

Aus der
Klinik und Poliklinik für Strahlentherapie und Radioonkologie
der Ludwig-Maximilians-Universität München
Direktor: Professor Dr. med. Claus Belka

Integrative characterisation and prediction of the radiation response in radiation oncology.



Kumulative Habilitationsschrift
zur Erlangung der Venia legendi
für das Fach Strahlenbiologie
vorgelegt von
Dr. rer. nat. Kristian Unger

2018

Contents

1	Preface	3
2	Scientific Background	4
	2.1 Cellular radiation response in the context of radiation therapy	4
	2.2 Improvement of therapeutic success by the use of high-dimensional data analysis	5
	2.3 Statistical analysis of high-dimensional data sets	7
3	Own research	10
	3.1 A 5-MicroRNA signature predicts Survival and Disease Control of Patients with Head and Neck Cancer negative for HPV-infection	10
	3.2 Prediction of the therapy response in glioblastoma	16
	3.3 A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer.	20
	3.4 Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer	24
	3.5 Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data	30
	3.6 Copy number aberrations from Affymetrix SNP 6.0 genotyping data - how accurate are the commonly used prediction approaches?	34
4	Publication list	44
	4.1 Original works as first or last author	44
	4.2 Original works as co-author	45
	4.3 Review articles	50
5	Acknowledgements	52
6	Reprints of the underlying original works	53

1 Preface

Radiation therapy along with surgery and chemo- or target-oriented therapy (antibody and "small molecules") represent the four pillars of oncological treatment of cancer patients whilst approx. 50% of all cancer patients receive some sort of radiotherapeutic treatment. Many tumour entities, including brain and head and neck tumours require radiation as part of the existing standard-of-care therapy schemes. Thus, the efficiency of a radiation therapy treatment represents a strong determinant of the overall prognosis for these diseases. The efficiency of radiation therapy is determined on the one hand by the intrinsic radiosensitivity of the normal tissue surrounding the tumour which restrains the maximum tolerated and thus applicable total dose. For the other, resistance of tumour cells upon irradiation in the course of radiation therapeutic treatment can limit its overall effect. Hence, one of the uttermost important objectives in molecular radiooncological research is to understand the mechanisms of the radiation response in normal and tumour tissue while the knowledge of which would provide the foundation for the identification of molecular target structures to be tackled by therapeutic agents that would allow the modulation of the radiation response in such a way radiation therapy becomes most efficient. This as a matter, of course only would work if the modulating agent specifically increases the radiation sensitivity of the tumour tissue and without also increasing that of the surrounding normal tissue.

The last decades of cancer research have undoubtedly shown that tumour biology is very much characterised by inter- and intraindividual heterogeneity. Modern therapeutic approaches have to account for this by providing individualised treatments. This requires the ability to stratify into groups that are likely to respond to individualised treatment approaches in the first place and to *a priori* identify patients being parts of these strata. This involves identification of prognostic markers in clinical, radiomics and molecular data of patients as part of retrospectively or prospectively collected cohorts. The latter can be efficiently done in high-dimensional molecular omics "big-data" sets that characterise the genome, transcriptome, proteome, metabolome etc. that are generated on clinical samples from patients of such cohorts in combination with clinical follow-up data followed by predictive modelling.

Through the mechanistic characterisation of the molecular radiation response and the discovery of prognostic markers predicting the therapeutic success of radiation treated cancer patients, important steps have been made forward. So, for head and neck cancer and glioblastoma prognostic markers were identified and independently validated. With regard to the understanding of the molecular radiation response both for normal and tumour tissue important insights have been generated. Combining these two sources of knowledge is likely to result in new personalised therapy concepts for radiation treated cancer. Another non-intended side-effect of radiotherapy is the induction of secondary malignancies in non-tumour tissue that receives irradiation dose as a side-effect. While epidemiologically proven, the underlying mechanisms and markers of radiation-induced secondary malignancies are still under investigation. We have demonstrated for radiation-induced thyroid and breast cancer that such mechanisms and markers exist.

The present cumulative habilitation thesis puts up clinical-translational scientific projects making extensive use of computational approaches - two of which are part of the present work - that were published in peer-reviewed papers that worked towards the understanding of the mechanisms of the radiation response and identified prognostic markers in radiation-therapy treated cancer patient cohorts and cohorts of patients with radiation-induced breast cancer.

2 Scientific Background

2.1 Cellular radiation response in the context of radiation therapy

The aim of primary definitive or adjuvant radio (chemo)therapy is to kill all dormant or dividing tumour cells by introducing DNA damage. Single-strand breaks (SSB), double-strand breaks (DSB) or base changes in the DNA of the cells are caused, whereby the biological effect of the radiation is mainly caused by unrepaired or incorrectly repaired DSBs. There are various mechanisms of DNA repair in affected cells. Besides single strand break repair (SSBR), base excision repair (BER) and homologous recombination (HR), non-homologous end-joining (NHEJ) plays the leading role [1]. Defective repair of DNA damage can lead to the loss of the unlimited ability for tumour cell division, whereby different forms of cell death (apoptosis, necrosis, autophagy, mitotic cell death) or senescence can be induced. Cells with non-repaired or mismatched DNA damage often go through several cell divisions before mitotic cell death (mitotic catastrophe) is activated. Serine protein kinase ATM (ataxia teleangiectasia mutated) is an important sensor of DNA damage and triggers various signalling cascades involved in the regulation of cell cycle control, cell cycle arrest, DNA repair and cell death [2]. In sufficiently irradiated tumour cells, cell death is directly activated because the large number of DNA damages exceeds the cell's repair capacity. The capacity for DNA repair of each cell type determines the specific intrinsic sensitivity of each cell type. Tumour cells bear large numbers of genetically enhanced mutations, such as inactivating mutations of the P53 gene, in the course of tumour evolution. Genetic alterations often affect signal molecules involved in DNA repair, as a result of which tumour cells usually have reduced repair capacity and are therefore more severely damaged than normal tissue cells as a result of being exposed to radiation. In addition, the normal tissue is exposed to a lower dose than the tumour tissue in the course of a therapeutic irradiation. This results in a survival advantage of normal cells with prior elimination of the tumor cells [3].

2.1.1 Approaches for the modulation of the radiation response

In order to modulate the sensitivity of normal and tumour tissue to radiation, targeted therapy can be applied at different levels, which essentially include DNA repair, cell cycle control, cell division and signalling pathways influencing cell death (e. g. NF- κ B or PI3K signalling pathways) and furthermore the microenvironment and a reduction of normal tissue toxicity [2]. In order to increase the radiation sensitivity of tumour cells without influencing the radiation sensitivity of normal tissue one strategy is the intensification of the oxygen effect by increasing the oxygen partial pressure in the breathing air in the context of a so-called hyperbaric oxygenation to increase the oxygen concentration in the tissue [4, 5]. The oxygen effect describes the amplification of the effect of ionizing radiation by the radicals and peroxides formed during radiolysis in the presence of oxygen, which attack the DNA in a chain reaction and fixate the DNA damage. The effect occurs mainly with the application of low-LET (linear energy transfer) radiation and is more pronounced the higher the oxygen concentration in the tissue. Hypoxic tumour areas, on the other hand, are characterised by increased resistance to radiation, which is why their elimination is, amongst other factors, decisive for the success of radiation therapy [6]. In addition, some chemotherapeutic agents and targeted substances for modulating radiation sensitivity and simultaneously damaging tumour cells have been successfully used in clinical applications such as e. g.B. Mitomycin-C, taxanes, antifolates, cisplatin, 5-FU, hydroxycarbamide and the EGFR inhibitor cetuximab in the radiochemotherapeutic treatment of head and neck tumours [7, 8, 9, 10]. Another chemotherapeutic agent that is used

in standard therapy of glioblastomas which is also radiosensitising is temozolomide [11, 12, 13]. Although the specific synergistic effects of these substances cannot be determined in detail by means of radiochemotherapy, cooperative effects that act simultaneously and at the same location on a tissue are differentiated from those that occur locally and possibly also at different times. In addition to the pharmacological mechanism of action, this distribution also depends to a large extent on the design of the therapy scheme [7, 10].

2.1.2 Therapy resistance

In the course of radiotherapeutic treatment of tumours, the occurrence of local and locoregional recurrences can often be observed, which is a strong limiting factor for the success of the therapy. Three factors can be blamed for the occurrence of local recurrences: the radiation resistance of hypoxic tumour cells, the ability of tumour cells to repopulate and intrinsic or acquired resistance towards radiation [2]. Since tumour cells exhibit a high degree of genomic instability, resistance to radiation can also be acquired in the course of tumour evolution and the enrichment of genetic alterations, e.g. mutation of P53, an important cell cycle control gene [7]. Accordingly, it is an important research objective to identify the markers and mechanisms underlying intrinsic and acquired radiation resistance in order to be able to use them for prognosis and targeted therapy.

2.1.3 Radiation-induced secondary malignancies

One side effect of radiotherapeutic treatment, particularly in the case of fractionated percutaneous photon-based radiotherapy is significant deposition of dose in normal tissue surrounding the tumour to be treated [14]. This problem is even more prominent the younger the patients are at the time of the radiotherapeutic treatment [15, 16]. While from the epidemiologic point of view there is no doubt about the inducibility of secondary cancers after radiation exposure the molecular mechanisms leading to radiation-induced carcinogenesis are not yet conclusively investigated [15]. However, in own research projects we could identify markers of radiation-induced thyroid and breast cancer. For thyroid cancer we identified a region on chromosomal band 7q11 which contains the gene CLIP2. 7q11 was exclusively gained in thyroid cancers of patients who were exposed to radiation from the Chernobyl accident fallout compared to a non-exposed control group. Further, expression of the CLIP2 protein was significantly increased in the exposed group compared with the unexposed group [17, 18, 19]. For breast cancer we recently demonstrated in a group of female patients who worked as clean-up workers at the Chernobyl power plant facility and were exposed to radiation that their breast cancer tissues showed attenuated expression of the transcription factor TRPS1 compared to a non-exposed control group. We further found miRNA hsa-miR-26b-5p as a likely regulator of TRPS1 as its expression was increased in the exposed compared to the non-exposed group [20]. In the same group of patients we identified a genomic copy number signature that allows prediction of the exposure status in breast cancer [21]. The breast cancer-related works are part of the present thesis.

2.2 Improvement of therapeutic success by the use of high-dimensional data analysis

Prior therapy of a tumour disease is the initial diagnosis, which must be followed by further diagnostic surveys through imaging (radiology), histology (pathology) and molecular markers (molecular pathology). Based on the overall diagnosis, a decision is made as to which therapy route is

chosen. The combination of surgical, radiotherapeutic, chemotherapeutic and new targeted treatment influences the clinical course of the disease and thus the therapeutic success. The process of individual diagnosis for the overall diagnosis, which is decisive for the therapy decision, is the key to an optimal therapy, in which the patients are assigned to a corresponding therapy scheme based on the diagnosis groups. The likelihood that in the course of standard treatment the chosen therapy route will lead to success is high for the majority of the assigned group. However, the clinical course at individual level is highly heterogeneous due to individual differences which cannot be resolved by radiological and pathological diagnostics. The aim of the analysis of clinically associated high-dimensional data sets is therefore to identify mechanisms of radiation sensitivity, molecular target structures for personalized therapy approaches and prognostic markers for therapy response.

