing on genes for which we also measured expression) were determined by mapping chromosomal locations of genes to the aCGH segments as described in [28]. The resulting $\log_2$-ratio gene copy number values were used to determine genes with increased or reduced copy number in radioresistant DU145 or LNCaP relative to their non-resistant counterpart using an absolute $\log_2$-ratio cutoff of 0.1 (Fig 2, S2 Table). The choice of this cutoff was motivated by moderately increased or decreased gene copy number alteration values comparing radioresistant to radiosensitive LNCaP. This choice did not influence the network inference and the computation of the network propagation matrix. This cutoff only defines a filter for the selection of candidate genes that were considered for more in-depth analyses. A heatmap representation that summarizes all gene copy number comparisons is shown in S2 Fig. aCGH data have been deposited in the Gene Expression Omnibus (GEO) database, accession no GSE134500.

Identification of differentially expressed genes

Gene expression levels of radioresistant and radiosensitive cell lines of DU145 and LNCaP were measured in three biological replicates. Experiments were done on Agilent's SurePrint G3 Human Gene Expression 8x60K v2 microarrays (Design ID: 039494, Agilent) and performed as described in [34]. Hybridization signals of 24,625 genes of all cell line specific experiments were quantile normalized [70]. Expression differences between the three radioresistant and the three radiosensitive LNCaP cell lines were not strong enough to enable a prediction of differentially expressed genes by standard t-tests with significant p-values after correction for multiple testing. Still, the t-test statistic, the p-value or the average log-ratio provide important information to rank genes according to their expression differences. We therefore used a specifically designed three-state Hidden Markov Model (HMM) to identify differentially expressed genes [35]. We trained two independent HMMs, one for DU145 and one for LNCaP, on the average gene expression $\log_2$-ratio profile comparing radioresistant to radiosensitive cell lines to account for cell line specific expression characteristics. This training was done with standard settings and initial state-specific means of -1.25 (underexpressed), 0 (unchanged), and 1.25 (overexpressed). We used state-posterior decoding to assign each gene to its most likely underlying state (underexpressed, unchanged, or overexpressed) in radioresistant relative to radiosensitive cell lines (S3 Table). Gene expression data have been deposited in the Gene Expression Omnibus (GEO) database, accession no GSE134500.

Inference of prostate cancer-specific gene regulatory network

We learned a prostate cancer-specific gene regulatory network to predict potential impacts of gene copy number alterations in DU145 and LNCaP on known radioresistance marker genes.

We downloaded aCGH profiles and gene expression data of 541 prostate cancer patients from TCGA and processed them as described in [28]. In addition, we removed all genes with very low constant or nearly non-variable expression across all patients and only kept genes with on average at least 1 transcription unit (normalized RSEM [71] counts from TCGA) per patient. Gene copy number and expression measurements of the remaining 14,780 genes were used to learn a gene regulatory network as outlined in [28] using the R package regNet [29]. Briefly, the expression of a specific gene was modeled as a linear combination of its copy number and the expression of all other genes. Lasso regression [72] in combination with cross validation and a significance test for lasso [73] were used to determine for each gene those predictors (e.g. gene-specific copy number or expression levels of other genes) that best explained the expression behavior of the considered gene across all prostate cancer patients. As done in [28], we focused on the most relevant links (p-values approximately zero) and removed spurious local regulators (local gene cutoff of 50) resulting in a prostate cancer-specific network with 60,447 activator and 2,105 inhibitor links between genes (S8 Table). We further confirmed that this network was capable to predict the expression of genes in cancer cell lines outperforming random networks of same complexity derived by degree-preserving network permutations (S3 Fig). The predictive power of our network was comparable to the predictive power of other networks that we had learned with the same lasso approach [28].

Impact quantification of gene copy number alterations on radioresistance markers

We applied network propagation [28] in combination with the prostate cancer-specific gene regulatory network to determine impacts of differentially expressed genes with underlying gene copy number alterations on the expression of known radioresistance marker genes. We used the R package regNet [29] to compute a specific impact matrix based on the cell line specific log-ratio gene copy number and expression profiles comparing the radioresistant cell line to its radiosensitive counterpart for DU145 and LNCaP separately. Each cell line specific impact matrix quantifies for each gene pair (a, b) how strong gene a acts on the expression of gene b by computing the impact that flows from gene a to gene b via all possible network paths in the prostate cancer-specific gene regulatory network connecting both genes under consideration of the predictive power of individual genes. More weight was given to genes with greater positive correlations than to genes with smaller positive correlations utilizing gene-specific correlation estimates obtained from cancer cell lines (S3 Fig). Next, we only considered potential radioresistance driver candidates focusing on genes with increased expression and underlying increased copy number and on genes with decreased expression and underlying decreased copy number in radioresistant versus radiosensitive cell lines (S4 Table). In total, 292 of 447 genes for DU145 and 40 of 66 genes for LNCaP that fulfilled these criteria were also expressed in prostate cancer samples of TCGA patients. We considered each candidate gene and determined its average impact on known differentially expressed cell line specific radioresistance markers (DU145: *CCL2*, *CLDN4*, *MRC2*, *SNAI2* overexpression; LNCaP: *CXCR4* underexpression in radioresistant vs. radiosensitive cell lines; S3 Table) from the cell line specific impact matrix (S5 Table). Finally, we determined which of the potential radioresistance driver candidates had significant impacts on these differentially expressed cell line specific radioresistance markers. Therefore, we computed corresponding average impacts under 10 random networks of same complexity as the original prostate cancer-specific network. These random networks were derived based on degree-preserving network permutations by exchanging active predictors between gene-specific linear models while keeping the number of incoming and outgoing links constant for each gene. We compared the driver gene specific

random impacts to the corresponding original impact by computing differences between the original impact and each corresponding random impact and used t-tests to determine which gene-specific differences in impact scores were significantly greater than zero. Genes with impacts significantly greater than under random networks were selected based on FDR-adjusted p-values (q-values) with $q < 0.01$ [44] leading to 162 potential driver candidates for DU145 and 27 for LNCaP (S5 Table).

Transfer of cell line specific radioresistance driver genes to prostate cancer patients

We analyzed the expression of potential radioresistance drivers of DU145 and LNCaP in prostate cancer patients from TCGA to identify marker candidates that distinguish between early and late relapse after adjuvant radiation therapy. Sufficient meta-information about initial treatment, treatment response and disease free survival were available for 214 of 541 prostate cancer patients of which 32 patients received radiation and 182 did not (S6 Table). All patients were either disease free or showed a relapse after initial treatment. In more detail, the majority of patients showed a complete remission (156 of 214), whereas other patients showed a stable disease (22 of 214), partial remission (23 of 214), or a progressive disease (13 of 214) after initial treatment. To determine marker genes that distinguish between early and late relapse, only genes with consistent expression behavior between radioresistant cell lines and irradiated patients were considered. Therefore, we translated the observed expression state of a potential marker candidate from the cell lines into a meaningful interpretation for irradiated tumor patients. We assumed that if a marker candidate was overexpressed (underexpressed) in the radioresistant compared to the radiosensitive cell line, then this overexpression (underexpression) may contribute to radioresistance. Consequently, irradiated patients with high (low) expression levels of this gene may show a faster relapse than irradiated patients with lower (higher) expression levels. Thus, a negative (positive) correlation between marker gene-specific expression in patients and disease free survival is expected.

To realize this, we first considered gene expression profiles of tumors before treatment to compute correlations between the expression of each potential radioresistance driver gene and the months until relapse (disease free survival) considering all 12 of 32 irradiated patients that had a relapse (S6 Table). Next, we compared the obtained gene specific correlations to the corresponding expression states observed for the cell lines and only kept those potential marker genes for further analysis that were in accordance with the transfer of the expression behavior from cell lines to tumors outlined above (overexpressed in radioresistant cell line vs. negative correlation between tumor expression and time until relapse, undexpressed in radioresistant cell line vs. positive correlation between tumor expression and time until relapse). This was fulfilled by 61 of 162 potential radioresistance driver genes from DU145 and for 14 of 27 from LNCaP (S5 Table). Finally, we analyzed each of these marker candidates for its potential to distinguish between early and late relapse of prostate cancer patients that received adjuvant radiation therapy. Therefore, we did a Kaplan-Meier analysis for each marker candidate where we tried to split the 32 irradiated TCGA prostate cancer patients into an early and late relapse group under consideration of the marker specific expression (R package 'survival' [74]). We determined an optimal gene expression cutoff for each marker for the separation into early and late relapse (disease free survival) by computing corresponding log-rank p-values with respect to the constraint that each group must contain at least eight patients. We selected all marker candidates with p-values less than 0.05 resulting in 10 markers from DU145 and 4 markers from LNCaP capable to distinguish between early and late relapse of irradiated prostate cancer patients for further analysis (S5 Table). The correlation between predicted and

experimentally measured expression levels of these 14 candidate markers was significantly greater than zero (t-test: $P < 0.019$) and at the level of individual genes also significantly better than for random networks of same complexity derived by degree-preserving network permutations (paired t-test: $P < 0.02$). Corresponding estimated conservative false discovery rates were between 14% and 22% [44] and more liberal estimates between 3% and 5% [45] (S5 Table). Random selections of genes would have resulted on average on 0.90 genes for LNCaP and 2.25 genes for DU145 with log-rank p-values less than 0.05 (95% confidence interval [0.88, 0.91] for LNCaP and [2.22, 2.27] for DU145), which is significantly less than the number of driver candidates predicted by our network-based approach.

In addition, we used the ExaLT algorithm [46] to compute exact permutational log-rank p-values for each optimal candidate gene-specific split between early and late relapse patients. Our initially computed approximate log-rank p-values (R package 'survival') varied only marginally from the exact permutational p-values, except for *FOXL1* (increase in log-rank p-value from 0.014 to 0.076), supporting that our selection of driver gene candidates based on the small cohort of irradiated patients was robust (S4 Fig). We also used Cox regression [47, 48] (R package 'survival') to analyze if our driver candidates were still informative for disease-free survival in the presence of currently used prognostic factors (age, clinical T-Stage, Gleason score, psa). The grouping information about early or late relapse derived form each individual driver candidate was important to model disease-free survival and reached more significant p-values than the other covariates for 13 of 14 candidate genes (S7 Fig, *AKR1B10*: clinical T-stage was slightly more significant than the grouping information derived from *AKR1B10* expression).

Further, we used the determined optimal marker gene-specific expression cutoffs to analyze the 182 non-irradiated TCGA prostate cancer patients to determine those markers that were exclusively associated with relapse of irradiated patients but not with relapse of non-irradiated patients.

Cell lines and culture conditions

Prostate cancer cell lines DU145, LNCaP and PC3 were purchased from the American Type Culture Collection (ATCC, Manassas, VA) and cultured according to the manufacturers recommendations in a humidified 37˚C incubator supplemented with 5% CO_2. DU145 and PC3 cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) (Sigma-Aldrich) and LNCaP cells in RPMI-1640 medium (Sigma-Aldrich) containing 10% fetal bovine serum (FBS, PAA Laboratories) and 1 mM L-glutamine (Sigma-Aldrich). The analyzed radioresistant cell lines of DU145 and LNCaP were established in [33] and further analyzed in [34]. In more detail, radioresistant cell sublines of DU145 and LNCaP had been generated by multiple fractions of 4 Gy X-ray irradiation until a total dose of more than 56 Gy was reached (Fig. 4a in [33]). Colony assays had been used to demonstrate the enhanced radioresistance of surviving cells (Fig. 4b in [33]). Corresponding age-matched non-irradiated radiosensitive parental cells were used as controls for radioresistant cell lines. All cell lines were genotyped using microsatellite polymorphism analysis and tested for mycoplasma directly before the experiments.