High-dimensional data sets and analyses can be used to calculate the so-called molecular network of radiation sensitivity. These networks can, for example, be calculated from microarray or next generation sequencing derived gene expression data sets of tissues with different radiation sensitivity and represent the interaction of genes associated with altered radiation sensitivity. By assigning genes to signalling pathways, molecular mechanisms of radiation sensitivity can be derived in the form of molecular networks. This in turn allows the most important subnetworks to be identified, whose manipulation has the potential to modulate the radiation sensitivity [22]. Such analyses would, in the most optimal case, result in the identification of a target structure, which would allow specific sensitisation of the tumour tissue but not the normal tissue [23].

One promising route towards achievement of the above-mentioned goals is the analysis of comprehensive data sets of different molecular levels from cells of normal and tumour tissue. These data sets are generated using microarray technology, mass spectrometry and next generation sequencing. The molecular levels genome, transcriptome and proteome, the interaction of which is described by the "Central Dogma of Molecular Biology", can be recorded and the molecular levels of epigenetics, which have a regulatory effect on genome, transcriptome and proteome, should be considered even importantly [24]. Such "big data", integrative data sets are characterised by the fact that several molecular levels are simultaneously recorded in the measurement of clinical tissue samples or cells of model systems and the interaction of these is taken into account in the analysis. So, the integration of miRNA and mRNA levels can be used to identify genes that are regulated by the miRNAs involved and thus allow drawing conclusions about the functional role of these miRNAs.

Biological systems can be studied on the molecular, cellular, organ, individual or population level. Understanding, describing, quantifying and analysing these levels as systems or parts of systems is the core of systems biology research approaches. The aim is to identify the elements that make up a biological system as globally as possible, systematically searching for connections between these elements and characterising the type of interaction. Like a circuit diagram of a technical device, an attempt is being made to represent the biological system as a regular and reproducible network that can precisely described. This requires that the elements of the molecular levels of cells representing normal tissue or tumour tissue are captured as completely as possible [25, 26]. The great advantage over traditional, purely association-based, descriptive biological research lies in the potential predictive power of systematically described systems. If the system is known, then it is also possible to make a statement about how the system will change in the event of a change (perturbation) of one or more components. These predictions are made with the help

of mathematical models adapted to the high-dimensional systems biology data.

2.3 Statistical analysis of high-dimensional data sets

2.3.1 Association analysis of integrative high-dimensional data sets

In association analysis, which is hypothesis-driven, patients are divided into groups that differ with respect to one parameter e.g. response to therapy. The measurement data of each data point is then assigned to the groups to be compared and subjected to a statistical test, which allows the null hypothesis to be checked and which is adequate for the test situation. Decisive for the selection of the test is, among other things, the distribution of the measured values, which must correspond to a certain distribution (e. g. normal distribution) for parametric tests, whereas this prerequisite does not have to be fulfilled for non-parametric tests. In addition, the data type (continuous or categorical data) plays an important role in the selection of a test. The final result of an association test is the probability for the rejection of the null hypothesis, which is represented by the so-called P-value. Many null hypotheses are tested in the association analysis of high-dimensional data sets due to the large number of data points. According to the measured data points, the number ranges from a few hundred (microRNA) to several million (Next Generation Sequencing)[27, 28].

Multiple Testing Problem

With normally distributed data and a significance level of 5%, a false positive rate (error I. type) of 5% is to be expected by definition, i.e. 5 false positive test results on 100 tests, 50 on 1,000 and 50,000 on 1 million. This problem is described as a so-called multiple test problem, which is addressed to reduce the number of false positives by correcting the P-values resulting from the statistical tests. The two best-known strategies have been developed and published by Carlo Emilio Bonferroni and the mathematicians Yoav Benjamini and Yosi Hochberg, whereas the false discovery rate (FDR) developed by Benjamini and Hochberg has established itself as a standard method in biostatistics because it is less conservative, i.e. tends to reject less null hypotheses [29, 30].

Static and dynamic data

Cells and tissues represent biological systems and are therefore dynamic. Thus, high-dimensional molecular measurements from tissues correspond to snapshots of dynamic processes and result in static data sets. Most of the clinical high-dimensional datasets contain one of these snapshots for each patient, whereby it must be left to chance as to the exact time at which this snapshot corresponds. However, since the molecular concentrations determined using high-throughput measurement technology correspond to the average over an entire tissue segment, it is assumed that found group differences represent general and thus also robust markers for a certain phenotype. Numerous studies have identified prognostic and predictive markers that are used in the diagnosis and therapy of tumour diseases. These include in the case of glioblastoma methylation of the promoter of the MGMT gene, in breast cancer a chromosomal rearrangement leading to the fusion protein HER2/neu or in squamous cell carcinoma in the head and neck area mutation of the P53 gene and especially HPV virus infection and activity [31, 32, 33, 34]. In contrast to static data sets, dynamic data sets need to be generated from living cells by performing several measurements at different times. Dynamic data sets are often generated in the frame of perturbation experiments in order to determine the effect of a stimulus (e. g. irradiation) on the rapidly reacting molecular levels, such as transcriptome, miRNA level or proteome. The answer can be quantified and described

over a whole period of time or, in the simplest case, a distinction can be made between answering sooner and later after perturbation [35]. Another application of dynamic data is the computation of gene regulatory association networks, which can be performed much more accurately from time-dependent transcriptome data than from static transcriptome data sets [36].

2.3.2 Interpretation of primary analysis results

In association analyses of high-dimensional data sets, especially gene expression data sets (mRNA array or RNAseq), a large number of genes correlated with the investigated groups are often found. In order to facilitate the interpretation of these often complex results, downstream analyses are carried out. A frequently used approach here is enrichment analysis, in which genes are grouped together into gene sets, resembling biological processes known from the literature. Specific accumulation of genes in these gene sets is tested and it is assumed that, in the event of a positive result, the underlying biological process is involved in the formation of the associated phenotype. In the context of a modified radiation response of tumour or normal cells, changes in genes that occur frequently in signalling pathways associated with DNA repair, stress response, senescence or apoptosis would be expected.

2.3.3 Prediction analysis in high-dimensional data setting

One of the central analyses of high-dimensional data is the development of prognostic signatures using prediction approaches. The aim of such analyses is to develop a signature consisting of several features (i. e. genes, miRNAs, proteins) from a high-dimensional data set, whose expression allows the calculation of a so-called "risk score", which in turn reflects the probability of a certain outcome such as therapeutic success or survival. The underlying signature is calculated from a high-dimensional data set in combination with clinical follow-up and endpoints (e. g. overall survival, tumour-specific survival, relapse-free survival). A particular challenge in the development of prognostic signatures is to keep the complexity (i. e. the number of elements used) of the signature low in a sense that the probability of an overfitting of the underlying model is kept low as well. Overfitting usually leads to the fact that the calculated signature cannot be validated in a different data set than the one in which it was developed ("validation surprise"). Cross-validation and the use of the Akaike Information Criterion (AIC) are frequently used methods for the reduction or avoidance of model overfitting [37, 38]. A key element of each signature determination is its validation in an independent data set. For this purpose, a training and a validation data set must be defined before the calculations are started. Validity can also be increased by testing the model against other independent data sets. The basis for the development of prognostic signatures is the Cox regression model, which estimates the influence of variables on the duration of an event (e. g. death or occurrence of a relapse). Possible variables included in the model represent the measuring points of a high-dimensional data set. In order to select from these variables those which, as part of a signature, reliably predict the occurrence of the considered endpoint, either linear regression methods in combination with regularization or a so-called step-by-step regression (e. g. forward selection, forward-selection backward-elimination, bidirectional elimination) are generally used [39, 40, 41, 42].

2.3.4 Reconstruction of gene regulatory networks

The reconstruction of gene regulatory association networks is helpful for the investigation of molecular mechanisms of radiation response. In the case of *de novo* reconstructed gene regulatory networks, it is advantageous to gain knowledge about the interaction of genes in the context of the radiation response without prior knowledge. This analysis is based on the assumption that direct or indirect interaction of genes is expressed in a high correlation of the expressions of the relevant interaction partners. The central elements of gene regulatory association networks are the transcription factors, which influence entire groups of genes in their expression behaviour. However, since correlations reflect causalities only to a limited extent and a pure correlation-based method leads to numerous false-positive interactions, the principle of partial correlation can be used. The final result of the analysis is a network with all calculated direct and indirect interaction partners [36]. The advantage of representation in the form of a network is that it can be analyzed with the help of network graph analysis. This includes, for example, the analysis using centrality measures, which allow the identification of the most important and central genes. Such genes could, in connection with the research of the molecular mechanisms of radiation response, be candidates for target structures that are suitable for the modulation of radiation sensitivity. In addition, so-called network modules can be used to identify which groups of genes are the most likely to interact. The detailed analysis of network modules can allow conclusions to be drawn on previously unknown mechanisms of action of genes [43]. In order to be able to use transcriptome data sets for the purpose of reconstructing gene regulatory association networks, they must meet certain requirements, so the number of replicates used should be as large as possible and the sample should be as uniform as possible [44].

3 Own research

The research projects being part of this thesis are from three different areas: Identification and characterisation of prognostic markers in head and neck squamous cell carcinoma (HNSCC) and glioblastoma, identification of markers predicting radiation exposure in radiation-induced breast cancer and bioinformatic/biostatistic tools for the time-resolved analysis of transcriptome data and the determination of genomic copy number data from widely used Affymetrix SNP 6.0 data. In the following all papers included are shortly described, summarised and potential implications for the field are discussed.

3.1 A 5-MicroRNA signature predicts Survival and Disease Control of Patients with Head and Neck Cancer negative for HPV-infection

Hess J*, Unger K*, Maihoefer C, Schüttrumpf L, Schneider L, Heider T, Weber P, Marschner S, Braselmann K, Kuger S, Pflugradt U, Baumeister P, Walch A, Woischke C, Kirchner T, Werner M, Werner K, Baumann M, Budach V, Combs SE, Debus J, Grosu AL, Krause M, Rödel C, Stuschke M, Zips D, Zitzelsberger H, Ganswindt U, Henke M, Belka C A 5-MicroRNA signature predicts Survival and Disease Control of Patients with Head and Neck Cancer negative for HPV-infection. *Clin Canc Res.* 2018 Aug 31. (IF: 10.2, *co-first authorship)

3.1.1 Background

The 5-year survival rate of patients with locally advanced head and neck squamous carcinoma (HNSCC) is approx. 50%. Standard-of-care comprises surgery in combination with radio(chemo)therapy (adjuvant treatment) or definitive radio(chemo)therapy treatment. The success of therapy is expressed by local and distant control of the tumour while in cases where radiotherapy is involved, tumour control is reduced by resistance to radiation. A lot of discussion about the mechanism of the development of recurrences is ongoing and alternative and complementary treatment concepts including immunotherapy have been proposed [45]. However, this requires knowledge of the strata that are likely to profit or not to profit from new therapy concepts. This, in turn, requires prognostic biomarkers that allow, either alone or, more likely, in combination with the established molecular and clinical prognostic markers, to classify patients in order to assort them into specific treatment strata. As for most cancer entities established molecular clinical parameters that would allow specific stratification are also missing for HNSCC. The only established molecular factor for HNSCC is HPV16 infection. Although, its predictive significance remains subject of discussion, HPV-associated HNSCC now is also considered being a distinct cancer entity [46, 47, 48]. Moreover, a number of other candidate prognostic markers have been proposed for HNSCC, however there is none that made it into clinical practice so far [49].