Sphere formation assay

To evaluate the self-renewal potential, cells were grown as non-adherent multicellular cell aggregates (spheres). Cells were plated at a density of 1,000 cells/2 mL/well in 6-well ultra-low attachment plates (Corning) in MEBM medium (Lonza) supplemented with 4 μg/mL insulin (Sigma-Aldrich), B27 (Invitrogen), 20 ng/mL EGF (Peprotech), and 20 ng/mL FGF (Peprotech). Media containing supplements were refreshed once a week and spheres with a

size > 100 μm were assayed after 14 days using Axiovert 25 microscope (Zeiss) or were automatically scanned using the Celigo S Imaging Cell Cytometer (Brooks).

Knockdown of *VGF* by siRNA transfection

For knockdown of *VGF* expression, cells were transfected with RNAiMAX (Life Technologies GmbH) according to the manufacturer's protocol. The siRNA target sequences were obtained from the Life Technologies website and corresponding RNA duplexes were synthesized by Eurofins. The sequences were VGF siRNA 1: sense GGAAGAAGCAGCUGAAGCUdCdT; antisense AGCUUCAGCUGCUUCUUCCdTdC and VGF siRNA 2: sense GGAGGAGCUG GAGAAUUACdAdT; antisense GUAAUUCUCCAGCUCCUCCdTdG for targeted knockdowns of *VGF*. Scrambled siRNA 1: sense UGCGCUAGGCCUCGGUUGCdTdT; antisense GCAACCGAGGCCUAGCGCAdTdT, scrambled siRNA 2: sense: AGGUAGUGUAAUCGC CUUGdTdT; antisense CAAGGCGAUUACACUACCUdTdT, and scrambled siRNA 3: sense GCAGCUAUAUGAAUGUUGUdTdT; antisense ACAACAUUCAUAUAGCUGCdTdT were used as negative control. In addition, knockdown efficiencies of VGF siRNA 1 and 2 were analyzed by RT-qPCR in comparison to scrambled siRNAs considering three biological replicates for DU145 and PC3 and two for LNCaP. Seven technical replicates were done for each biological replicate.

Clonogenic cell survival assay

Cells were plated at a density of 500 cells/well in 6-well plates in complete medium and irradiated with doses of 2, 4 and 6 Gy of 200 kV X-rays (Yxlon Y.TU 320; dose rate 1.3 Gy/min at 20 mA) filtered with 0.5 mm Cu. The absorbed dose was measured using a Duplex dosimeter (PTW). After 10 days, the colonies were fixed with 10% formaldehyde (VWR) and stained with 0.05% crystal violet (Sigma-Aldrich). Colonies containing > 50 cells were counted using a stereo microscope (Zeiss). The plating efficiency (PE) was calculated as ratio between the number of colonies and the number of cells plated. The surviving fraction (SF) was calculated as ration between the PE of irradiated cells divided by PE of corresponding non-irradiated control cells. We also learned linear-quadratic (LQ) models to obtain a functional representation of the surviving fraction for each cell line using the R package 'CFAssay' [54] with standard settings (S11 Fig). We did not consider higher irradiation doses of 8 or 10 Gy in our experiments, because only few cells survived at 6 Gy especially for LNCaP and PC3.

Supporting information

S1 Text. Literature analysis and discussion of identified driver candidates.
(PDF)

S1 Fig. Technical flow chart.
(PDF)

S2 Fig. Heatmap representation of copy number alterations.
(PDF)

S3 Fig. Predictive power of network for cancer cell lines.
(PDF)

S4 Fig. Comparison of approximate and exact log-rank p-values.
(PDF)

S5 Fig. Marker gene-based separation of prostate cancer patients into early and late relapse groups.
(PDF)

S6 Fig. Copy number alteration levels of driver candidate genes.
(PDF)

S7 Fig. Cox regression results for modeling of disease-free survival.
(PDF)

S8 Fig. DU145, LNCaP and PC3: Monolayer vs. Spheres.
(PDF)

S9 Fig. Validation of VGF in prostate cancer cell line PC3.
(PDF)

S10 Fig. Western blots and RT-qPCR analysis.
(PDF)

S11 Fig. LQ model fits of clonogenic survival.
(PDF)

S1 Table. DNA copy number segmentation profiles of DU145 and LNCaP.
(SEG)

S2 Table. Gene copy number data of DU145 and LNCaP.
(XLS)

S3 Table. Gene expression data of DU145 and LNCaP.
(XLS)

S4 Table. Differentially expressed genes with directly underlying copy number alterations for DU145 and LNCaP.
(XLS)

S5 Table. Impacts of differentially expressed genes with directly underlying copy number alterations on known radioresistant marker genes.
(XLS)

S6 Table. Clinical information of irradiated and non-irradiated prostate cancer patients from TCGA.
(XLS)

S7 Table. Data of *VGF* validation experiments.
(XLS)

S8 Table. Connectivity table of prostate cancer-specific gene regulatory network.
(TSV)

Acknowledgments

We thank Vasyl Lukiyanchuk (OncoRay) for assistance with RT-qPCR and discussion of gene expression and aCGH data.

Author Contributions

Conceptualization: Michael Seifert, Anna Dubrovska.

Data curation: Michael Seifert, Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner, Barbara Klink, Anna Dubrovska.

Formal analysis: Michael Seifert.

Funding acquisition: Anna Dubrovska.

Investigation: Michael Seifert, Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner, Barbara Klink, Anna Dubrovska.

Methodology: Michael Seifert.

Project administration: Michael Seifert, Anna Dubrovska.

Resources: Michael Seifert, Claudia Peitzsch, Barbara Klink, Anna Dubrovska.

Software: Michael Seifert.

Supervision: Michael Seifert, Anna Dubrovska.

Validation: Michael Seifert, Claudia Peitzsch, Ielizaveta Gorodetska, Caroline Börner, Barbara Klink, Anna Dubrovska.

Visualization: Michael Seifert.

Writing – original draft: Michael Seifert, Anna Dubrovska.

Writing – review & editing: Michael Seifert.

References

1. Johansson S, Astrom L, F S, Isacsson U, Montelius A, Turesson I. Hypofractionated proton boost combined with external beam radiotherapy for treatment of localized prostate cancer. Prostate Cancer. 2012; 2012:654861. https://doi.org/10.1155/2012/654861 PMID: 22848840
2. Pahlajani N, Ruth KJ, Buyyounouski MK, Chen DY, Horwitz EM, Hanks GE, et al. Radiotherapy doses of 80 Gy and higher are associated with lower mortality in men with Gleason score 8 to 10 prostate cancer. Int J Radiat Oncol Biol Phys. 2012; 82(5):1949–56. https://doi.org/10.1016/j.ijrobp.2011.04.005 PMID: 21763081
3. Zietman AL, Bae K, Slater JD, Shipley WU, Efstathiou JA, Coen JJ, et al. Randomized trial comparing conventional-dose with high-dose conformal radiation therapy in early-stage adenocarcinoma of the prostate: long-term results from proton radiation oncology group/american college of radiology 95-09. J Clin Oncol. 2010; 28(7):1106–11. https://doi.org/10.1200/JCO.2009.25.8475 PMID: 20124169
4. Bonkhoff H. Factors implicated in radiation therapy failure and radiosensitization of prostate cancer. Prostate Cancer. 2012; 2012:593241. https://doi.org/10.1155/2012/593241 PMID: 22229096
5. Chang L, Graham PH, Hao J, Bucci J, Cozzi PJ, Kearsley JH, et al. Emerging roles of radioresistance in prostate cancer metastasis and radiation therapy. Cancer Metastasis Rev. 2014; 33(2-3):469–96. https://doi.org/10.1007/s10555-014-9493-5 PMID: 24445654
6. Corn PG. The tumor microenvironment in prostate cancer: elucidating molecular pathways for therapy development. Cancer Manag Res. 2012; 4:183–93. https://doi.org/10.2147/CMAR.S32839 PMID: 22904640
7. McAllister MJ, Underwood MA, Leung HY, Edwards J. A review on the interactions between the tumor microenvironment and androgen receptor signaling in prostate cancer. Transl Res. 2019; 206:91–106. https://doi.org/10.1016/j.trsl.2018.11.004 PMID: 30528321
8. Barker HE, Paget JT, Khan AA, Harrington KJ. The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. Nat Rev Cancer. 2015; 15(7):409–25. https://doi.org/10.1038/nrc3958 PMID: 26105538
9. Leroi N, Lallemand F, Coucke P, Noel A, Martinive P. Impacts of Ionizing Radiation on the Different Compartments of the Tumor Microenvironment. Front Pharmacol. 2016; 7:78. https://doi.org/10.3389/fphar.2016.00078 PMID: 27064581
10. Di Lorenzo G, Buonerba C, Kantoff PW. Immunotherapy for the treatment of prostate cancer. Nat Rev Clin Oncol. 2011; 8(9):551–61. https://doi.org/10.1038/nrclinonc.2011.72 PMID: 21606971

11. Finkelstein SE, Salenius S, Mantz CA, Shore ND, Fernandez EB, Shulman J. Combining immunotherapy and radiation for prostate cancer. Clin Genitourin Cancer. 2015; 13(1):1–9. https://doi.org/10.1016/j.clgc.2014.09.001 PMID: 25450032
12. Cordes LM, Gulley JL, Madan RA. The evolving role of immunotherapy in prostate cancer. Curr Opin Oncol. 2016; 28(3):232–40. https://doi.org/10.1097/CCO.0000000000000281 PMID: 26977847
13. Tang L, Wei F, Wu Y, He Y, Shi L, Xiong F, et al. Role of metabolism in cancer cell radioresistance and radiosensitization methods. J Exp Clin Cancer Res. 2018; 37(1):87. https://doi.org/10.1186/s13046-018-0758-7 PMID: 29688867
14. Chaiswing L, Weiss HL, Jayswal RD, Clair DKS, Kyprianou N. Profiles of Radioresistance Mechanisms in Prostate Cancer. Crit Rev Oncog. 2018; 23(1-2):39–67. https://doi.org/10.1615/CritRevOncog.2018025946 PMID: 29953367
15. Palacios DA, Miyake M, Rosser CJ. Radiosensitization in prostate cancer: mechanisms and targets. BMC Urol. 2013; 13:4. https://doi.org/10.1186/1471-2490-13-4 PMID: 23351141
16. An J, Chervin AS, Nie A, Ducoff HS, Huang Z. Overcoming the radioresistance of prostate cancer cells with a novel Bcl-2 inhibitor. Oncogene. 2007; 26(5):652–61. https://doi.org/10.1038/sj.onc.1209830 PMID: 16909121
17. Lai CH, Chang CS, Liu HH, Tsai YS, Hsu FM, Yu YL, et al. Sensitization of radio-resistant prostate cancer cells with a unique cytolethal distending toxin. Oncotarget. 2014; 5(14):5523–34. https://doi.org/10.18632/oncotarget.2133 PMID: 25015118
18. Baldwin P, Van De Ven AL, Seitzer N, Clohessy S, Cormack RA, Makrigiorgos M, et al. Radiosensitization in prostate cancer: mechanisms and targets. Int J Radiat Oncol Biol Phys. 2016; 96(2):E595.
19. Chen YA, Lien HM, Kao MC, Lo UG, Lin LC, Lin CJ, et al. Sensitization of Radioresistant Prostate Cancer Cells by Resveratrol Isolated from Arachis hypogaea Stems. PLoS One. 2017; 12(1):e0169204. https://doi.org/10.1371/journal.pone.0169204 PMID: 28081154
20. Hoey C, Ray J, Jeon J, Huang X, Taeb S, Ylanko J, et al. miRNA-106a and prostate cancer radioresistance: a novel role for LITAF in ATM regulation. Mol Oncol. 2018; 12(8):1324–1341. https://doi.org/10.1002/1878-0261.12328 PMID: 29845714
21. Elshafae SM, Hassan BB, Supsavhad WP ans Dirksen W, Camiener RY, Ding H, et al. Gastrin-releasing peptide receptor (GRPr) promotes EMT, growth, and invasion in canine prostate cancer. Prostate. 2016; 76(9):796–809. https://doi.org/10.1002/pros.23154 PMID: 26939805
22. Carlsson SV, Kattan MW. On Risk Estimation versus Risk Stratification in Early Prostate Cancer. PLoS Med. 2016; 13(8): e1002100. https://doi.org/10.1371/journal.pmed.1002100 PMID: 27482892
23. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015; 163(4):1011–25. https://doi.org/10.1016/j.cell.2015.10.025 PMID: 26544944
24. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. Cell. 2015; 161(5):1215–1228. https://doi.org/10.1016/j.cell.2015.05.001 PMID: 26000489
25. Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N Engl J Med. 2016; 375(5):443–53. https://doi.org/10.1056/NEJMoa1603144 PMID: 27433846
26. Hieronymus H, Murali R, Tin A, Yadav K, Abida W, Moller H, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife. 2018;pii: e37294. https://doi.org/10.7554/eLife.37294 PMID: 30178746
27. Mateo J, Boysen G, Barbieri CE, Bryant HE, Castro E, Nelson PS, et al. DNA Repair in Prostate Cancer: Biology and Clinical Implications. Eur Urol. 2017; 71:417–421. https://doi.org/10.1016/j.eururo.2016.08.037 PMID: 27590317
28. Seifert M, Friedrich B, Beyer A. Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. Genome Biology. 2016; 17:204. https://doi.org/10.1186/s13059-016-1058-1 PMID: 27716417
29. Seifert M, Beyer A. regNet: An R package for network-based propagation of gene expression alterations. Bioinformatics. 2018; 34(2):308–311. https://doi.org/10.1093/bioinformatics/btx544
30. Gladitz J, Klink B, Seifert M. Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. Acta Neuropathol Commun. 2018; 6:49. https://doi.org/10.1186/s40478-018-0544-y PMID: 29890994
31. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013; 10:1108–115. https://doi.org/10.1038/nmeth.2651 PMID: 24037242
32. Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015; 47(2):106–114. https://doi.org/10.1038/ng.3168 PMID: 25501392