3.1.2 Summary

We set out to search for a prognostic signature predicting the risk for local and distant recurrence in the course of HNSCC therapy. In a discovery cohort of 85 HPV-negative HNSCC patients established by the German Consortium for Translational Cancer Research Radiation Oncology Group

(DKTK-ROG) we identified a 5-miRNA signature predicting recurrence of the disease and survival endpoints. We were able to validate our results in an independent cohort of HPV-negative patients who were recruited at the Department of Radiation Oncology at the LMU University Clinics. Using recursive partitioning analysis (RPA) we were further able to build a decision tree integrating the prognostic miRNA signature with established clinical parameters and identified four prognostically distinct groups as a potential aid in therapy guidance. In addition a transcriptome-supported miRNA-mRNA target prediction network provides putative insights into the molecular networks the miRNA signature is embedded in.

3.1.3 Methods, Results and Discussion

3.1.3.1 Patients The two cohorts exclusively contained HPV-negative HNSCC patients who had undergone surgical resection of the tumour followed by radio(chemo)therapy. All patients were diagnosed with histologically proven HNSCC of the hypopharynx, oropharynx, or the oral cavity. The DKTK-ROG discovery cohort contained patients who were treated from 2005-2011 at one of the eight DKTK partner sites [50]. The monocentric validation cohort comprised 85 patients treated between 2008 and 2013 at the LMU Department of Radiation Oncology.

3.1.3.2 miRNA microarray expression profiling Total RNA with preserved small RNA fraction was used for global miRNA expression profiling using Agilent microarrays. After quality assessment and filtering the data set comprised the profiles of 1.046 miRNAs across 161 patients.

3.1.3.3 Prognostic model A robust likelihood-based survival modelling forward-selection approach was used for the selection of features i.e. miRNAs in the discovery set to be included in a prognostic cox-proportion hazard model predicting freedom from recurrence [51]. The maximum number of miRNAs allowed in the model was 20 and iterative random assignment of 4 folds for cross-validation was 10 times repeated. The best model was chosen according to the Akaike Information Criterion (AIC) and contained the following five miRNAs: hsa-let-7g-3p, hsa-miR-6508-5p, hsa-miR-210-5p, hsa-miR-4306, and hsa-miR-7161-3p. Risk scores were calculated for each patient after linear combination of signature miRNA expressions and prediction model cox-proportional hazard coefficients. The median risk score of the discovery set was used as a threshold for the definition of high- and low-risk patients in the validation set. Fig 1 shows the performance of the model in the discovery and validation sets (although not meaningful the log-rank test p-value for the discovery set was included for demonstration purposes). In addition to the endpoint freedom from recurrence which was used to build and train the model the signature also predicted other endpoints such as recurrence-free survival, overall survival and disease-specific survival.

In order to assess performance of the model with regard to sensitivity and specificity of the risk factor, receiver-operator characteristics (ROC) analysis was conducted. The analysis was performed in comparison to other clinical parameters that also showed association with freedom from recurrence in univariate cox-proportional hazard analysis. After five years follow-up (freedom from recurrence) the risk factor showed a better AUC compared with T-stage, lymphovascular invasion (LVI) and extracapsular extension (ECE, Figure 2).

In order to test for independence of the miRNA risk factor from the other clinical parameters, multivariate cox analysis was performed in which T-stage, LVI and ECE were included as covariates. Overall the model significantly predicted freedom from recurrence for DKTK-ROG with a hazard-ratio of 5.55 (95% CI 2.09-14.79, P=0.0006) and for LMU-KKG with a hazard-ratio of 3.94 (95% CI

1.23-12.59, P=0.02082).

These data and results demonstrate the independent prognostic relevance of the 5-miRNA signature and a superior prognostic value compared with established clinical parameters. However, prognostic relevance alone does not mean clinical relevance which requires that the prognostic marker aids in therapeutic decision making in concert with established clinical parameters. An approach to define prognostic groups considered for different treatment is the generation of decision trees using recursive partitioning analysis (RPA). We applied RPA in a pooled analysis on a dataset combining DTKK-ROG and LMU-KKG on the parameters 5-miRNA signature risk factor, T-stage, N-stage and ECE. The analysis revealed four distinct groups that significantly differed in prognosis. The groups identified by RPA could build the basis for the conception of personalised treatment approaches while in a first step the group with the best prognosis could be considered for therapeutic deintensification- and that with the worst prognosis for intensification strategies.

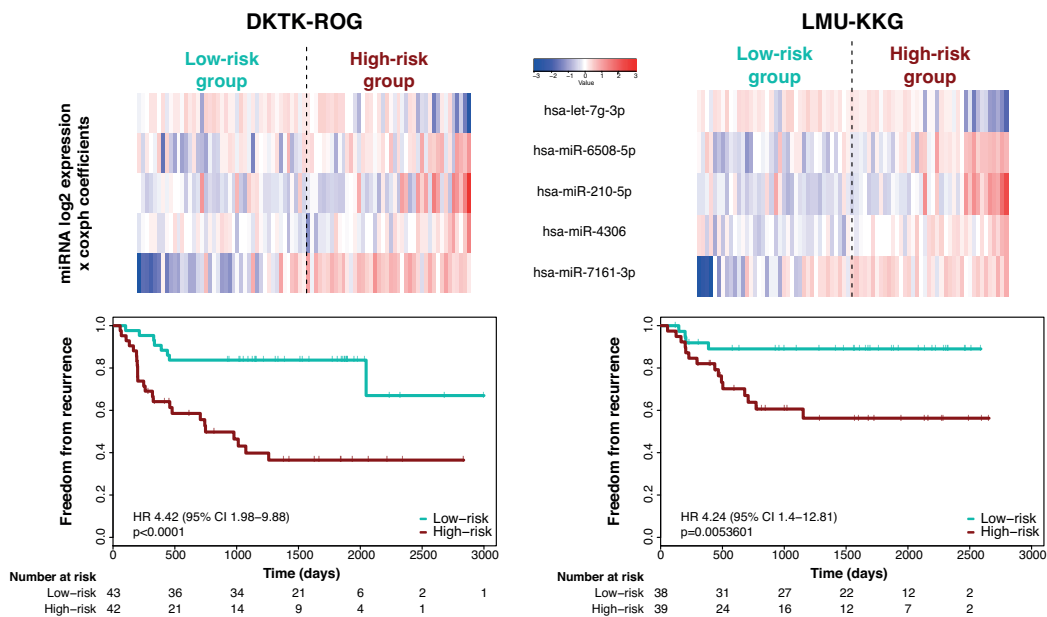


Figure 1: Freedom from recurrence stratified by risk according to the five-miRNA-signature: miRNA expression and Kaplan-Meier curves in the training and validation set. Heat map colors indicate scaled miRNA log₂ expression values multiplied by the Cox proportional hazard coefficients (coxph) from low (blue) to high (red) on a scale from -3 to 3 for each of the five signature miRNAs in the training (left panel) and validation set (right panel). Kaplan-Meier curves for the endpoint freedom from recurrence for HNSCC patients of the training (DTKK-ROG sample; left panel) and validation set (LMU-KKG sample; right panel) stratified into low- and high-risk patients according to the five-miRNA-signature. P-values are derived by log-rank test.

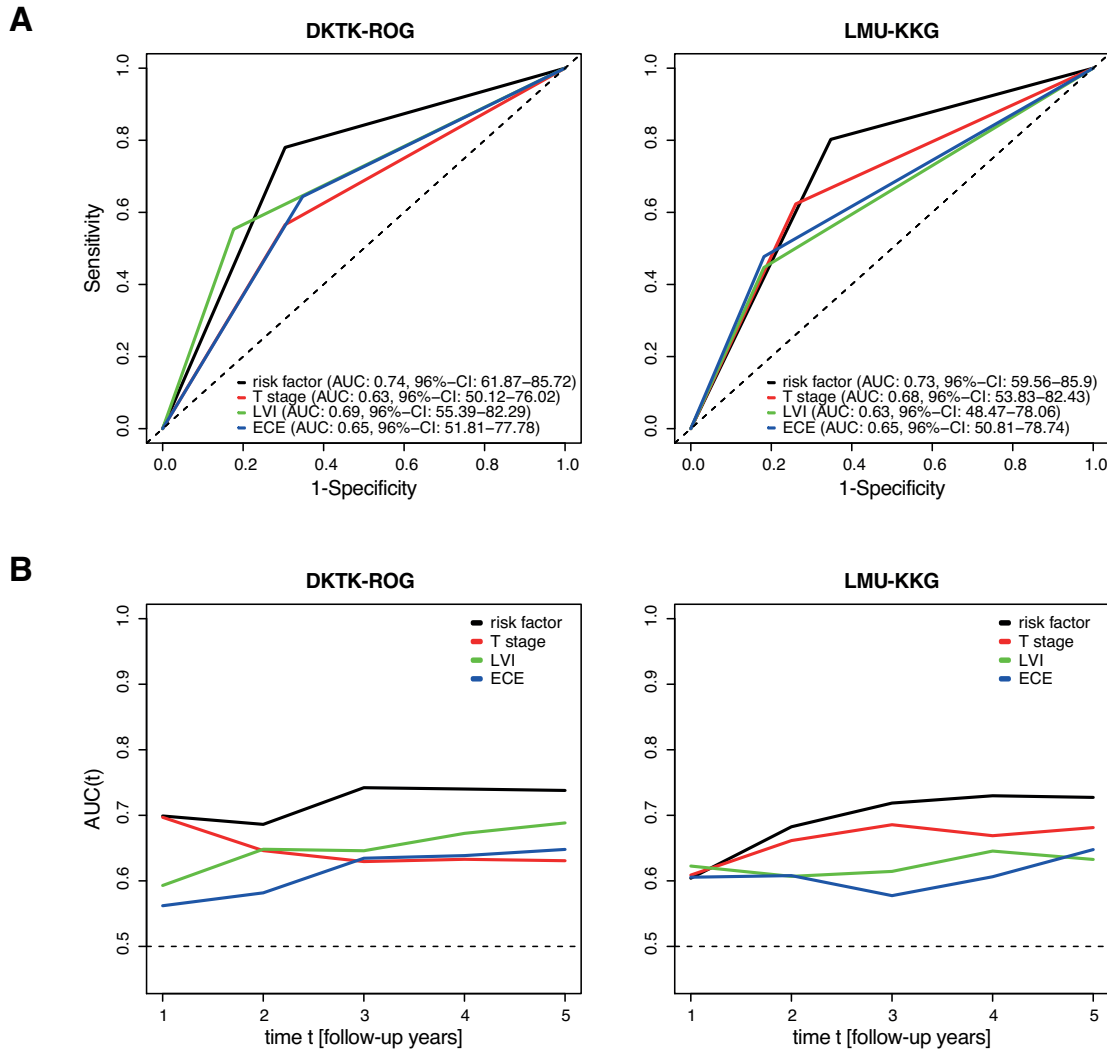


Figure 2: Time-dependent ROC curve analysis for the prediction of freedom from recurrence. (A) Time-dependent receiver operating characteristics (ROC) curves for the prediction of freedom from recurrence in the training (left panel) and validation set (right panel) at five follow-up years. The area under the curve (AUC) and the 95% CI of the five-miRNA-signature derived risk factor, TNM T stage, lymphovascular invasion (LVI), and extracapsular extension (ECE) are shown. (B) ROC curve analysis was performed for freedom from recurrence in the training (left panel) and validation set (right panel) at follow-up years 1-5. The area under the curve (AUC) of the five-miRNA-signature derived risk factor, TNM T stage, lymphovascular invasion (LVI), and extracapsular extension (ECE) are shown over time (at years 1-5).