33. Cojoc M, Peitzsch C, Kurth I, Trautmann F, Kunz-Schughart LA, Telegeev GD, et al. Aldehyde Dehydrogenase Is Regulated by beta-Catenin/TCF and Promotes Radioresistance in Prostate Cancer Progenitor Cells. Cancer Res. 2015; 75(7):1482–94. https://doi.org/10.1158/0008-5472.CAN-14-1924

34. Peitzsch C, Cojoc M, Hein L, Kurth I, K M, Trautmann F, et al. An Epigenetic Reprogramming Strategy to Resensitize Radioresistant Prostate Cancer Cells. Cancer Res. 2016; 76:2637–51. https://doi.org/10.1158/0008-5472.CAN-15-2116 PMID: 26984757

35. Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. Autoregressive Higher-Order Hidden Markov Models: Exploiting Local Chromosomal Dependencies in the Analysis of Tumor Expression Profiles. PLoS One. 2014; 9(6): e100295. https://doi.org/10.1371/journal.pone.0100295 PMID: 24955771

36. Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database. 2010; 2010:baq020. https://doi.org/10.1093/database/baq020 PMID: 20689021

37. Goffart N, Lombard A, Lallemand F, Kroonen J, Nassen J, Di Valentin E, et al. CXCL12 mediates glioblastoma resistance to radiotherapy in the subventricular zone. Neuro Oncol. 2017; 19(1):66–77. https://doi.org/10.1093/neuonc/now136 PMID: 27370398

38. Hoang DT, Iczkowski KA, Kilari D, See W, Nevalainen MT. Androgen receptor-dependent and -independent mechanisms driving prostate cancer progression: Opportunities for therapeutic targeting from multiple angles. Oncotarget. 2017; 8(2):3724–3745. https://doi.org/10.18632/oncotarget.12554 PMID: 27741508

39. Wang T, Huang J, Vue M, Alavian MR, Goel HL, Altieri DC, et al. $\alpha v\beta 3$ Integrin Mediates Radioresistance of Prostate Cancer Cells through Regulation of Survivin. Mol Cancer Res. 2019; 17(2):398–408. https://doi.org/10.1158/1541-7786.MCR-18-0544

40. Michna A, Schötz U, Selmansberger M, Zitzelsberger H, Lauber K, Unger K, et al. Transcriptomic analyses of the radiation response in head and neck squamous cell carcinoma subclones with different radiation sensitivity: time-course gene expression profiles and gene association networks. Radiat Oncol. 2016; 11:94. https://doi.org/10.1186/s13014-016-0672-0 PMID: 27455841

41. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483(7391):603–7. https://doi.org/10.1038/nature11003 PMID: 22460905

42. Tanaka Y, Kanda M, Sugimoto H, Shimizu D, Sueoka S, Takami H, et al. Translational implication of Kallmann syndrome-1 gene expression in hepatocellular carcinoma. Int J Oncol. 2015; 46(6):2546–54. https://doi.org/10.3892/ijo.2015.2965 PMID: 25892360

43. Liu J, Cao W, Chen W, Xu L, Zhang C. Decreased expression of Kallmann syndrome 1 sequence gene (KAL1) contributes to oral squamous cell carcinoma progression and significantly correlates with poorly differentiated grade. J Oral Pathol Med. 2015; 44(2):109–14. https://doi.org/10.1111/jop.12206 PMID: 25060050

44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995; 57:289–300.

45. Storey JD. A direct approach to false discovery rates. J R Statist Soc B. 2002; 64(3):479–498. https://doi.org/10.1111/1467-9868.00346

46. Vandin F, Papoutsaki A, Raphael BJ, Upfal E. Accurate computation of survival statistics in genome-wide studies. PLoS Comput Biol. 2015; 11(5):e1004071. https://doi.org/10.1371/journal.pcbi.1004071 PMID: 25950620

47. Cox D. Regression Models and Life-Tables. J Roy Stat Society Series B. 1972; 34(2):187–220.

48. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics. 1982; 10(4):1100–1120. https://doi.org/10.1214/aos/1176345976

49. Hayashi M, Bernert H, Kagohara LT, Maldonado L, Brait M, et al. Epigenetic inactivation of VGF associated with Urothelial Cell Carcinoma and its potential as a non-invasive biomarker using urine. Oncotarget. 2014; 5(10):3350–61. https://doi.org/10.18632/oncotarget.1949 PMID: 24830820

50. Hwang W, Chiu YF, Kuo MH, Lee KL, Lee AC, Yu CC, et al. Expression of Neuroendocrine Factor VGF in Lung Cancer Cells Confers Resistance to EGFR Kinase Inhibitors and Triggers Epithelial-to-Mesenchymal Transition. Cancer Res. 2017; 77(11):3013–26. https://doi.org/10.1158/0008-5472.CAN-16-3168 PMID: 28381546

51. Marwitz S, Heinbockel L, Scheufele S, Nitschkowski D, Kugler C, Perner S, et al. Epigenetic modifications of the VGF gene in human non-small cell lung cancer tissues pave the way towards enhanced expression. Clin Epigenetics. 2017; 9:123. https://doi.org/10.1186/s13148-017-0423-6 PMID: 29209432

52. Wang X, Prager BC, Wu Q, Kim LJY, Gimple RC, Shi Y, et al. Reciprocal Signaling between Glioblastoma Stem Cells and Differentiated Tumor Cells Promotes Malignant Progression. Cell Stem Cell. 2018; 22(4):514–528. https://doi.org/10.1016/j.stem.2018.03.011 PMID: 29625067

53. Pastrana E, Silva-Vargas V, Doetsch F. Eyes Wide Open: A Critical Review of Sphere-Formation as an Assay For Stem Cells. Cell Stem Cell. 2011; 8(5):486–498. https://doi.org/10.1016/j.stem.2011.04.007 PMID: 21549325

54. Braselmann H, Michna A, Heß J, Unger K. CFAssay: statistical analysis of the colony formation assay. Radiat Oncol. 2015; 10:223. https://doi.org/10.1186/s13014-015-0529-y PMID: 26537797

55. Levi A, Eldridge JD, Paterson BM. Molecular cloning of a gene sequence regulated by nerve growth factor. Science. 1985; 229(4711):393–5.

56. Bartolomucci A, Possenti R, Levi A, Pavone F, Moles A. The role of the vgf gene and VGF-derived peptides in nutrition and metabolism. Genes Nutr. 2007; 2(2):169–80. https://doi.org/10.1007/s12263-007-0047-0 PMID: 18850173

57. Shimazawa M, Tanaka H, Ito Y, Morimoto N, Tsuruma K, Kadokura M, et al. An inducer of VGF protects cells against ER stress-induced cell death and prolongs survival in the mutant SOD1 animal models of familial ALS. PloS One. 2010; 5(12):e15307. https://doi.org/10.1371/journal.pone.0015307 PMID: 21151573

58. Lewis JE, Brameld JM, Jethwa PH. Neuroendocrine Role for VGF. Front Endocrinol (Lausanne). 2015; 6:3. https://doi.org/10.3389/fendo.2015.00003

59. Severini C, Ciotti MT, Biondini L, Quaresima S, Rinaldi AM, Levi A, et al. TLQP-21, a neuroendocrine VGF-derived peptide, prevents cerebellar granule cells death induced by serum and potassium deprivation. J Neurochem. 2008; 104(2):534–44. https://doi.org/10.1111/j.1471-4159.2007.05068.x PMID: 18173805

60. Lu Y, Wang C, Xue Z, Li C, Zhang J, Zhao X, et al. PI3K/AKT/mTOR signaling-mediated neuropeptide VGF in the hippocampus of mice is involved in the rapid onset antidepressant-like effects of GLYX-13. Int J Neuropsychopharmacol. 2014; 18(5), pii:pyu110. https://doi.org/10.1093/ijnp/pyu110

61. Wu W, Hodges E, Redelius J, Höög C. A novel approach for evaluating the efficiency of siRNAs on protein levels in cultured cells. Nucleic Acids Res. 2004; 32(2):e17. https://doi.org/10.1093/nar/gnh010 PMID: 14739231

62. Hong SW, Jiang Y, Kim S, Li CJ, Lee DK. Target gene abundance contributes to the efficiency of siRNA-mediated gene silencing. Nucleic Acid Ther. 2014; 24(3):192–8. https://doi.org/10.1089/nat.2013.0466 PMID: 24527979

63. Pauli C, Hopkins BD, Prandi D, Shaw R, Fedrizzi T, Sboner A, et al. Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. Cancer Discov. 2017; 7(5):462–477. https://doi.org/10.1158/2159-8290.CD-16-1154 PMID: 28331002

64. Liu Y, Mi Y, Mueller T, Kreibich S, Williams EG, et al. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat Biotechnol. 2019; 37(3):314–322. https://doi.org/10.1038/s41587-019-0037-y PMID: 30778230

65. Aguilera TA, Rafat M, Castellini L, Shehade H, Kariolis MS, Hui AB, et al. Reprogramming the immunological microenvironment through radiation and targeting Axl. Nat Commun. 2016; 7:13898. https://doi.org/10.1038/ncomms13898 PMID: 28008921

66. Slovin SF, Higano CS, Hamid O, Tejwani S, Harzstark A, Alumkal JJ, et al. Ipilimumab alone or in combination with radiotherapy in metastatic castration-resistant prostate cancer: results from an open-label, multicenter phase I/II study. Ann Oncol. 2013; 24(7):1813–21. https://doi.org/10.1093/annonc/mdt107 PMID: 23535954

67. Liebscher S, Koi L, Löck S, Muders MH, Krause M. The HIV protease and PI3K/Akt inhibitor nelfinavir does not improve the curative effect of fractionated irradiation in PC-3 prostate cancer in vitro and in vivo. Clin Transl Radiat Oncol. 2017; 2:7–12. https://doi.org/10.1016/j.ctro.2016.12.002 PMID: 29657993

68. Abou-El-Ardat K, Seifert M, Becker K, Eisenreich S, Lehmann M, Hackmann K, et al. Comprehensive molecular characterization of multifocal glioblastoma proves its monoclonal origin and reveals novel insights into clonal evolution and heterogeneity of glioblastomas. Neuro Oncol. 2017; 19(4):546–57. https://doi.org/10.1093/neuonc/now231 PMID: 28201779

69. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007; 23(6):657–63. https://doi.org/10.1093/bioinformatics/btl646 PMID: 17234643

70. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19(2):185–193. https://doi.org/10.1093/bioinformatics/19.2.185 PMID: 12538238

71. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. https://doi.org/10.1186/1471-2105-12-323 PMID: 21816040

72. Tibshirani R. Shrinkage and Selection via the Lasso. J R Stat Soc B. 1996; 58:267–288.

73. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. 2014; 42:413–468. https://doi.org/10.1214/13-AOS1175 PMID: 25574062

74. Therneau TM. A Package for Survival Analysis in S; 2015. https://CRAN.R-project.org/package=survival.

5 Discussion

My habilitation thesis contains seven selected publications that are all focused on different important topics in computational cancer omics data analysis. The overarching link between these publications is given by the application of a newly developed concept for gene network inference to identify potential major regulators that distinguish cancer subtypes in combination with a newly developed concept for network propagation to quantify impacts of altered genes on clinically relevant characteristics. I started to work on these topics in 2012 and contributed substantially to all included publications that were published over the last years.

When I started with this work it was already widely accepted that cancer is a complex genetic disease that is driven by combinations of mutated genes that alter cellular hallmark pathways that contribute to cancer development (Hanahan and Weinberg (2011)). Due to great efforts to analyze cancer genomes of thousands of patients, frequently and rarely mutated genes had largely been cataloged for all major types of human cancer (Vogelstein et al. (2013); Lawrence et al. (2014); The Cancer Genome Atlas Research Network (2013b)). One important finding was that the vast majority of mutated genes only occurred in some patients, whereas generally only few frequently mutated genes were observed for specific cancer types (Vogelstein et al. (2013); The Cancer Genome Atlas Research Network (2013b)).

Importantly, functional roles and clinical implications of frequently mutated genes can be studied within large cancer cohorts with the help of existing statistical methods, but methods for the analysis of rarely mutated genes were largely missing. It was also not possible to analyze the impact of all mutated genes of a patient-specific cancer on clinically relevant characteristics. This whole situation was further complicated by the fact that many cancers also showed DNA copy number alterations, chromosomal instability and epigenetic alterations (Berdasco and Esteller (2010); Hanahan and Weinberg (2011); Ciriello et al. (2013); Zack et al. (2013)). All these alterations can contribute to the existences of different cancer subtypes and complex alterations of cancer transcriptomes.

This huge complexity of cancer genomes strongly complicates the identification of

driver mutations from the specific set of mutations that is found in an individual cancer. Novel computational strategies were urgently required to predict major regulators that distinguish different cancer subtypes and to determine how these genes impact on pathogenesis and therapy response. A promising way to realize this was to consider cancer as a disease of cellular pathways and networks and to utilize this idea to develop novel computational approaches for the analysis of individual cancer patients (Krogan et al. (2015)). A first great success in this direction was reached by Chuang et al. (2007) for the classification of the metastatic potential of breast cancer with the help of protein interaction networks. Over the years, such approaches for the analysis of cancer data were more widely used and extended leading to the development of the research field of network medicine (Barabási et al. (2011)).

An important contribution to computational network medicine was made by Hofree et al. (2013) utilizing network propagation of individual cancer mutations for the identification of clinically relevant cancer subtypes that were composed of tumors that only rarely shared the same gene mutations. Also Leiserson et al. (2015) used network propagation for a pan-cancer analysis revealing cancer-relevant sub-networks that included many genes that were only rarely mutated across different cancer types. Both approaches utilized existing protein or gene interaction networks to analyze gene mutation data of individual cancers, but they did not include tumor gene expression profiles, which can provide important information about the structure and activity of gene regulatory networks, cancer subtypes or signaling pathway alterations.

To fill this gap, I developed a novel computational approach to directly learn cancer-specific gene regulatory networks from molecular data of individual cancers or cancer cells with the goal to use the resulting network for network propagation under consideration of individual tumor expression profiles (Seifert et al. (2016)). This computational framework enabled for the first time an in-depth analysis of all tumor-specific gene copy number and expression alterations on clinically relevant characteristics for individual patients. The focus on the combined analysis of gene copy number and corresponding gene expression profiles was driven by the specific research questions of the publications included in this habilitation thesis, but the learned cancer type-specific gene regulatory networks may also be considered for the propagation of gene mutations as done in Hofree et al. (2013) and Leiserson et al. (2015).

I started to develop my computational approach for network inference in 2012. In a first application study, we utilized the network inference approach to identify potential major regulators within the large gene expression signature that we had determined

to distinguish astrocytomas in childhood (pilocytic astrocytomas) from those in adulthood (diffuse and anaplastic astrocytomas and glioblastomas) (Seifert et al. (2015), see Section 4.1). The derived signature-specific gene regulatory network had purely been learned based on gene expression data. Predicted major regulators had known functions in important biological processes including brain development, cell cycle, proliferation, apoptosis and epigenetic regulation. Expression differences between both groups of astrocytomas were mainly explained by DNA methylation changes and gene copy number alterations. Overall, this in-depth study represents one of the first large-scale computational comparisons of molecular data of all four major astrocytoma types (Louis et al. (2007)) substantially extending related studies in the early 2000s (Rickman et al. (2001); Hunter et al. (2002); Rorive et al. (2006)). Moreover, the integration of the gene network inference approach clearly demonstrated the potential to predict major regulators out of hundreds of differentially expressed genes. Such major regulators can contribute to a better understanding of differences between astrocytoma types and also provide a basis for additional experimental studies.

Motivated by this first success, I suggested to utilize my network inference approach for the identification of major regulators that distinguish subtypes of histologically classified oligodendrogliomas (Lauber et al. (2018), see Section 4.2). In this study, we learned oligodendroglioma-specific gene regulatory networks that distinguished oligodendrogliomas with a 1p/19q co-deletion and an IDH mutation from those that only had an IDH mutation. Importantly, we found that networks learned from gene expression and corresponding gene copy number profiles were clearly better suited to predict the expression levels of individual signature genes than networks that were only learned from gene copy number data alone. Focusing on the most significant highly recurrent links between signature genes, the predicted major regulators had different important biological functions in cytoskeleton remodeling, apoptosis and neural development. Interestingly, the obtained network also showed characteristic differences of several HOX and SOX transcription factors between both oligodendroglioma subgroups. This suggested that different glioma stemness programs were potentially active in each subgroup, which was also supported by single cell oligodendroglioma transcriptome analyses that were published during the work on this study (Tirosh et al. (2016); Venteicher et al. (2017)). Overall, this study demonstrated that gene copy number profiles alone are sufficient to derive known molecular subgroups of histologically classified oligodendrogliomas. The identified molecular signatures and major regulators provide a good basis for further studies and experimental validations. Specifically focusing on

the network approach, this study also showed that the predictive power of the network inference approach can be improved substantially by the integrative analysis of gene expression and corresponding gene copy number profiles of signature genes in comparison to only using gene copy number data alone. In addition, important methodological advancements in comparison to Seifert et al. (2015) were that the network inference was repeated 100 times on different training sets along with the evaluation of the prediction quality of the individual networks on their corresponding test sets. This enabled us to focus on the most stable links between genes to improve the generalization capacity of the final network. Since network inferences are typically time consuming (Seifert et al. (2016); Seifert and Beyer (2018)), such large numbers of repeated network inferences were only possible for gene expression signatures but not for the genome-wide network inferences approaches presented in this thesis.

Similarly, I also suggested to utilize my network inference approach to predict major regulators that distinguished short- from long-lived DNMT3A-mutant acute myeloid leukemia patients (Lauber et al. (2020), see Section 4.3). In this study, we learned gene regulatory networks for the gene expression signature that distinguished both patient classes. Network inference was done based on gene expression and microRNA profiles. The integration of microRNAs as predictors slightly improved the network prediction quality and also slightly increased the number of predictable signature genes. We again also considered separations into training and test sets and repeated the network inference 100 times to focus on the most significant stable links. An important improvement compared to Lauber et al. (2018) was that we were now better able to account for multiple testing in relation to the network links using FDR-adjusted p-values based on the method by Benjamini and Hochberg (1995) as implemented in regNet (Seifert and Beyer (2018)). In my two prior studies Seifert et al. (2015) and Lauber et al. (2018), we could only focus on the most significant network links at the detection limit of the covariance test due to the implicit and undocumented rounding of the p-values to four digits in the R package covTest by Lockhart et al. (2013). I modified the implementation of the covariance test to overcome this limitation. The obtained gene regulatory network contained potential major regulators including several genes and microRNAs that were already known to be involved in the pathogenesis of acute myeloid leukemia. Interestingly, we also identified novel candidate genes with known functions in the regulation of hematopoiesis, cell cycle, cell differentiation and immunity. Moreover, we could also show that our revealed gene mutation and expression signatures were also predictive for independent DNMT3A-mutant acute myeloid leukemia patients from other cohorts.

Further, our findings also suggest that our predictions can be used to improve currently used clinical prognostic scoring systems (Döhner et al. (2010, 2017)). This is in good accordance and further extends recent findings by Herold et al. (2020).

Thus, the three included publications (Seifert et al. (2015); Lauber et al. (2018, 2020)) clearly showed that network inference enables a more detailed analysis of gene expression signatures to identify potential major regulators that distinguish cancer subtypes. Such major regulators are frequently associated with clinically relevant characteristics and also potentially influence the expression behavior of many other signature genes. Therefore, major regulators determined by the network inference approach can represent promising targets for additional perturbation experiments in the wet lab.

In the following, the discussion shifts from networks obtained for gene expression signatures to genome-wide network inference and network propagation. In 2012, I started to develop the network-based framework for the joint analysis of gene expression and copy number profiles of individual cancers with the goal to enable a risk stratification of altered genes on clinically relevant characteristics. The underlying mathematical concepts for network inference and network propagation, which also formed the general basis of all network-based analysis in the studies that are part of this habilitation thesis, were published in Seifert et al. (2016) along with applications to different cancer types and in-depth validation studies (see Section 4.4). In this study, we learned a genome-wide gene regulatory network based on gene expression and corresponding gene copy number data of 768 cancer cell lines. Importantly, we demonstrated that this network was highly predictive for expression of genes of more than 4,500 cancer samples of 13 different cancers. In addition, we further showed that this network can quantify impacts of patient-specific gene copy number alterations on patient survival with the help of the newly developed network propagation algorithm. Such patient-specific risk classifications of individual gene copy number alterations are only possible with the help of the specifically designed network propagation algorithm and cannot be realized with existing statistical tests or methods for the analysis of gene copy number data. The value of our network propagation approach, which integrates direct and indirect impacts of gene copy number alterations on the expression of clinically relevant target genes, had also been demonstrated by a comparison to a closely related reduced network approach that only accounted for direct impacts. This finding also supports that local network neighborhood approaches should better be replaced by network propagation algorithms (Cowen et al. (2017)).