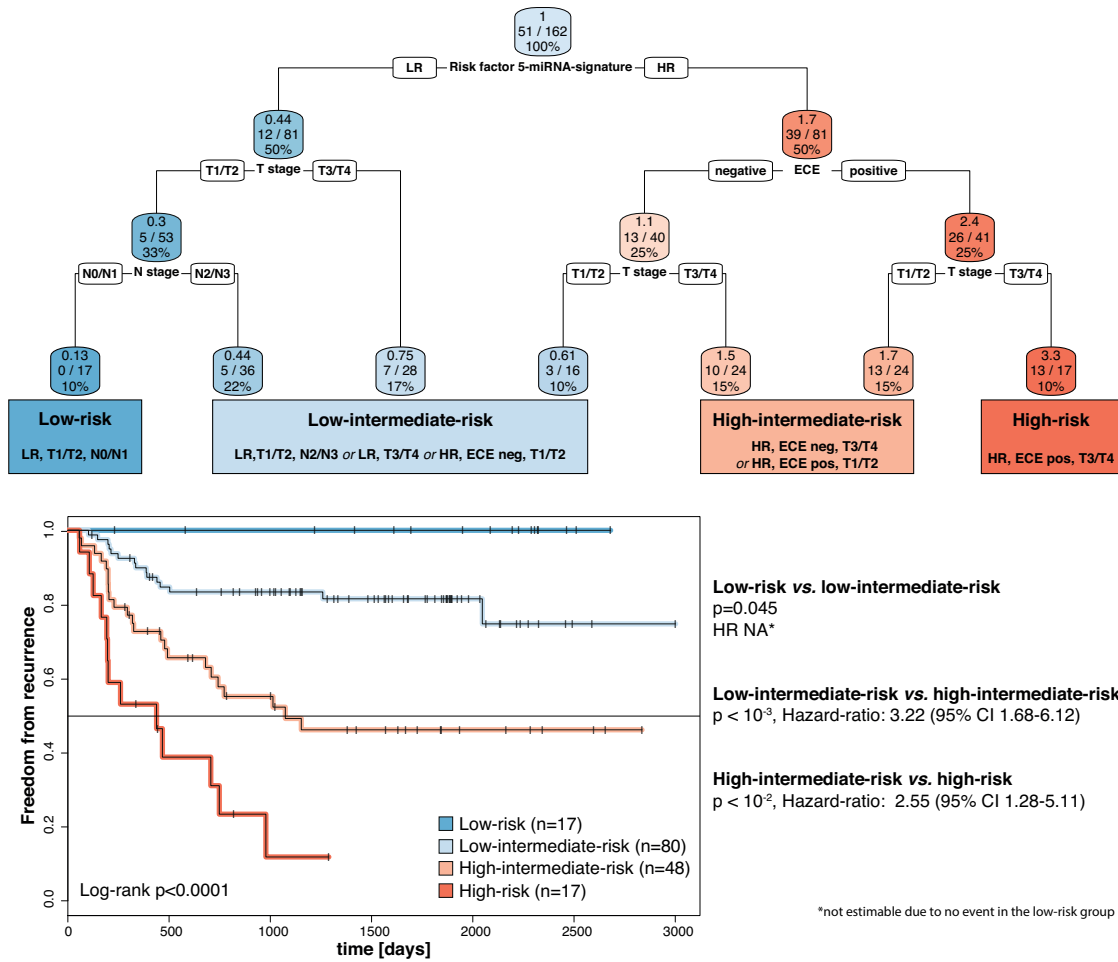


Figure 3: **Risk groups for recurrence identified by recursive partitioning analysis (RPA).** RPA tree and risk groups for recurrence combining the parameters five-miRNA-signature (high-risk, low-risk), ECE (negative, positive), T stage (T1/T2, T3/T4), and N stage (N0/N1, N2/N3) in the pooled HNSCC data set (n=162). Each node shows the predicted probability of recurrence (locoregional or distant failure; color code low to high: blue-red), the number of events for the total number of patients, and the percentage of observations in the node. Kaplan-Meier curves for the endpoint freedom from recurrence for the four identified risk groups “low-risk”, “low-intermediate risk”, “high-intermediate risk”, and “high-risk”. Multivariate and pairwise comparisons are shown. P-values are derived by log-rank test.

3.1.4 Conclusion

In the study we identified a prognostic 5-miRNA signature in a multicentric cohort followed by validation in a monocentric cohort. The signature was independent of established clinical prognosticators. Moreover, integration with established clinical prognostic parameters resulted in four prognostically distinct groups that could be considered for personalised therapeutic concepts.

3.1.5 Outlook

In a next step thorough exploration of the mechanistic molecular impact of the signature will be carried out in order to investigate the possibilities of specifically molecularly targeting the revealed prognostic groups. This involves *in vitro* and *in vivo* characterisation of cells with regulated expression of the signature miRNAs on the one hand. On the other hand transcriptome analysis of tumour specimens followed by correlation with the 5-miRNA signature risk score will be carried out. Another ongoing project investigates the intra-tumour heterogeneity of tissue sections of patients from the LMU-KKG cohort for whom the 5-miRNA signature risk score is known.

3.2 Prediction of the therapy response in glioblastoma

Niyazi M, Pitea A, Mittelbronn M, Steinbach J, Sticht C, Zehentmayr F, Piehlmaier D, Zitzelsberger H, Ganswindt U, Rödel C, Lauber K, Belka C, Unger K A 4-miRNA signature predicts the therapeutic outcome of glioblastoma. *Oncotarget*. 2016 Jul 19;7(29):45764-45775. (IF: 5.2)

3.2.1 Background

Glioblastoma (GBM) represent the most aggressive form of gliomas and in the standard setting including patients younger than 70 years with a Karnofsky performance score (KPS) greater 60 are treated by surgical resection and adjuvant radiochemotherapy followed by maintenance chemotherapy with temozolomide [12, 11]. However, high rate of recurrence is mostly responsible for the 5-year overall survival rate of 10% [52, 53]. Beside the established clinical prognostic markers including age, sex and KPS the only relevant molecular prognosticator is methylation of the promoter region of the MGMT gene. MGMT promoter methylation as such, however, lacks sufficient prognostic power for changing therapeutic decisions.

In an explorative study we set out to identify a molecular signature that enables stratification of GBM patients into prognostically significant groups for tailored GBM treatment approaches. We decided for miRNA as the molecular level of interest as it is known that miRNAs with a high degree of promiscuity target and regulate several mRNA species encoding for proteins involved in various signaling pathways [54]. Further, miRNA expression profiles can be easily accessed from formalin-fixed paraffin-embedded (FFPE) tissue which are gathered in the frame of clinical routine diagnostics carried out on stereotactic biopsies or surgically resected tumour tissues [55].

3.2.2 Summary

We profiled global miRNA expression patterns using microarray technology in a standard-treated retrospective cohort of 36 GBM patients. A prognostic cox-proportional hazard model was generated after applying iterative forward-selection feature selection in combination with overall survival as clinical outcome endpoint. A signature consisting of four miRNAs was identified that significantly predicted overall survival in a retrospective validation cohort ($n = 58$) that was matched for age, sex and MGMT promoter methylation status. The signature was independent from age, sex and MGMT methylation status and identified a high- and low-risk group that differed in the risk for death in the discovery and validation cohorts. The signature was technically validated in the discovery cohort by qRT-PCR. At the functional level matched miRNA and transcriptome data were used for correlation analyses in order to identify genes that are likely to be regulated by the signature miRNAs.

3.2.3 Methods, Results and Discussion

The FFPE sections of resected tumour tissues from 36 GBM patients were subjected to total RNA isolation followed by miRNA expression profiling using Febit human miRNA microarrays. The resulting data were quality filtered and quantile normalised. In conjunction with overall survival time and censoring status the data were subjected to iterative forward-selection for feature selection using the R package rbsurv [56]. Based on the Akaike Information Criterion (AIC) which provides a

trade-off between complexity and model deviance four miRNAs were selected: hsa-let-7a-5p, hsa-let-7b-5p, hsa-miR-125a-5p and hsa-miR-615-5p. A risk score was built using linear combination and high- and low-risk groups were defined using its median. Although not meaningful, in the discovery cohort, the high-risk and low risk groups showed a hazard-ratio of 3.8 (95%-CI 2.03-12.85, log-rank test p-value: 0.0001, Fig. 4). In the validation cohort this finding could be replicated after calculating the risk scores using linear combination of the model coefficients and expression of signature miRNAs and defining high- and low-risk groups using the median threshold as defined in the discovery cohort. In the validation cohort, the risk for death in the high-risk group was 2.1 times higher compared to the low-risk group (95%-CI 1.13-3.91, log-rank test p-value: 0.02, Fig. 5). Moreover, in both cohorts the signature was independent of sex, age and MGMT promoter methylation status. The latter is important with regard to a potential integration of the signature and MGMT promoter methylation which has the potential to identify strata with even more extreme differences in prognosis.

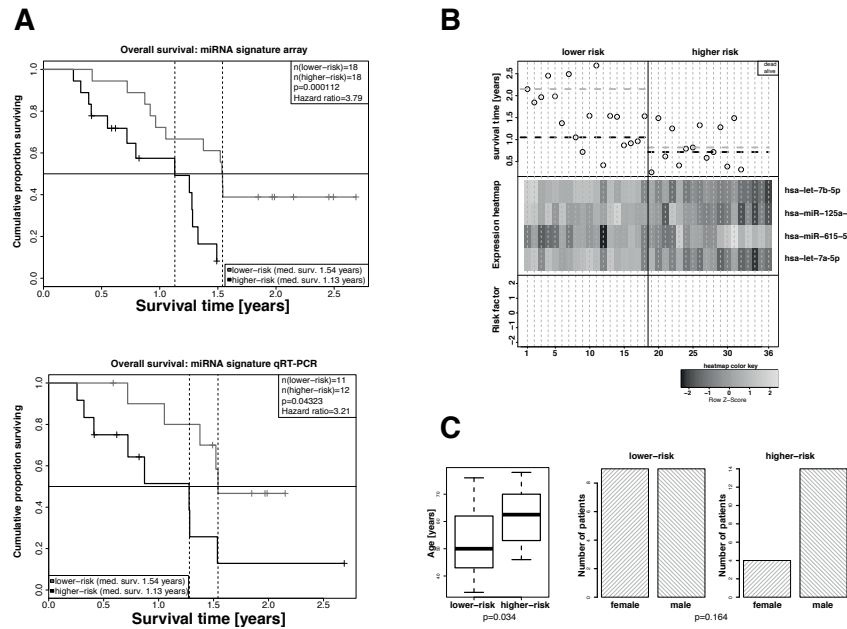


Figure 4: **4-miRNA signature as prognostic marker in the retrospective GBM cohort.** A. Kaplan-Meier overall survival analysis of high-risk and low-risk GBM patients. High-risk and low-risk patients were stratified based on the risk factors calculated from the cox-proportional hazard coefficients and the miRNA expression levels as measured in the microarray (left panel, 36 patients) or by qRT-PCR analyses (bottom panel, 19 patients). Hazard ratios and p-values were calculated by log-rank test. B. Overall survival (top panel), hierarchical cluster heat map of miRNA array expression levels (middle panel), and risk factors calculated on the basis of miRNA expression values and cox-proportional hazard coefficients (bottom panel) for all patients. miRNAs hsa-let-7a-5p, hsa-let-7b-5p and hsa-miR-125a-5p in patients of the higher-risk group show a tendency towards lower expression and that of hsa-miR-615-5p a tendency towards higher expression. The median risk factor value was used to classify high-risk and low-risk patients. C. Distribution of age (left panel) and sex (middle and right panels) in high-risk and low-risk GBM patients. Statistical comparison was performed by Student's t-test or Fisher's exact test. The patients of the lower-risk group were statistically significant older compared with that of the lower-risk group. The differences in the numbers of male and female patients of the lower- and higher-risk groups were not statistically significant.