Overall, the developed network-based framework (Seifert et al. (2016)) contributes to

a patient-specific risk classification including the possibility to identify the most important altered genes along with their downstream targets. Such a personalized analysis has not been possible before for the integrative analysis of gene expression and gene copy number data. Related approaches based on networks (e.g. Akavia et al. (2010); Carro et al. (2010); Jörnsten et al. (2011)) or genetic linkage analysis (Adler et al. (2006)) had already been developed before to identify cancer-relevant major regulators, but none of these methods can realize impact quantifications for rarely mutated genes. Our approach can analyze all altered genes of a patient-specific tumor including frequent and rare alterations. Further, in contrast to the network propagation studies by Hofree et al. (2013) and Leiserson et al. (2015), our network approach does not rely on existing protein or gene interaction networks. Our considered networks were directly learned for cancer cell lines or specific cancer types that potentially better reflect the presence and activity of regulatory links between genes in a specific cancer type than existing networks compiled based on data from cells of different tissues. The required network inference step increases the computational demands. Genome-wide network inference and network propagation can only be efficiently done on a compute server with many cores in parallel (e.g. network inference took in total about 140 days of computing time for one network instance and network propagation analysis took about 24 hours per patient). Nevertheless, the great success of our study showed that it is worth to invest this to improve predictions for individual patients and to learn more about the roles of rarely mutated genes.

Focusing on clinically relevant predictions, the application of our network-based approach (Seifert et al. (2016)) led to the following novel insights. We observed that the predicted regulatory links between genes were surprisingly well conserved across tumors from different tissues, which had not been reported before. We also found that up to 100 patient-specific gene copy number alterations influence the survival of a patient. Interestingly, this number of genes is strongly greater than traditional assumptions that up to 10 genes contribute to development of a tumor (Vogelstein et al. (2013)). Thus, many mutated genes can contribute to the aggressiveness of a tumor. This observation is in accordance with Davoli et al. (2013). We further discovered that some gene copy number alterations also have beneficial contributions that increased patient survival. I also made a similar observation within the frame of the glioblastoma case study of my R package regNet (Seifert and Beyer (2018)). Most important for personalized cancer medicine is that our study also showed that rare gene copy number alterations can be as important as frequent gene copy number alterations. Finally, we did not stop at

the descriptive level and found that genomic features explained why certain genes with high impact were actually rarely affected by gene copy number alterations in specific cancer types. All these results were obtained from data that underwent rigorous quality checks. We also performed thousands of computations to analyze the robustness of our results. We further validated each step and established clinical relevance by the usage of independent test cohorts.

The great value and the generally broad scope of application of the network-based approach in Seifert et al. (2016) further contributed to finish the development of the R package regNet (Seifert and Beyer (2018), see Section 4.5). Initial regNet source code developments started in 2012. Major parts of this code were used to realize the network analysis that are part of all publications included in this habilitation thesis. regNet represents a user-friendly tool that implements the basic network inference algorithm and corresponding network propagation algorithms developed in Seifert et al. (2016). The implemented data flow management together with the pre-defined automatic naming scheme of created files allows users to realize their own studies. Since regNet is open-source, extensions or adaptations of the source code for specific requirements are easily possible. The two included case studies with code examples demonstrate how regNet can be used to identify major regulators within a signature of differentially expressed genes and how regNet can be used to realize impact quantifications on a genome-wide scale.

The first study that we fully realized with the help of the R package regNet had the goal to identify novel cancer gene candidates within the region of the 1p/19q co-deletion of oligodendrogliomas (Gladitz et al. (2018), see Section 4.6). Therefore, we learned oligodendroglioma-specific gene regulatory networks using publicly available gene expression and copy number data of 178 patients. We learned 10 genome-wide networks on different training sets and evaluated their predictive power on the corresponding test sets and independent oligodendroglioma samples from other studies. We used these networks to compute impacts of differentially expressed genes within the region of the 1p/19q co-deletion on cancer-relevant signaling and metabolic pathways. Comparisons to impacts of random networks of same complexity enabled us to predict 8 genes with strong impact on signaling pathways and 14 genes with strong impact on metabolic pathways. Literature analysis suggested that many of these genes have the potential to push or counteract oligodendroglioma development. Overall, this network-based study suggested novel driver gene candidates that could contribute to a better understanding of the pathology of the 1p/19q co-deletion.

The 1p/19q co-deletion, which has been in the main focus of the analysis by Gladitz et al. (2018), is an important clinically relevant molecular marker for oligodendrogliomas (Louis et al. (2016)). Nevertheless, only little progress was made in the identification of potential driver genes within the region of the 1p/19q co-deletion to better understand alterations of molecular mechanisms that drive oligodendroglioma development (Bettegowda et al. (2011); Eisenreich et al. (2013)). The main challenge was that hundreds of genes on 1p and 19q are affected by the combined loss of one copy of these chromosomal arms. Therefore, we could not simply distinguish between driver and passenger genes utilizing standard statistical tests or bioinformatics methods for gene expression and copy number data analysis. All oligodendrogliomas show almost identical co-deletions, which does not allow to narrow down to specific chromosomal regions on 1p or 19q to pinpoint potential driver genes. Further, hundreds of genes on 1p and 19q are differentially expressed, which does not allow to select driver genes without biological prior knowledge about altered molecular processes that drive oligodendroglioma development. Thus, this complex situation defined an ideal clinically relevant application study to further highlight the potential of the underlying network approach (Seifert et al. (2016)). Network propagation applied to oligodendroglioma-specific gene regulatory networks enabled us to predict candidate genes that are potentially associated with the development of oligodendrogliomas. This greatly extends the prior analysis of oligodendrogliomas that is part of this habilitation thesis (Lauber et al. (2018)). Our study also suggests candidate genes for functional validations that could be tested in wet lab experiments. However, cell cultures or xenografts of oligodendrogliomas did not exist for experimental validations at the time of our study, but recent progress by Exner et al. (2019) suggests that oligodendroglioma xenografts might become available for validation experiments.

In addition, I also considered the R package regNet for the analysis of prostate cancer cell lines to identify novel candidate genes associated with radioresistance and relapse of prostate cancer patients (Seifert et al. (2019), see Section 4.7). We learned a genome-wide prostate cancer-specific gene regulatory network from publicly available gene expression and copy number profiles of 541 prostate cancer patients. Considering differentially expressed genes with directly underlying copy number alterations from two radioresistant prostate cancer cell lines, we used this network to compute impacts on known radioresistance marker genes. We predicted 14 potential driver genes that were able to separate irradiated prostate cancer patients into early and late relapse groups. In-depth experimental validations of one selected candidate gene suggested

that our network-based approach is able to identify genes that potentially contribute to radioresistance of prostate cancer.

Radiation therapy is a very effective treatment for many prostate cancer patients, but patients can relapse due to radioresistance of prostate cancer cells that survive the maximal radiation dose that can safely be delivered to the tumor (Chang et al. (2014); Chaiswing et al. (2018)). To identify genes involved in radioresistance, radioresistant prostate cancer cell lines are typically compared to their radiosensitive parental cells (Cojoc et al. (2015); Peitzsch et al. (2016)). Since hundreds or thousands of genes are usually affected by irradiation induced cell line-specific DNA copy number alterations, the key challenge was to find a strategy how one can distinguish between driver and passenger alterations. This cannot be realized by existing statistical tests or basic methods for the analysis of gene expression and copy number profiles, because of the very low number of cell lines that are typically considered and also because each cell line has its unique alterations that require an individual analysis of each cell line to better explore the set of potential radioresistant driver genes. Thus, this was again an ideal use case for the developed network-based approach (Seifert et al. (2016)). Network propagation enabled us to predict driver gene candidates with the help of a prostate cancer-specific gene regulatory network. Together with the study by Gladitz et al. (2018), this study demonstrated again the great value and broad applicability of the approaches for network inference and network propagation developed in Seifert et al. (2016) for the integrative analysis of genome-wide gene expression and copy number profiles to predict clinically relevant marker candidates.

In summary, the overarching connection between the different manuscripts that form that basis of this habilitation thesis has been the developed network-based approach together with its different applications for the integrative analysis of molecular cancer data. Still, methodological advancements like the integration of stability selection (Meinshausen and Bühlmann (2010)) or the modeling of interactions between selected genes (Lim and Hastie (2015)) could potentially improve the current network approach. Also a transfer of the network approach to single cell data has great potential. First studies in these directions have already been done by Leote et al. (2019) to replace dropouts in single cell RNA-seq data. It would also be very interesting to extend the oligodendroglioma study by Gladitz et al. (2018) to the single cell level using data from Tirosh et al. (2016). Obviously, the presented studies are not the end, but just the starting point of a new strategy for the analysis of molecular cancer data. I have substantially contributed to these developments with my research over the last years and hope to continue to further extend and apply this promising approach in future projects.

English summary

Cancer is a very complex genetic disease driven by combinations of mutated genes. This complexity strongly complicates the identification of driver genes and puts enormous challenges to reveal how they influence cancerogenesis, prognosis or therapy response. Thousands of molecular profiles of the major human types of cancer have been measured over the last years. Apart from well-studied frequently mutated genes, still only little is known about the role of rarely mutated genes in cancer or the interplay of mutated genes in individual cancers. Gene expression and mutation profiles can be measured routinely, but computational methods for the identification of driver candidates along with the prediction of their potential impacts on downstream targets and clinically relevant characteristics only rarely exist. Instead of only focusing on frequently mutated genes, each cancer patient should better be analyzed by using the full information in its cancer-specific molecular profiles to improve the understanding of cancerogenesis and to more precisely predict prognosis and therapy response of individual patients. This requires novel computational methods for the integrative analysis of molecular cancer data.

A promising way to realize this is to consider cancer as a disease of cellular networks. I have developed a novel network-based approach for the integrative analysis of molecular cancer data over the last years. This approach directly learns gene regulatory networks form gene expression and copy number data and further enables to quantify impacts of altered genes on clinically relevant downstream targets using network propagation. In this habilitation thesis, I summarize the results of seven publications to which I have contributed substantially as first or last author. All publications have their focus on the integrative analysis of molecular data of different cancer types along with the overarching connection to the application of the newly developed network-based approach. In the first three included publications, networks were learned to identify major regulators that distinguish cancer subtypes enabling a more detailed analysis of characteristic gene expression signatures of known astrocytoma entities, of revealed oligodendroglioma subtypes, and of subgroups of acute myeloid leukemia patients with profound survival differences. Next, I introduce the central publication of this habilita-

tion thesis that combines network inference with network propagation. I demonstrate the great value of this approach by quantifying potential direct and indirect impacts of rare and frequent gene copy number alterations on survival of individual cancer patients. Further, I introduce the publication of my R package regNet that represents a user-friendly implementation of the network-based approach. Finally, I included two publications that further strongly highlight the value of the developed network-based approach for the personalized analysis of individual cancer omics profiles. In more detail, I show how we predicted cancer gene candidates within the region of the 1p/19q co-deletion of oligodendrogliomas and I introduce how I determined driver candidates associated with radioresistance and relapse of prostate cancer patients.

All seven publications that form the core of this habilitation thesis are embedded into a brief introduction that motivates the scientific background and the major objectives of this thesis. The scientific background is briefly going from the hallmarks of cancer over the complexity of cancer genomes down to the importance of networks in cancer. This also includes a short introduction of the mathematical concepts that underlie the developed network inference and network propagation algorithms that were used in the different publications. Further, I briefly motivate and summarize my studies before the presentation of the original publications. The habilitation thesis is completed with a general discussion of the major results with a specific focus on the utilized network-based data analysis strategies. Major biologically and clinically relevant findings of each included publication are also briefly summarized.