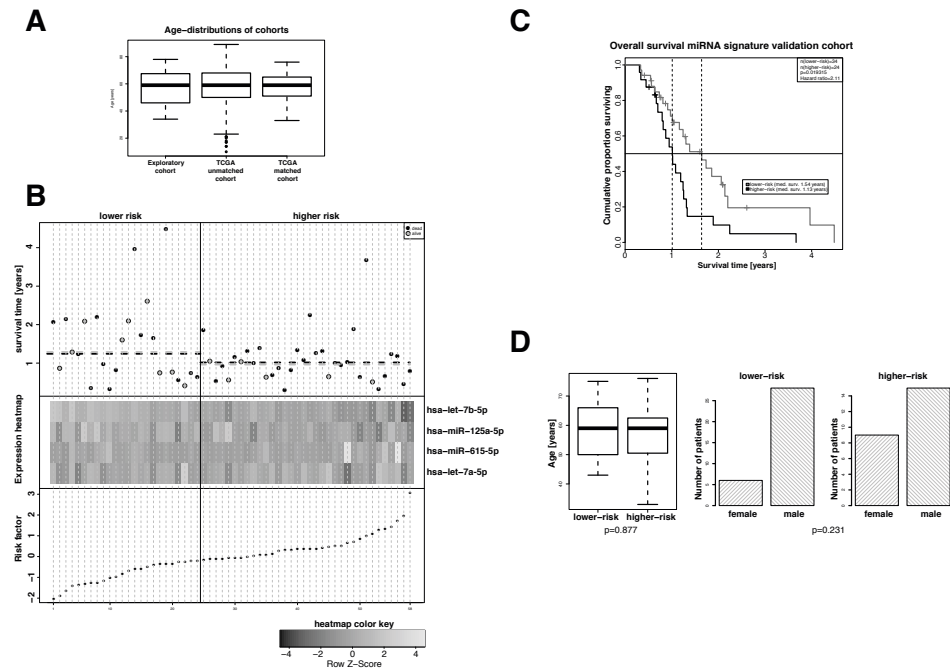


Figure 5: Validation of 4-miRNA GBM signature in retrospective validation set. A. Age distribution in the exploratory cohort and the TCGA GBM cohort before and after age matching. B. Overall survival (top panel), hierarchical cluster heat map of miRNA expression levels (middle panel), and risk factors for patients of the age- and sex-matched TCGA GBM cohort. The median risk factor value was used to classify high-risk and low-risk patients. miRNAs hsa-let-7a-5p, hsa-let-7b-5p and hsa-miR-125a-5p in the high-risk group show a slight tendency towards lower expression and that of hsa-miR-615-5p a slight tendency towards higher expression. C. Kaplan-Meier overall survival analyses of high- and low-risk patients of the matched TCGA GBM cohort. Classification was performed on the basis of the risk factors calculated from the cox-proportional hazard coefficients and the miRNA expression levels. Hazard ratios and p-values were calculated by log-rank test. D. Distribution of age (left panel) and sex (right panel) in high-risk and low-risk patients of the age- and sex-matched TCGA GBM cohort. Student's t-test and Fisher's exact test were employed for statistical comparison as depicted.

3.2.4 Conclusion

In the study we were able to identify a prognostic 4-miRNA signature that allows to identify high- and low-risk GBM patients. This signature has the potential to identify patients which would profit from therapy options different from the current standard-of-care setting.

3.2.5 Outlook

In order to prove validity of the miRNA signature the level of independent validation has to be increased by testing in further miRNA data sets from standard-treated GBM patients. Further, validation in a prospective setting is required. Another aspect is the evaluation of signature expression in blood plasma samples in order to test potential prognostic significance as liquid biopsy marker.

3.3 A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer.

Wilke CM, Braselmann H, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Walch AK, Selmansberger M, Samaga D, Weber P, Schneider L, Fend F, Bösmüller HC, Zitzelsberger H, **Unger K** A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer. *Int J Cancer*. 2018 Sep 15;143(6):1505-1515. (IF: 7.3)

3.3.1 Background

Breast cancer is known to be a heterogeneous disease which is associated with a number of risk factors such as life style, smoking, age. In addition, incidence of the disease has been associated with exposure to ionising radiation at the epidemiology level [57, 58, 59]. However, at the molecular level no radiation-specific mechanisms of radiation-associated breast cancer tumourigenesis have been identified so far. In the course of the clean-up activities after the Chernobyl reactor accident on the 26th April 1986 so-called "liquidators" were employed to remove nuclear waste from the reactor facilities. These clean-up workers were exposed to significant doses of ionising radiation in the range of only a few to hundreds of milligrays. Amongst these, also female workers have been employed and exposed and increased breast cancer incidence rates have been reported for most regions that were contaminated with radioactive fallout in the aftermath of the Chernobyl accident including oblasts in Russia, Belarus and the Ukraine [60, 61, 62].

3.3.2 Summary

We set up a comparative study on a cohort of female breast cancer patients from women who were exposed to ionising radiation in the course of Chernobyl clean-up activities and a case-by-case matched control cohort of sporadic female breast cancer patients who were not exposed and who were from the same regions of residence. We selected genomic copy number as the molecular level of interest since the DNA double strand break is the primary relevant molecular lesion in cells caused by ionising irradiation. A signature comprising nine different genomic copy number regions was established in a randomly selected training subset of the data which was subsequently validated in the remaining data.

3.3.3 Methods, Results and Discussion

FFPE sections from 68 patients were used for the generation of the training data set and that from the remaining 68 patients for the validation data set. All sections were reviewed by a pathologist for the purpose of diagnosis and typing of the estrogen- and progesteron-receptor expression and HER2 after immunohistochemistry staining. Further, the cellularity was determined and tumour regions marked prior macrodissection.

Macrodissected tumour tissues were subjected to DNA extraction using the Qiagen AllPrep kit. The genomic DNA of the tumours was labelled with Cy3-dCTP and that of pooled male reference DNA (Promega) with Cy5-dCTP using random-prime labelling prior hybridisation on Agilent 8x60k human aCGH arrays. After washing the slides were scanned using an Agilent array scanner.

The resulting spot intensities were imported into the R statistical platform. Log₂ ratios were built after correction for spatial artefacts and median normalisation [63]. The copy number profiles were

subjected to circular binary segmentation, copy number calling and copy number regions calculation using the R packages CGHcall and CGHregions [64, 65]. In order to build a mathematical model predicting exposure status a multivariate logistic regression approach was used while exposure status was used as response variable and genomic copy number status (-1 loss, 0 normal and 1 gain) of the signature regions were used as independent variables. The signature regions were identified by a stepwise-forward-selection/backward-elimination feature selection approach. Akaike information criterion (AIC) was used for selection of the best performing model. A signature consisting of nine genomic copy regions on chromosomal bands on chromosomal bands 7q11.22–11.23, 7q21.3, 16q24.3, 17q21.31, 20p11.23–11.21, 1p21.1, 2q35, 2q35 and 6p22.2 was identified. The performance of the signature is illustrated in Figure 6. From the 68 patients (34 exposed and 34 non-exposed) in the validationset 45 were predicted as exposed and 23 as non-exposed. This results in a true-positive rate of 0.794 and a false-positive rate of 0.529. The ROC area under the curve (AUC) was 0.617 and thereby greater than an area of 50% as it would be expected just by chance (Figure 7).

The signature is, therefore, with some statistical uncertainties able to predict breast cancer that is likely of having developed in the course of exposure to ionising radiation. The uncertainties in prediction can be explained by the fact that the range of radiation doses the exposed patients have received was wide and real ground truth with regard to known radiation-induced and spontaneous breast cancer cases cannot be formulated. However, the results suggest that a molecular signature of genomic copy number changes differentiates radiation-exposed and non-exposed breast cancers.

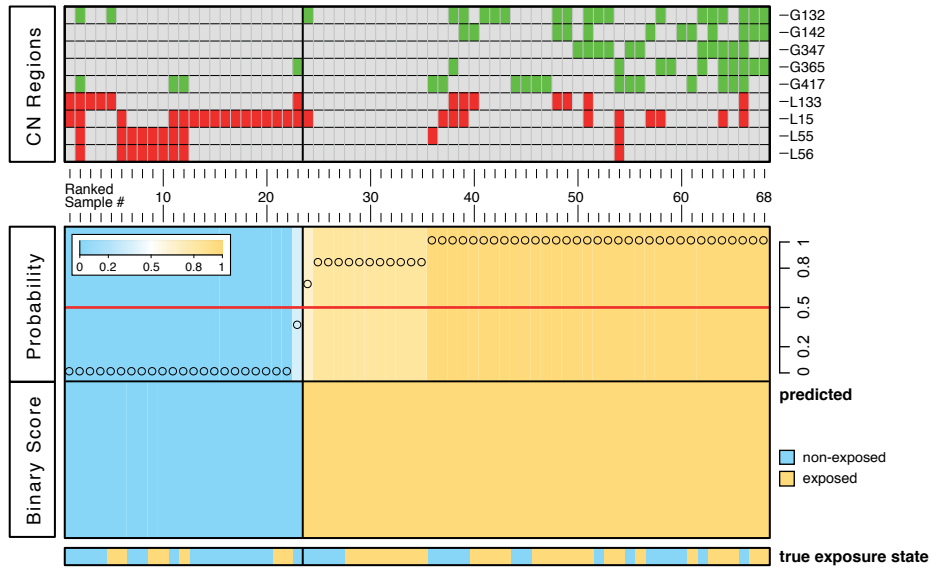


Figure 6: **Performance of 9-CNA signature in validation set.** Copy number gains (green) and losses (red) are shown in the top panel. The risk score on the probability scale is indicated in the middle panel while data are sorted according to the ascending order of the risk score per patient. Patients with probabilities greater than 0.5 are predicted as exposed, otherwise as non-exposed (middle panel, right and left side, respectively). The exposure status of each patient is indicated in the low panel, thus on the right yellow cases mark true positives, blue cases mark false positives. On the left side yellow cases mark false negatives, blue cases mark true negatives.

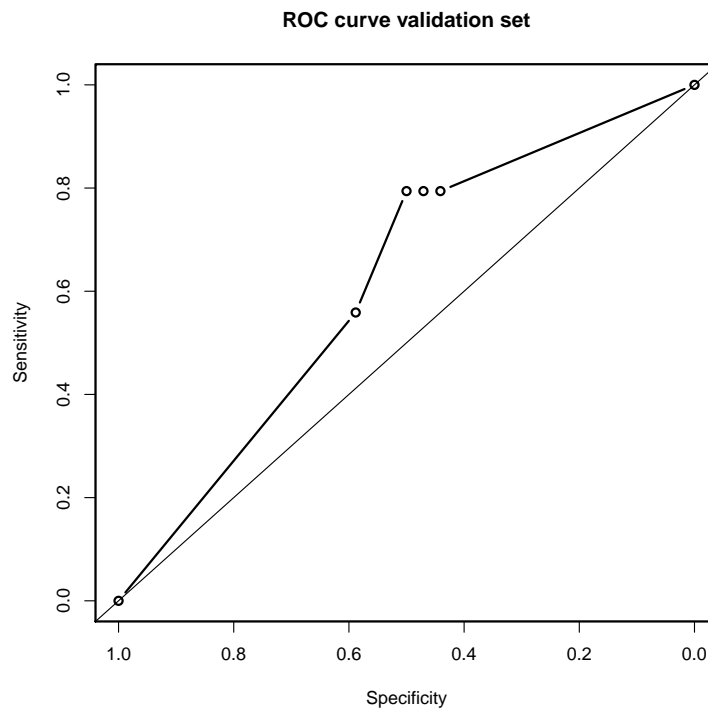


Figure 7: **Receiver Operator Characteristics (ROC) analysis.** ROC analysis of the logistic regression model using the 9-CNA signature fitted on the training set and evaluated on the validation set. Each point (circles) corresponds to a probability cutoff level decreasing from left to right, given by the steps visualized in Figure 6. Points are connected by straight lines.

3.3.4 Conclusion

We identified a genomic copy number signature that allows the differentiation of breast cancers from patients who were exposed and those who were not exposed to ionising radiation. It bears the potential to study radiation-specific molecular features of radiation-induced breast cancer such as secondary breast cancer that develops in the course of medical radiation.

3.3.5 Outlook

In order to translate the results of the study into the clinical setting, verification of the signature in a clinically derived cohort of secondary breast cancers that developed the disease after exposure to therapeutic radiation of preceding malignant diseases such as tumours of the lung.

3.4 Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer

Wilke CM, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Feuchtinger A, Walch AK, Selmansberger M, Braselmann H, Schneider L, Pitea A, Steinhilber J, Fend F, Bösmüller HC, Zitzelsberger H, **Unger K** *Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. Int J Cancer.* 2018 Feb 1;142(3):573-583. (IF: 7.3)

3.4.1 Background

Long-term data on the survivors of the Japan atomic bombings have shown, amongst other cancer entities, an increase in breast cancer that was associated with the exposure to ionising radiation [57]. Moreover, breast cancer also is amongst the diseases that have been shown to be increased in incidence after the Chernobyl accident in 1986 [60]. The important role of epigenetics and post-transcriptional regulation via miRNAs has been increasingly acknowledged in the last decades and a number of miRNAs have been shown to be associated with breast cancer carcinogenesis. We took up this knowledge and conducted a study in which we studied the role of these miRNAs in radiation-associated breast cancer that has developed in patients who were exposed to ionising radiation in the course of their role as female clean-up workers aka "liquidators" after the Chernobyl reactor accident in 1986 in comparison to spontaneous breast cancer cases from residence-matched non-exposed controls. The main aim of our study was to determine a potential radiation-specific role of the miRNAs and even importantly of proteins these miRNAs are likely to regulate.

3.4.2 Summary

We identified breast cancer associated miRNAs by literature research and tested these for differential expression in a cohort of 77 exposed and 77 non-exposed breast cancer patients from the Ukraine. The patient cohort was randomly split into a discovery (n=76) and validation (n=78) set. From the tested miRNAs only hsa-miR-26b-5p was significantly higher expressed in the exposed compared to the unexposed patients. The protein TRPS1, which is one of the transcriptional targets of hsa-miR-26b-5p was significantly lower expressed in the exposed compared to unexposed cases. TRPS1, which is a transcription factor, would be particularly suitable as a radiation marker in breast cancer since it would be technically feasible to detect from diagnostic routine samples. After siRNA knockdown of the TRPS1 gene in a breast cancer cell culture model we identified genes playing a role in DNA-repair, cell cycle, mitosis, cell migration, angiogenesis and EMT pathways.

3.4.3 Methods, Results and Discussion

From the 77 breast cancer patients who were exposed and from the 77 patients who were not exposed to radiation formalin-fixed paraffin-embedded tissue sections were available. The tissue sections were assessed by a pathologist for diagnosis and the definition of tumour regions prior macrodissection and isolation of DNA and total RNA including small RNAs using the Qiagen All-Prep kit. Further, the pathologist provided estrogen, progesterone, HER2, c-kit, cytokeratin 5/6, P53, Ki-67 and BRCA1/2 status. Further, for all exposed cases the doses the patients received was reconstructed using the RADRUE method [66]. The majority of tumours was diagnosed as invasive

carcinoma of no special type (NST). Further types diagnosed were invasive lobular carcinoma (ILC), intracystic papillary breast carcinoma and breast carcinomas with medullary features.

In order to find miRNAs repeatedly reported to play a role in breast cancer carcinogenesis in the context of radiation a PubMed research was conducted which resulted the miRNAs hsa-miR-26b-5p, hsa-miR-99b-5p, hsa-miR-221-3p and hsa-miR-222-3p for which the expression was subsequently determined using qRT-PCR. Only hsa-miR-26b-5p could be shown to be overexpressed in the exposed cases compared with the unexposed (Fig. 8).

The expression of TRPS1 which is a transcriptional target of hsa-miR-26b-5p was determined by immunohistochemistry on FFPE slides and was significantly reduced in exposed compared to unexposed cases. This finding suggests a regulatory effect of hsa-miR-26b-5p on TRPS1 protein expression (Fig. 9 and Fig. 10).

Since expression of the transcription factor TRPS1 was specifically down-regulated in exposed cases we strove for investigating the mechanistic impact of TRPS1 on the transcriptional level by siRNA knock-down in a radiation-transformed breast cancer cell culture model B42-16 and by correlation analyses using the TCGA breast cancer data set [67]. The transcriptome was characterised using Agilent human gene expression array analysis comparing B42-16 cells after TRPS1 siRNA knockdown with scrambled controls. Pathway enrichment of differentially upregulated genes revealed DNA-repair, cell cycle and mitosis and that of down-regulated genes cell migration, angiogenesis and EMT. Using the TCGA breast cancer data set we generated a TRPS1-centered correlation network (Fig. 11). Pathway enrichment analysis of the network revealed mostly apoptosis related pathways. The transcriptome analysis results suggest involvement of TRPS1 important cancer hallmarks. Thus, TRPS1 could be a early radiation-induced event in the carcinogenesis of the breast after exposure to ionising radiation.

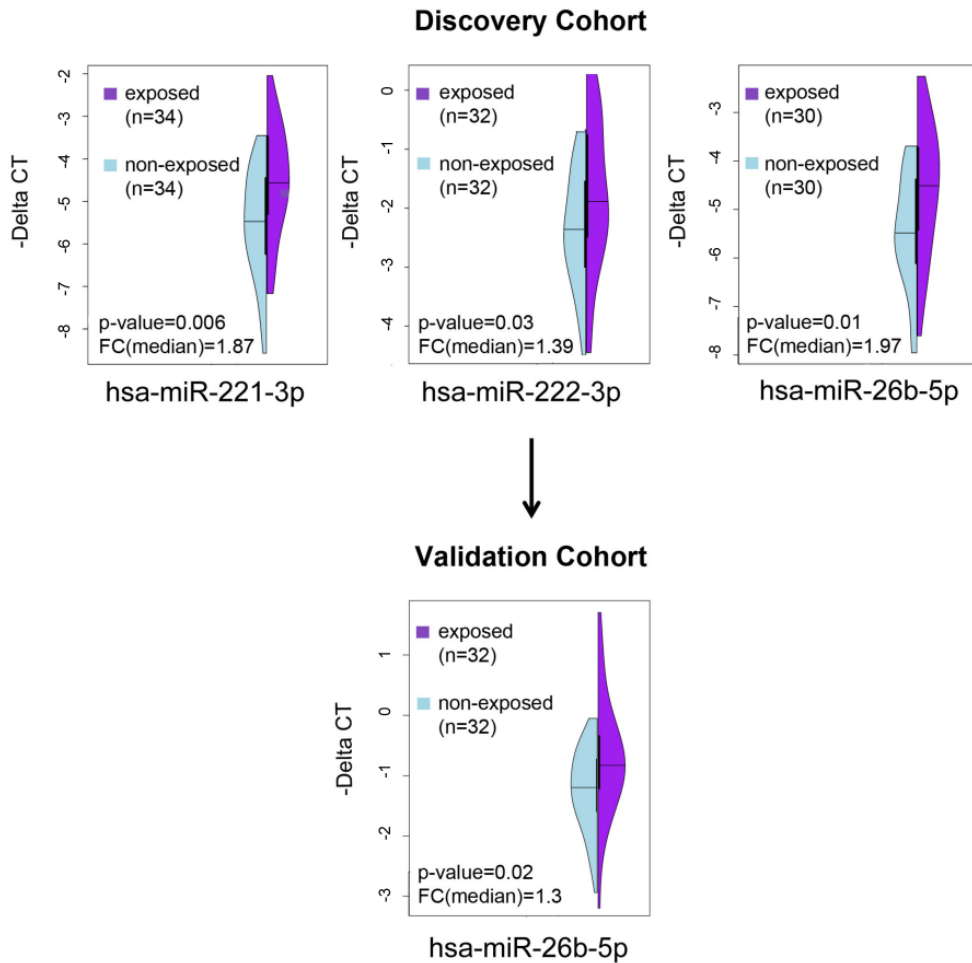


Figure 8: **Expression of breast cancer associated miRNAs.** Violin plots displaying the expressions of hsa-miR-26b-5p, hsa-miR-221-3p and hsa-miR-222-3p in the Chernobyl discovery cohort and hsa-miR-26b-5p in the Chernobyl validation cohort measured by qRT-PCR ($-\Delta\text{CT}$ values) are shown (right panel). The nonexposed control group is labeled in light blue and the exposed group in purple. The middle dark line represents the median of expression values. The vertical black line represents the interquartile.