In summary, in this habilitation thesis I demonstrate the great value of the application of network-based data analysis for the prediction of clinically relevant alterations in molecular cancer profiles. The underlying methodological developments and the predictions in the different application studies represent important contributions to the research field of computational cancer medicine. This is also supported by experimental validations of predictions that were made by collaboration partners for two included publications to strengthen the biological relevance our findings and to highlight the potential of the network-based approach. Overall, the included methodological developments and the specific application studies represent new promising strategies for the integrative analysis of molecular data from individual cancer patients.

Deutsche Zusammenfassung

Krebs ist eine sehr komplexe genetische Krankheit, die durch Mutationen von Genen im Erbgut ausgelöst wird. Die Vielzahl von Mutationen und deren komplexes Zusammenspiel erschwert die Identifizierung von veränderten Genen, die zur Krebsentstehung beitragen, erheblich. Diese Komplexität macht auch die Ermittlung von genspezifischen Einflüssen auf die Prognose oder das Therapieverhalten zu einer sehr großen Herausforderung. In den letzten Jahren wurden zwar tausende molekulare Profile der wichtigsten Krebsarten gemessen, aber abgesehen von meist gut untersuchten häufig mutierten Genen ist bisher kaum etwas über die Bedeutung von selten mutierten Genen oder deren Zusammenspiel mit anderen mutierten Genen bekannt. Mittlerweile können Genexpressions- und Mutationsprofile routinemäßig gemessen werden, aber an innovativen rechnergestützten Verfahren, die diese Daten umfassend analysieren und aus der Vielzahl an molekularen Veränderungen potenziell ursächliche Gene und deren Auswirkung auf andere Gene und klinisch relevante Eigenschaften direkt ermitteln können, mangelt es bisher noch. Anstatt sich, wie bisher meist üblich, hauptsächlich auf häufig mutierte Gene zu konzentrieren, sollten zukünftig möglichst alle verfügbaren Informationen von krebsspezifischen molekularen Profilen genutzt werden, um jeden Krebspatienten noch stärker individuell betrachten zu können. Dies kann zu einem besseren Verständnis der molekularen Ursachen der Krebsentstehung beitragen und darüber hinaus eine präzisere Prognose und bessere Therapieauswahl für Patienten ermöglichen. Für die Umsetzung dieses Vorgehens ist die Entwicklung von innovativen rechnergestützten Verfahren für die integrative Analyse von molekularen Krebsdaten notwendig.

Ein vielversprechender Ansatz zur Realisierung dieses Konzepts ist die Betrachtung von Krebs als eine Erkrankung von zellulären Regulationsnetzwerken. Ich habe in diesem Kontext in den letzten Jahren ein netzwerkbasiertes Verfahren für die integrative Analyse von molekularen Krebsdaten entwickelt. Dieses Verfahren lernt Genregulationsnetzwerke direkt aus Tumorexpressions- und Kopiezahldaten von Genen und ermöglicht darüber hinaus die Quantifizierung von Einflüssen von veränderten Genen auf klinisch relevante Zielgene durch Netzwerkflussanalysen. Im Rahmen dieser Ha-

bilitationsschrift fasse ich die Ergebnisse von sieben Publikationen zusammen, zu denen ich als Erst- oder Letztautor maßgeblich beigetragen habe. Der Fokus liegt dabei auf der integrativen Analyse molekularer Daten für verschiedene Krebsarten, wobei alle Arbeiten übergreifend durch die spezifische Anwendung des von mir entwickelten netzwerkbasierten Verfahrens miteinander in Verbindung stehen. In den ersten drei aufgenommenen Publikationen wurden Netzwerke gelernt, um Hauptregulatoren zu bestimmen, die Subtypen von Krebserkrankungen gezielt unterscheiden. Dies ermöglichte uns eine detailliertere Analyse charakteristischer Genexpressionssignaturen von bereits bekannten Astrozytomentitäten und identifizierten Oligodendrogliomsubtypen. Weiterhin war es dadurch möglich, Patienten mit akuter myeloischer Leukämie, die starke Unterschiede in der Überlebenszeit aufwiesen, besser zu charakterisieren. Im Anschluss daran stelle ich die zentrale Publikation dieser Habilitationsschrift vor, die Netzwerkinferenz mit Netzwerkflussanalysen verknüpft. Dabei demonstriere ich den großen Nutzen dieses Ansatzes durch die Quantifizierung von potenziell existierenden direkten und indirekten Einflüsse von seltenen und häufigen Genkopiezahlveränderungen auf das Überleben von Krebspatienten. Nachfolgend präsentiere ich die Publikation zu dem von mir entwickelten R Paket regNet, das eine nutzerfreundliche Implementierung des netzwerkbasierten Verfahrens beinhaltet. Abschließend habe ich noch zwei Publikationen aufgenommen, die nochmals stark den großen Nutzen des entwickelten netzwerkbasierten Verfahrens für die personalisierte Analyse von molekularen Krebsprofilen hervorheben. Dabei zeige ich, wie wir vielversprechende Krebsgenkandidaten im Bereich der 1p/19q Kodeletion von Oligodendrogliomen bestimmt haben. Ich demonstriere darüber hinaus auch, wie ich potenzielle Kandidatengene, die mit der Entwicklung einer Radioresistenz und damit verbundenen Rezidiven von Prostatakarzinomen in Verbindung stehen können, ermittelt habe.

Alle sieben Publikationen, die den Kern dieser Arbeit bilden, werden zu Beginn der Habilitationsschrift durch eine Einführung motiviert, die den wissenschaftlichen Hintergrund und die Hauptziele der Arbeit darlegt. Dabei führt der wissenschaftliche Hintergrund von einem kurzen Überblick über die Hauptmerkmale der Krebsentstehung über die Komplexität von Krebsgenomen bis hin zur Bedeutung von zellulären Netzwerken bei Krebs. Dies beinhaltet auch eine kurze Einführung in die mathematischen Konzepte, die den entwickelten Netzwerkinferenz- und Netzwerkflussalgorithmen zu Grunde liegen und im Rahmen der Publikationen angewendet wurden. Zudem werden die durchgeführten Studien motiviert und deren Resultate kurz zusammengefasst. Die Habilitationsarbeit wird mit einer allgemeinen Diskussion der wichtigsten Ergebnisse

abgeschlossen, wobei ein besonderer Schwerpunkt auf den eingesetzten netzwerkbasierten Datenanalysestrategien liegt. Dabei werden auch die wichtigsten biologisch und klinisch relevanten Ergebnisse der sieben Publikationen kurz zusammengefasst.

Die Habilitationsschrift demonstriert den großen Nutzen der Anwendung von netzwerkbasierten Datenanalysestrategien für die Prädiktion klinisch relevanter Veränderungen in molekularen Krebsprofilen. Die durchgeführten methodischen Entwicklungen und deren konsequente Anwendung auf spezifische medizinische Fragestellungen stellen wichtige Beiträge im Forschungsbereich der rechnergestützten Krebsforschung dar. Dies wird auch durch die experimentelle Validierungen von Prädiktionen gestützt, die von Kooperationspartnern für zwei Veröffentlichungen durchgeführt wurden, um die biologische und medizinische Relevanz unserer Ergebnisse zu stärken und das Potenzial des netzwerkbasierten Ansatzes noch deutlicher hervorzuheben. Somit stellt diese Arbeit einen wichtigen Beitrag zur Entwicklung neuer vielversprechender Strategien für die integrative Analyse molekularer Daten von Krebspatienten dar.

Bibliography

Abou-El-Ardat, K., Seifert, M., Becker, K., Eisenreich, S., Lehmann, M., Hackmann, K., et al. (2017). Comprehensive molecular characterization of multifocal glioblastoma proves their monoclonal origin and reveals novel insights into clonal evolution and heterogeneity of glioblastomas. *Neuro-Oncology*, 19(4):546–557.

Adler, A. S., Lin, M., Horlings, H., Nuyten, D. S. A., van de Vijver, M. J., and Chang, H. Y. (2006). Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet.*, 38(4):421–30.

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017.

Alfonso, J. C. L., Talkenberger, K., Seifert, M., Klink, B., Hawkins-Daarud, A., Swanson, K. R., et al. (2017). The biology and mathematical modelling of glioma invasion: a review. *J. R. Soc. Interface*, 14:20170490.

Aspuria, P. P., Lunt, S. Y., Väremo, L., Vergnes, L., Gozo, M., Beach, J. A., et al. (2014). Succinate dehydrogenase inhibition leads to epithelial-mesenchymal transition and reprogrammed carbon metabolism. *Cancer Metab*, 2:21.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., et al. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol.*, 21(11):1337–42.

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet.*, 12(1):56–68.

Barabási, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet.*, 5(2):101–13.

Barker, H., Paget, J. T., Khan, A. A., and Harrington, K. J. (2015). The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. *Nat Rev Cancer*, 15(7):409–25.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–7.

Baysal, B. E. and Maher, E. R. (2015). 15 YEARS OF PARAGANGLIOMA: genetics and mechanism of pheochromocytoma-paraganglioma syndromes characterized by germline SDHB and SDHD mutations. *Endocr Relat Cancer*, 22(4):T71–82.

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

Berdasco, M. and Esteller, M. (2010). Aberrant epigenetic landscape in cancer: How cellular identity goes awry. *Dev. Cell*, 19(5):698–711.

Bettegowda, C., Agrawal, N., Jiao, Y., Sausen, M., Wood, L. D., Hruban, R. H., et al. (2011). Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science*, 333(6048):1453–1455.

Biedermann, J., Preussler, M., Conde, M., Peitzsch, M., Richter, S., Wiedemuth, R., et al. (2019). Mutant IDH1 Differently Affects Redox State and Metabolism in Glial Cells of Normal and Tumor Origin. *Cancers*, 11(12):2028.

Bonkhoff, H. (2012). Factors implicated in radiation therapy failure and radiosensitization of prostate cancer. *Prostate Cancer*, 2012:593241.

Brohée, S. and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488.

Bulfone, A., Smiga, S., Shimamura, K., Peterson, A., Puelles, L., and Rubenstein, J. (1995). T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron*, 15:63–78.

Cairncross, J. G., Ueki, K., Zlatescu, M. C., Lisle, D. K., Finkelstein, D. M., Hammond, R. R., et al. (1998). Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *J Natl Cancer Inst*, 90(19):1473–9.

Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):1011–25.

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Synder, E. Y., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(21):318–325.

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39:D685–690.

Chai, L., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Comput Biol Med.*, 48:55–65.

Chaiswing, L., Weiss, H. L., Jayswal, R. D., Clair, D. K. S., and Kyprianou, N. (2018). Profiles of Radioresistance Mechanisms in Prostate Cancer. *Crit Rev Oncog.*, 23(1-2):39–67.

Challen, G. A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J., et al. (2011). Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet*, 44(1):23–31.

Chang, L., Graham, P. H., Hao, J., Bucci, J., Cozzi, P. J., Kearsley, J. H., et al. (2014). Emerging roles of radioresistance in prostate cancer metastasis and radiation therapy. *Cancer Metastasis Rev*, 33(2-3):469–96.

Chen, X., Liu, M. X., and Yan, G. Y. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst.*, 8(7):1970–8.

Cho, D. Y., Kim, Y. A., and Przytycka, T. M. (2012). Chapter 5: Network biology approach to complex diseases. *PLoS Comput Biol.*, 8(12):e1002820.

Chuang, H., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol.*, 3:140:175–185.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, 22(2):398–406.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, 47(10):1127–1133.

Cohen, A., Holmen, S., and Colman, H. (2013). IDH1 and IDH2 mutations in gliomas. *Curr Neurol Neurosci Rep*, 13(5):345.

Cojoc, M., Peitzsch, C., Kurth, I., Trautmann, F., Kunz-Schughart, L. A., Telegeev, G. D., et al. (2015). Aldehyde Dehydrogenase Is Regulated by beta-Catenin/TCF and Promotes Radioresistance in Prostate Cancer Progenitor Cells. *Cancer Res*, 75(7):1482–94.