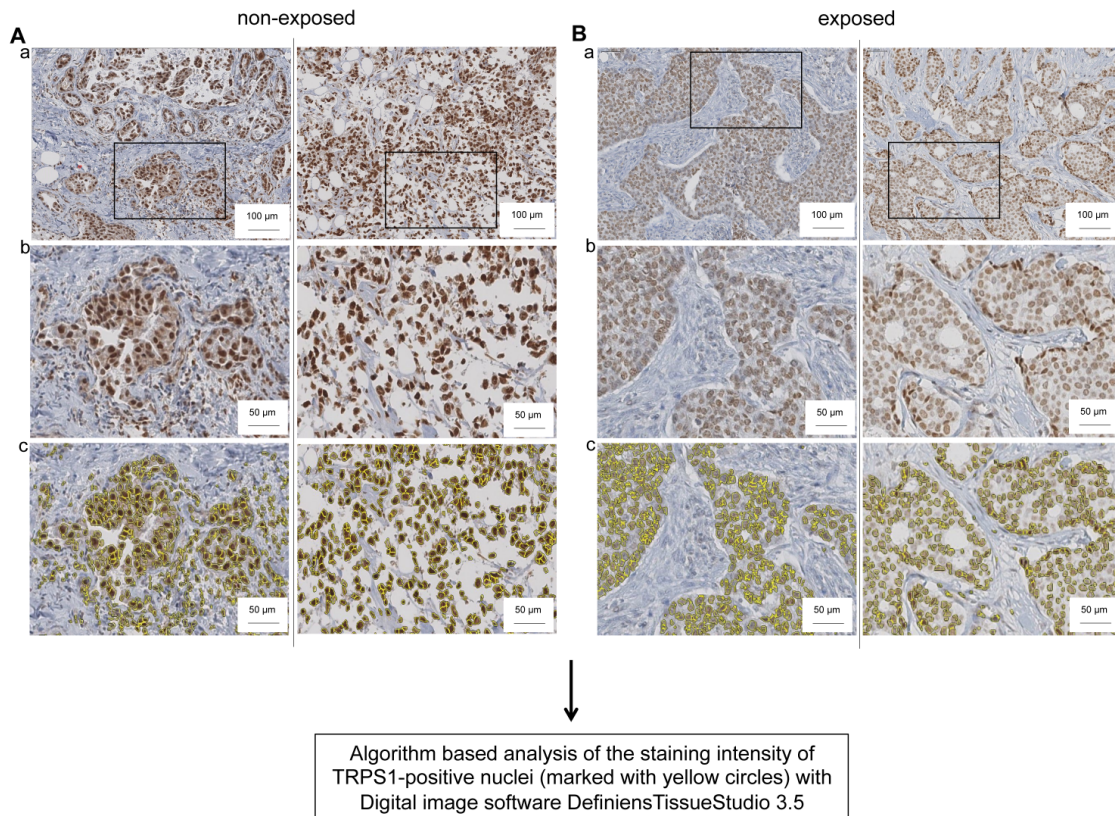


Figure 9: **TRPS1 immunohistochemistry staining** Digital image analysis of immunohistochemically stained FFPE tumour sections from non-exposed and exposed breast cancer samples using an antibody against TRPS1. (A/B) Two representative immunohistochemically stained breast carcinoma cases are shown for nonexposed (A) and exposed (B) cases. Image details of Aa and Ba (black frames) are shown in Ab and Bb. Detection and quantification of TRPS1-stained nuclei was performed using the digital image software Definiens. Nuclei of tumour cells, for which the staining intensities were calculated based on the algorithm, are labeled in yellow (Ac, Bc).

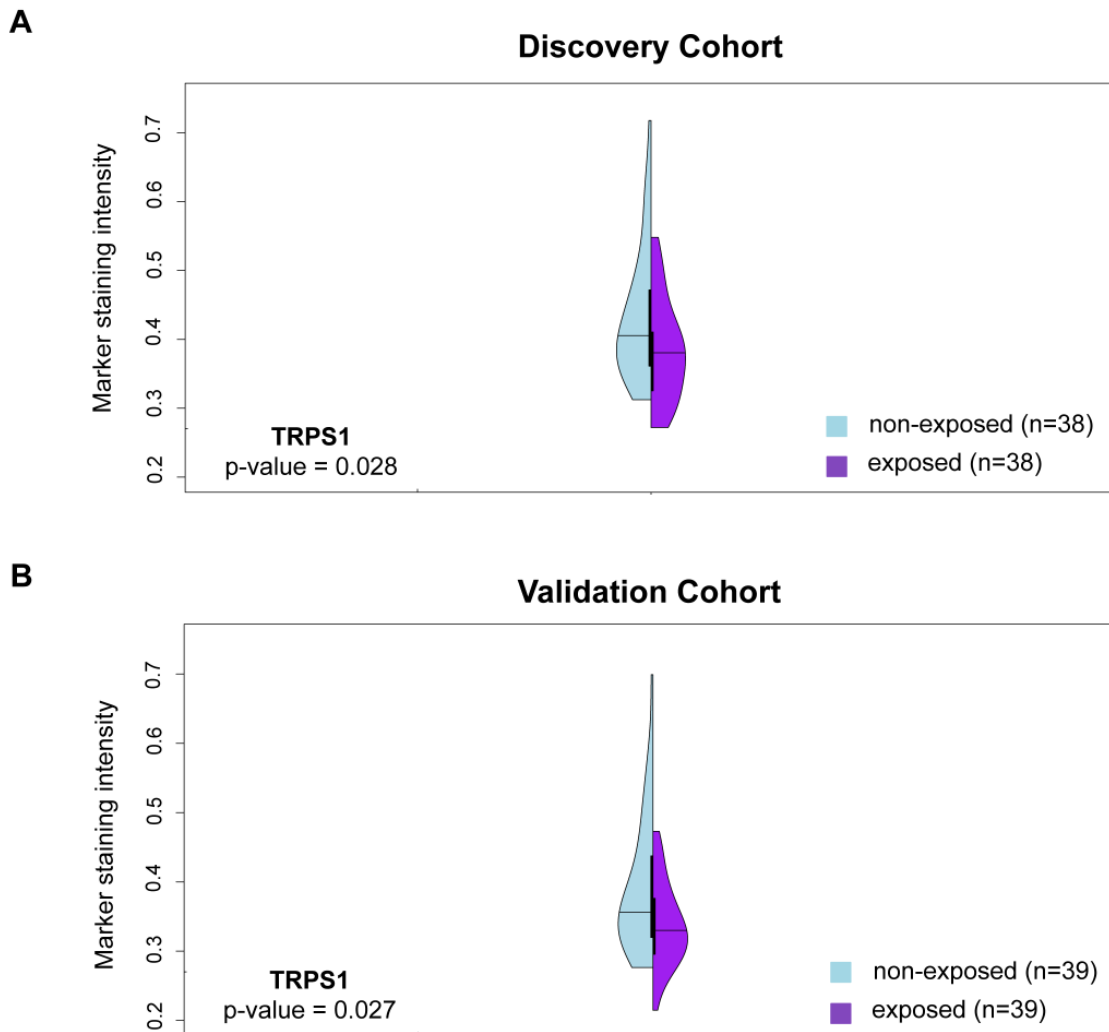


Figure 10: **Differential TRPS1 protein expression in exposed and unexposed cases.** Significantly increased TRPS1 protein expression represented by the marker staining intensity was observed in breast cancer tissues from the nonexposed groups (light blue) compared to the exposed groups (purple) in the discovery (a, $p = 0.028$) and validation cohorts (b, $p = 0.027$). p values were calculated using the partial differential test considering intertumour heterogeneity.

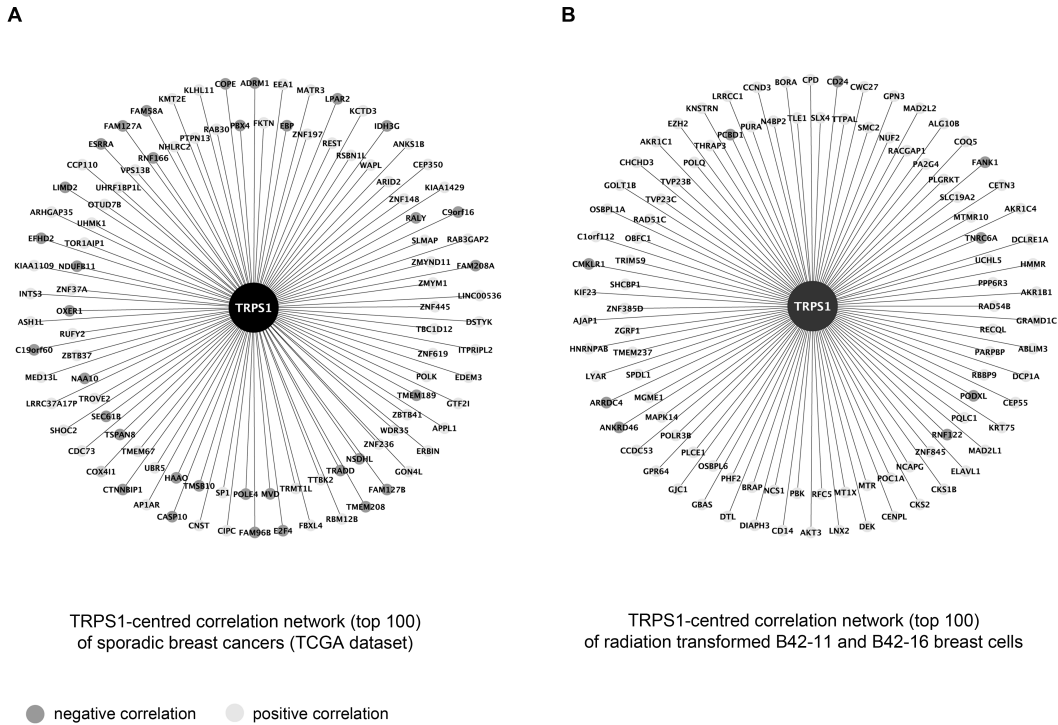


Figure 11: **TRPS1-centered correlation networks.** Top 100 correlating genes with an FDR <0.05. The expression of genes labeled with dark grey circles showed negative correlation with TRPS1 expression and that of genes labeled with light grey circles showed positive correlation with TRPS1 expression. (a) TRPS1-centered correlation network based on global mRNA expression data from matched sporadic breast cancers of the publicly available TCGA dataset. (b) TRPS1-centered correlation network based on microarray gene expression data from B42-11 and B42-16 untransfected, scrambled-siRNA transfected and TRPS1-downregulated cells.

3.4.4 Conclusion

We were able to identify the radiation markers miRNA hsa-miR-26-5p and the protein TRPS1 which have the potential to serve as a marker for radiation-induced secondary breast cancer in the clinical setting.

3.4.5 Outlook

In order to study a potential role as markers for secondary, radiation-induced breast cancer prospective validation in clinically derived cohorts is required. Moreover, further *in vitro* and *in vivo* experiments are necessary to further investigate the mechanistic role of TRPS1 and hsa-miR-26b-5p in radiation-associated breast carcinogenesis.

3.5 Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data

Michna A, Braselmann H, Selmansberger M, Dietz A, Hess J, Gomolka M, Hornhardt S, Blüthgen N, Zitzelsberger H, **Unger K**. *Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data*. **PLoS One**. 2016 Aug 9;11(8). (IF: 2.8)

3.5.1 Background

The characterisation of the transcriptome and its analysis plays an important role in modern molecular biology while the majority of data sets still is conducted on steady-state static datasets that simply work out differential expression based on t-testing approaches. However, time-resolved dynamic transcriptomic analyses are likely to be much more informative but require tailored approaches to analyse differential expression over time. One major approach to extract and analyse the information from time-resolved transcriptome data from another angle is the reconstruction of gene regulatory networks (GRN) that allow gathering insights into the interplay of genes expressed by network modules. And in addition, the resulting networks and network modules can be analysed towards the involvement of molecular pathways. In this study we developed the R package SplineTimeR which provides a workflow for the conduction of the afore described computational tasks.

The analysis workflow of the study was established using time-resolved transcriptome data from lymphoblastoid cells lines that were generated in the frame of the LUCY (Lung Cancer in the Young) study and which differed in radiation sensitivity. Results on differential expression over time after gamma irradiation with 1 Gy and 10 Gy, as well as pathway analysis of the reconstructed GRN is presented in the study.