Coons, S. W., Johnson, P. C., Scheithauer, B. W., Yates, A. J., and Pearl, D. K. (1997). Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. *Cancer*, 79:1381–1393.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet.*, 18(9):551–562.

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther.*, 138(3):333–408.

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., and Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat Rev Microbiol.*, 8(10):717–29.

Deshmukh, H., Yu, J., Shaik, J., MacDonald, T. J., Perry, A., Payton, J. E., et al. (2011). Identification of transcriptional regulatory networks specific to pilocytic astrocytoma. *BMC Med Genomics*, 4:57.

Di Lorenzo, G., Buonerba, C., and Kantoff, P. W. (2011). Immunotherapy for the treatment of prostate cancer. *Nat Rev Clin Oncol.*, 8(9):551–61.

Ding, J., McConechy, M. K., Horlings, H. M., Ha, G., Chun Chan, F., Funnell, T., et al. (2015). Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.*, 6:8554.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129(4):424–447.

Döhner, H., Estey, E. H., Amadori, S., Appelbaum, F. R., Büchner, T., Burnett, A. K., et al. (2010). Diagnosis and Management of Acute Myeloid Leukemia in Adults: Recommendations From an International Expert Panel, on Behalf of the European LeukemiaNet. *Blood*, 115(3):453–74.

Döhner, H., Weisdorf, D. J., and Bloomfield, C. D. (2015). Acute myeloid leukemia. *N Engl J Med*, 373:1136–1152.

Eisenreich, S., Abou-El-Ardat, K., Szafranski, K., Campos Valenzuela, J. A., Rump, A., Nigro, J. M., et al. (2013). Novel CIC point mutations and an exon-spanning, homozygous deletion identified in oligodendroglial tumors by a comprehensive genomic approach including transcriptome sequencing. *PLoS ONE*, 8(9):e76623.

Exner, N. D., Valenzuela, J. A. C., Abou-El-Ardat, K., Miletic, H., Huszthy, P. C., Radehaus, P. M., et al. (2019). Deep sequencing of a recurrent oligodendroglioma and the derived xenografts reveals new insights into the evolution of human oligodendroglioma and candidate driver genes. *Oncotarget*, 10(38):3641–3653.

Fouad, Y. A. and Anaei, C. (2017). Revisiting the hallmarks of cancer. *Am J Cancer Res*, 7(5):1016–1036.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22.

Gladitz, J., Klink, B., and Seifert, M. (2018). Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. *Acta Neuropathol Commun.*, 6:49.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100:57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144:646–674.

Hastie, T. and Efron, B. (2013). lars: Least angle regression, lasso and forward stagewise, R package lars, https://cran.r-project.org/package=lars.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira1, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, 10(8):551–564.

Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Hum. Mol. Genet.*, 18:R1–R8.

Herold, T., Rothenberg-Thurley, M., Grunwald, V. V., Janke, H., Goerlich, D., Sauerland, M. C., et al. (2020). Validation and refinement of the revised 2017 European LeukemiaNet genetic risk stratification of acute myeloid leukemia. *Leukemia*.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods*, 10(11):1108–1115.

Hunter, S., Young, A., Olson, J., Brat, D. J., Bowers, G., Wilcox, J. N., et al. (2002). Differential expression between pilocytic and anaplastic astrocytomas: identification of apolipoprotein D as a marker for low-grade, non-infiltrating primary CNS neoplasms. *J Neuropathol Exp Neurol.*, 61:275–281.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res.*, 18(4):644–52.

Jansen, M., Yip, S., and Louis, D. N. (2010). Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. *The Lancet Neurology*, 9(7):717–726.

Jenkins, R. B., Blair, H., Ballman, K. V., Giannini, C., Arusell, R. M., and Law, M. (2006). A t(1;19)(q10;p10) mediates the combined deletions of 1p and 19q and predicts a better prognosis of patients with oligodendroglioma. *Cancer Res*, 66(20):9852–61.

Johansson, S., Astrom, L., F., S., Isacsson, U., Montelius, A., and Turesson, I. (2012). Hypofractionated proton boost combined with external beam radiotherapy for treatment of localized prostate cancer. *Prostate Cancer*, 2012:654861.

Jones, D. T. et al. (2013). Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.*, 45:927–932.

Jones, D. T., Gronych, J., Lichter, P., Witt, O., and Pfister, S. M. (2012). MAPK pathway activation in pilocytic astrocytoma. *Cell Mol Life Sci.*, 69:1799–1811.

Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T., et al. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.*, 7:486.

Kaltenbach, H. M., Dimopoulos, S., and Stelling, J. (2009). Systems analysis of cellular networks under uncertainty. *FEBS Lett.*, 583(24):3923–30.

Kamoun, A., Idbaih, A., Dehais, C., Elarouci, N., Carpentier, C., Letouzé, E., et al. (2016). Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. *Nature Communications*, 7:11263.

Klapproth, E., Dickreuter, E., Zakrzewski, F., Seifert, M., Petzold, A., Dahl, A., et al. (2018). Whole exome sequencing identifies mTOR and KEAP1 as potential targets for radiosensitization of HNSCC cells refractory to EGFR and β1 integrin inhibition. *Oncotarget*, 9:18099-18114.

Krogan, N. J., Lippman, S., Agard, D. A., and Ashworth, A.and Ideker, T. (2015). The cancer cell map initiative: defining the hallmark networks of cancer. *Mol Cell*, 58(4):690–8.

Labussiere, M., Idbaih, A., Wang, X.-W., Marie, Y., Boisselier, B., Falet, C., et al. (2010). All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. *Neurology*, 74(23):1886–1890.

Lauber, C., Correia, N., Trumpp, A., Rieger, M. A., Dolink, A., Bullinger, L., Roeder, I., and Seifert, M. (2020). Survival differences and associated molecular signatures of DNMT3A-mutant acute myeloid leukemia patients. *Scientific Reports*, 10:12761.

Lauber, C., Klink, B., and Seifert, M. (2018). Comparative analysis of histologically classified oligodendrogliomas reveals characteristic molecular differences between subgroups. *BMC Cancer*, 18:399.

Lawrence, M., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

Lazebnik, Y. (2010). What are the hallmarks of cancer? *Nature*, 10:232–233.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, 47(2):106–114.

Leote, A. C., Wu, X., and Beyer, A. (2019). Network-based imputation of dropouts in single-cell RNA sequencing data. *bioRxiv*, page doi: https://doi.org/10.1101/611517.

Letouzé, E., Martinelli, C., Loriot, C., Burnichon, N., Abermil, N., and Ottolenghi, C. (2013). SDH mutations establish a hypermethylator phenotype in paraganglioma. *Cancer Cell*, 23(6):739–52.

Li, H. an Xuan, J., Wang, Y., and Zhan, M. (2008). Inferring regulatory networks. *Front Biosci.*, 13:263–75.

Lim, M. and Hastie, T. (2015). Learning Interactions via Hierarchical Group-Lasso Regularization. *J Comput Graph Stat*, 24(3):627–654.

Lin, B., Lee, H., Yoon, J.-G., Madan, A., Wayner, E., Tonning, S., et al. (2015). Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene. *Oncotarget*, 6(7):5369–5381.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2013). covTest: Computes covariance test for adaptive linear modelling, R package covTest, https://cran.r-project.org/src/contrib/archive/covtest/.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Stat.*, 42(2):413–468.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, 114(2):97–109.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820.

Mäder, L., Blank, A. E., Capper, D., Jangsong, J., Baumgarten, P., Wirsik, N. M., et al. (2018). Pericytes/vessel-associated mural cells (VAMCs) are the major source of key epithelial-mesenchymal transition (EMT) factors SLUG and TWIST in human glioma. *Oncotarget*, 9:24041–24053.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods*, 9(8):796–804.

Marbach, D., Mattiussi, C., and Floreano, D. (2009). Combining multiple results of a reverse-engineering algorithm: application to the dream five-gene network challenge. *Ann N Y Acad Sci*, 1158:102–13.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, 107(14):6286–6291.

Mardis, E. (2014). The translation of cancer genomics: time for a revolution in clinical cancer care. *Genome Med.*, 6(3):22.

Masiero, M., Simoes, F. C., Han, H. D., Snell, C., Peterkin, T., Bridges, E., et al. (2013). A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis. *Cancer Cell*, 24(2):229–241.

Mateo, J., Boysen, G., Barbieri, C. E., Bryant, H. E., Castro, E., Nelson, P. S., et al. (2017). DNA Repair in Prostate Cancer: Biology and Clinical Implications. *Eur Urol*, 71:417–421.

McAllister, M. J., Underwood, M. A., Leung, H. Y., and Edwards, J. (2019). A review on the interactions between the tumor microenvironment and androgen receptor signaling in prostate cancer. *Transl Res.*, 206:91–106.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B*, 72(4):417–473.

Mitra, K., Carvunis, A. R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet.*, 14(10):719–32.

Newman, M. E. J. (2010). *Networks: An Introduction*. Number ISBN: 978-0199206650. Oxford University Press.

Noble, W., Kuang, R., Leslie, C., and Weston, J. (2005). Identifying remote protein homologs by network propagation. *FEBS J.*, 272(20):5119–28.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–522.

Ohgaki, H. and Kleihues, P. (2005). Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. *J Neuropathol Exp Neurol*, 64:479–489.

Ohgaki, H. and Kleihues, P. (2009). Genetic alterations and signaling pathways in the evolution of gliomas. *Cancer Sci.*, 100:2235–2245.

Ohgaki, H. and Kleihues, P. (2013). The definition of primary and secondary glioblastoma. *Clin Cancer Res.*, 19:764–772.

Pahlajani, N., Ruth, K. J., Buyyounouski, M. K., Chen, D. Y., Horwitz, E. M., Hanks, G. E., et al. (2012). Radiotherapy doses of 80 Gy and higher are associated with lower mortality in men with Gleason score 8 to 10 prostate cancer. *Int J Radiat Oncol Biol Phys*, 82(5):1949–56.

Peitzsch, C., Cojoc, M., Hein, L., Kurth, I., K., M., Trautmann, F., et al. (2016). An epigenetic reprogramming strategy to resensitize radioresistant prostate cancer cells. *Cancer Res*, 76:2637–51.

Ploen, G. G., Nederby, L., Guldberg, P., Hansen, M., Ebbesen, L. H., Jensen, U. B., et al. (2014). Persistence of DNMT3A mutations at long-term remission in adult patients with AML. *Br J Haematol.*, 167(4):478–486.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA*, 99:12963–12968.

Reiss, D. J., Baliga, N. S., and Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7:280.

Renneville, A., Boissel, N., Nibourel, O., Berthon, C., Helevaut, N., Gardin, C., et al. (2012). Prognostic significance of DNA methyltransferase 3a mutations in cytogenetically normal acute myeloid leukemia: a study by the Acute Leukemia French Association. *Leukemia*, 26(6):1247–1254.

Ribeiro, A. F. T., Pratcorona, M., Erpelinck-Verschueren, C., Rockova, V., Sanders, M., Abbas, S., et al. (2012). Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood*, 119(24):5824–5831.

Rickman, D. S., Bobek, M. P., Misek, D. E., Kuick, R., Blaivas, M., Kurnit, D. M., et al. (2001). Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res.*, 61:6885–6895.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47.

Rorive, S., Maris, C., Debeir, O., Sandras, F., Vidaud, M., Bièche, I., et al. (2006). Exploring the distinctive biological characteristics of pilocytic and low-grade diffuse astrocytomas using microarray gene expression profiles. *J Neuropathol Exp Neurol.*, 65:794–807.

Ruffalo, M., Koyutürk, M., and Sharan, R. (2015). Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer. *PLoS Comput Biol.*, 11(12):e1004595.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol.*, 18(12):1257–61.

Seifert, M. (2010). Extensions of Hidden Markov Models for the analysis of DNA microarray data. *PhD Thesis*, http://dx.doi.org/10.25673/325.

Seifert, M., Abou-El-Ardat, K., Friedrich, B., Klink, B., and Deutsch, A. (2014). Autoregressive Higher-Order Hidden Markov Models: Exploiting Local Chromosomal Dependencies in the Analysis of Tumor Expression Profiles. *PLoS One*, 9:e100295.

Seifert, M. and Beyer, A. (2018). regNet: An R package for network-based propagation of gene expression alterations. *Bioinformatics*, 34(2):308–11.

Seifert, M., Friedrich, B., and Beyer, A. (2016). Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. *Genome Biology*, 17:204.

Seifert, M., Garbe, M., Friedrich, B., Mittelbronn, M., and Klink, B. (2015). Comparative transcriptomics reveals similarities and differences between astrocytoma grades. *BMC Cancer*, 15:952.

Seifert, M., Peitzsch, C., Gorodetska, I., Börner, C., Klink, B., and Dubrovska, A. (2019). Network-based analysis of prostate cancer cell lines reveals novel marker gene candidates associated with radioresistance and patient relapse. *PLoS Comput Biol*, 15(11): e1007460.

Seifert, M., Schackert, G., Temme, A., Schröck, E., Deutsch, A., and Klink, B. (2020). Molecular Characterization of Astrocytoma Progression Towards Secondary Glioblastomas Utilizing Patient-Matched Tumor Pairs. *Cancers*, 12(6):1696.

Shah, M. Y. and Licht, J. D. (2011). DNMT3A mutations in acute myeloid leukemia. *Nat Genet*, 43(4):289–290.

Sharan, R., Ulitsky, I., and R., S. (2007). Network-based prediction of protein function. *Mol Syst Biol.*, 3:88.

Shnaps, O., Perry, E., Silverbush, D., and Sharan, R. (2016). Inference of personalized drug targets via network propagation. *Pac Symp Biocomput.*, 21:156–67.

Sun, Y., Shen, H., Xu, T., Yang, Z., Qiu, H., Sun, A., et al. (2016). Persistent DNMT3A mutation burden in DNMT3A mutated adult cytogenetically normal acute myeloid leukemia patients in long-term remission. *Leuk Res.*, 49:102–107.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45:D362–368.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244.

The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615.

The Cancer Genome Atlas Research Network (2012a). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525.

The Cancer Genome Atlas Research Network (2012b). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337.

The Cancer Genome Atlas Research Network (2012c). Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70.

The Cancer Genome Atlas Research Network (2013a). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, 368:2059–2074.

The Cancer Genome Atlas Research Network (2013b). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, 45:1113–1120.

The Cancer Genome Atlas Research Network (2014a). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513:202–209.

The Cancer Genome Atlas Research Network (2014b). Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159:676–690.

The Cancer Genome Atlas Research Network (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.*, 372(26):2481–2498.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B*, 58(1):267–288.

Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., et al. (2016). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(539):309–313.

Tomasetti, C. and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81.

Tonn, J.-C., Westphal, M., Rutka, J. T., and Grossman, S. A. (2005). Neuro-oncology of CNS tumors. *Springer*, ISBN: 978-3540258339.

Towner, R. A., Jensen, R. L., Colman, H., Vaillant, B., Smith, N., Casteel, R., et al. (2013). ELTD1, a potential new biomarker for gliomas. *Neurosurgery*, 72(1):77–91.

Turcan, S., Rohle, D., Goenka, A., Walsh, L. A., Fang, F., Yilmaz, E., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483(7390):479–483.

van den Bent, M. J. (2010). Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathologica*, 120(3):297–304.

van't Veer, L., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol.*, 6(1):e1000641.

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, 355(6332):eaai8478.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17:98 – 110.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558.

Wang, C., Funk, C. C., Eddy, J. A., and Price, N. D. (2013). Transcriptional analysis of aggressiveness and heterogeneity across grades of astrocytomas. *PLoS One*, 8:e76694.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–9.

Yang, L., Rau, R., and Goodell, M. A. (2015). DNMT3A in haematological malignancies. *Nat Rev Cancer*, 15(3):152–165.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45(10):1134–1140.

Zakrzewski, F., Gieldon, L., Rump, A., Seifert, M., Grützmann, K., Krüger, A., et al. (2019). Targeted capture-based NGS is superior to multiplex PCR-based NGS for hereditary BRCA1 and BRCA2 gene analysis in FFPE tumor samples. *BMC Cancer*, 19:396.

Zhang, W., Ma, J., and Ideker, T. (2018). Classifying tumors by supervised network propagation. *Bioinformatics*, 34(13):i484–i493.

Ziegler, J., Pody, R., Coutinho de Souza, P., Evans, B., Saunders, D., Smith, N., et al. (2017). ELTD1, an effective anti-angiogenic target for gliomas: preclinical assessment in mouse GL261 and human G55 xenograft glioma models. *Neuro-Oncology*, 19(2):175–185.

Zietman, A. L., Bae, K., Slater, J. D., Shipley, W. U., Efstathiou, J. A., Coen, J. J., et al. (2010). Randomized trial comparing conventional-dose with high-dose conformal radiation therapy in early-stage adenocarcinoma of the prostate: long-term results from proton radiation oncology group/american college of radiology 95-09. *J Clin Oncol*, 28(7):1106–11.

Danksagung

Meine Begeisterung für die Entwicklung und Anwendung von computerbasierten Verfahren für die innovative Analyse von molekularen Tumordaten mit Hilfe von Netzwerken wurde entscheidend durch meinen Wechsel nach Dresden geweckt. In der damaligen Arbeitsgruppe von Prof. Dr. Andreas Beyer am Biotechnologischen Zentrum (BIOTEC) hatte ich als junger Postdoc die Gelegenheit, mich tiefer in aktuelle Problemstellungen der Krebsforschung und Bioinformatik einzuarbeiten. Dabei konnte ich die Grundlagen für das Netzwerkinferenz- und das Netzwerkflussverfahren legen, welche eine bedeutende Rolle für die vorliegende Habilitationsschrift gespielt haben. Darüber hinaus danke ich Prof. Dr. Andreas Beyer für die exzellente wissenschaftliche Betreuung, den stetigen Austausch von Ideen, die umfassende Unterstützung und die Förderung über die eigentliche Projektlaufzeit hinaus.

Weiterhin bin ich Prof. Dr. Andreas Deutsch sehr dankbar, dass er mich, nach dem Wechsel von Prof. Dr. Andreas Beyer nach Köln, bei sich in seiner Arbeitsgruppe am Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) aufgenommen hat. Dies hat es mir erlaubt, das begonnene Forschungsprojekt, welches an den Standort Dresden geknüpft war, erfolgreich fortzuführen. Die am ZIH verfügbaren Compute-Server haben mir die Berechnung von genomweiten Genregulationsnetzwerken und Netzwerkflussanalysen ermöglicht. In dieser Zeit ist auch, durch den regen interdisziplinären Austausch, die enge Zusammenarbeit mit Dr. Barbara Klink entstanden.

Für die Chance als Gruppenleiter eine Bioinformatik Core Unit am Institut für Medizinische Informatik und Biometrie (IMB) aufzubauen und dabei meine Kenntnisse im Bereich der Krebsforschung weiterzuentwickeln und zielgerichtet in eine Vielzahl von Projekten einzubringen, bin ich Prof. Dr. Ingo Röder sehr dankbar. Ich danke Prof. Dr. Ingo Röder darüber hinaus für den kollegialen Umgang miteinander, die umfassende Unterstützung und Förderung meiner beruflichen Laufbahn, strategische Ratschläge und die Möglichkeit zum unabhängigen Arbeiten an verschiedenen Projekten. Dies hat letztendlich die vorliegende Habilitationsschrift erst möglich gemacht und meine persönliche Entwicklung in den letzten Jahren entscheidend mitgeprägt.

Für die umfassenden Einblicke in die Hirntumorforschung und die daraus resultierende enge Zusammenarbeit an medizinisch relevanten Fragen bin ich Dr. Bar-

bara Klink sehr dankbar. Unsere gemeinsame Arbeit an verschiedenen Projekten, die zum Großteil auch als Publikationen in diese Habilitationsschrift mit eingegangen sind, demonstrieren sehr deutlich, dass die interdisziplinäre Zusammenarbeit zwischen Bioinformatikern und Medizinern neue Erkenntnisse und Hypothesen für zukünftige Experimente erarbeiten kann. Der offene und freundschaftliche Umgang miteinander und die gegenseitige Wertschätzung bei der interdisziplinären Herangehensweise haben diese Erfolge erst möglich gemacht. Ich danke Dr. Barbara Klink darüber hinaus sehr, dass sie sich, gemeinsam mit Frau Prof. Dr. Evelin Schröck, sehr für meinen Wechsel ans IMB und meinen Verbleib in Dresden eingesetzt hat.

Für eine ebenso gute interdisziplinäre Zusammenarbeit im Bereich der Radioresistenz von Prostatakarzinomen möchte ich mich bei Prof. Dr. Anna Dubrovska und Dr. Claudia Peitzsch bedanken. Nur durch unsere gemeinsame interdisziplinäre Analyse der Zellliniendaten war es uns letztendlich möglich, neue Kandidatengene zu identifizieren, die mit der Ausprägung einer Radioresistenz in Verbindung stehen. Ganz besonders möchte ich mich für die durchgeführten Validierungsexperimente bedanken. Diese haben ausgezeichnet dazu beigetragen, das große Potenzial des von mir entwickelten netzwerkbasierten Verfahrens zur Bestimmung der Kandidatengene noch deutlicher hervorzuheben.

Ein ganz besonderer Dank geht an die Gruppenmitglieder meiner Bioinformatik Core Unit. Insbesondere die intensive Zusammenarbeit mit Josef Gladitz, um neue Krebsgenkandidaten für Oligodendrogliome zu ermitteln, war ein sehr erfolgreiches Promotionsprojekt, das auch in die vorliegende Habilitationsschrift mit eingeflossen ist. Weiterhin danke ich Dr. Chris Lauber für die Zusammenarbeit bei der Identifikation von Oligodendrogliomsubtypen und für die Zusammenarbeit bei der Bestimmung von molekularen Signaturen, die das Überleben von AML Patienten mit DNMT3A Mutationen besser vorhersagen können. Auch diese beiden Arbeiten haben einen wichtigen Anteil an der vorliegenden Habilitationsschrift.

Ich möchte mich auch bei allen bisher noch nicht genannten Co-Autoren der Publikationen, die in diese Habilitationsschrift eingeflossenen sind, für die gute Zusammenarbeit bedanken. Ebenso möchte ich mich bei allen Mitarbeiterinnen und Mitarbeitern des IMB bedanken. Die anregenden Fachdiskussionen, der kollegiale Umgang, die Administration der Compute-Server und die gute Arbeitsatmosphäre waren wichtige Bausteine auf dem Weg zur vorliegenden Arbeit.

Abschließend möchte ich mich noch ganz herzlich bei meiner Frau, Ulrike Seifert, und meiner Tochter, Theresa Seifert, für den großen Rückhalt bedanken, der mir den notwendigen Freiraum zur Vollendung dieser Arbeit ermöglicht hat.

Erklärung über die eigenständige Abfassung der Arbeit

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Habilitationsschrift stehen, und dass die vorgelegte Arbeit weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Habilitation oder eines anderen Prüfungsverfahrens vorgelegt wurde. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt an der Entstehung der vorliegenden Arbeit beteiligt waren.

Dresden, 11.05.2021
Dr. Michael Seifert