3.5.2 Summary

We developed the R package SplineTimeR which provides functions required for an analysis workflow for time-resolved transcriptome data and apply this workflow in order to work out the transcriptome response of a radiation-normal and -hypersensitive lymphoblastoid cell line. The developed natural cubic spline regression modelling (NCSRМ) approach for time-resolved differential transcriptomics showed superior performance in comparison with existing approaches.

3.5.3 Methods, Results and Discussion

A graphical outline of the study is shown in Fig. 12.

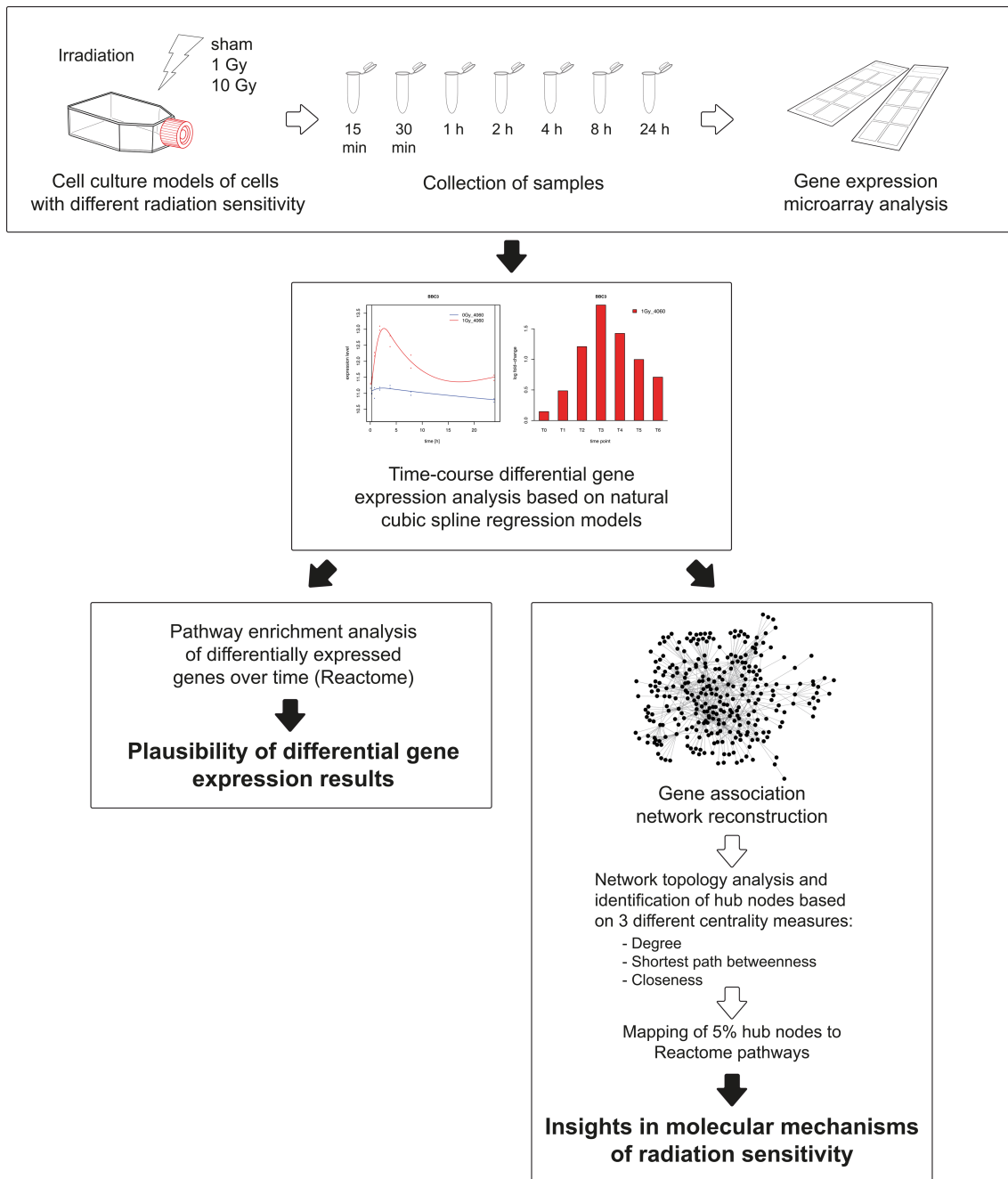


Figure 12: Graphical outline of the study.

The transcriptome profiles of biological replicates of the normal and the hypersensitive cell lines were generated 15 min, 30 min, 1 h, 2 h, 4 h, 8 h and 24 h after irradiation or sham irradiation. For this Agilent human gene expression microarrays were used. The profiles of the duplicates were averaged and then analysed for differential gene expression between irradiated and sham control. We applied the NCSR approach on the data sets and revealed a sparse response at the gene expression level of the normal sensitive cells after 1 Gy (7 genes) compared to that of the hypersen-

sitive cells (2335 genes). An example for NCSRSM differential expression of the gene BBC3 is shown in Fig. 13. After 10 Gy irradiation both cell lines showed a massive transcriptomic response: 3892 genes for the normal and 6019 genes for the hypersensitive cells. For assessing plausibility of the results pathway enrichment analysis was conducted for the differentially expressed genes. This resulted in no pathway for the normal sensitive cell line after 1 Gy and in hundreds of pathways for the other comparisons which all showed cell cycle and cell division related pathways at the top of the lists.

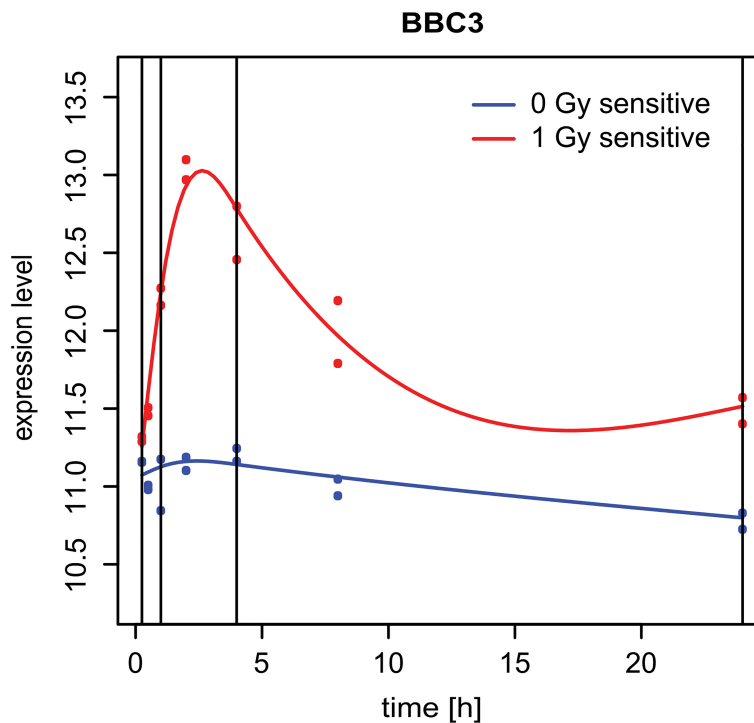


Figure 13: NCSRSM fit for gene BBC3 in the hypersensitive cells The blue line represents the fitted model for the control (0 Gy) and read line that for the irradiated group (1 Gy). Blue and red dots represent the measured expression levels of the biological replicates. Vertical lines represent the endpoints and interior knots correspond to the 0.33- and 0.66-quantiles.

The expression profiles of these genes were then subjected to GRN reconstruction using the GeneNet approach which resulted in networks that in size (number of nodes/edges) were proportional to the number of underlying differentially expressed genes: normal sensitive 1 Gy no network, hypersensitive 1 Gy 1140/12198 (nodes/edges), normal sensitive 10 Gy 2735/84695 (nodes/edges) and hypersensitive 10 Gy 3483/114629 (nodes/edges). We identified the top most important genes (i.e. nodes) from the networks using a combined network analysis metric and subjected these genes to pathway enrichment analysis. Normal sensitive cells after 10 Gy and hypersensitive cells after 1 Gy showed strong association with cellular senescence while hypersensitive cells after 10 Gy showed enrichment of apoptosis related pathways.

We also evaluated the differential expression results and reconstructed GRNs that were received after NCSRSM and another established method for the analysis of time-course transcriptome data BETR (Bayesian Estimation of Temporal Regulation)[68]. Overall, NCSRSM detected more genes as

differentially expressed compared to BETR but there was a great overlap between NCRSM and BETR results. Also when comparing the top hub genes of the GRNs there was a good overlap between enriched pathways. In line with the literature BETR seems to underestimate the number of truly differentially expressed genes and thereby is likely to miss potentially important information. Also, in contrast to BETR NCRSM is more flexible and tolerant when it comes to selection of time-points and missing data. This is an important feature since it is not a rare scenario that the design of transcriptomic data analysis is not optimised for the purpose of differential expression analysis. In a further comparison where we compared the overlap of GRNs reconstructed from NCRSM and BETR identified differentially expressed genes with the interactions as known from the Reactome Interaction Network database. Except one network we saw significantly better overlap between NCRSM derived GRNs with Reactome than with that of the BETR derived ones [69].

3.5.4 Conclusion

We established a workflow for the differential gene expression, GRN reconstruction and functional interpretation of the GRNs of time-resolved transcriptome data. Our approach is flexible and tolerant against frequently occurring experimental uncertainties such as suboptimal study design or missing data points. Analysis of the biological data set revealed differences and common properties in the molecular radiation response. We provide the workflow as a Bioconductor R package and thereby allow public and free access to it for the scientific community.

3.5.5 Outlook

Since GRNs are derived from transcriptome data the identified interactions between genes could reflect direct or indirect relationships of the expressed proteins. Thus, the physical interaction between nodes identified in the GRNs and their biological meaning will be subject of future studies.

3.6 Copy number aberrations from Affymetrix SNP 6.0 genotyping data - how accurate are the commonly used prediction approaches?

Pitea A, Kondofersky I, Sass S, Theis FJ, Mueller NS, Unger K. Copy number aberrations from Affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches? *Briefings in Bioinformatics*. 2018;doi:10.1093/bib/bby096, (IF: 6.3)

3.6.1 Background

Genomic copy number changes (CNA) are frequently occurring in most cancer entities and different approaches are available for typing CNAs while the method of choice is array comparative genomic hybridisation (aCGH). However, also other sources of data which are not primarily for this purpose can be used for typing CNAs including DNA methylation array data, exome and whole genome sequencing data and SNP (short nucleotide polymorphism) array data. In our study we focussed on Affymetrix SNP 6.0 data generated from clinical tumour specimen. Since we are working in the field we used the Affymetrix SNP 6.0 data on the TCGA (The Cancer Genome Atlas) head and neck squamous cell carcinoma (HNSCC) data set [70]. In addition and in order to be able to assess the performance of different approaches we used a simulated Affymetrix SNP 6.0 like dataset. The main aim of the study was, to compare the performance of established analysis approaches including OncoSNP, ASCAT, CGHcall, genoCNA and GISTIC with each other. Further, we improved the tumour-derived data tailored approach CGHcall for the purpose which we called CGHcall*. Moreover, we compared the CNA results of the assessed analysis approaches on the TCGA HNSCC data set with published CNA data on HNSCC.

3.6.2 Summary

In our study we comprehensively compared the performance of the copy number calling algorithms OncoSNP, ASCAT, CGHcall, genoCNA and GISTIC and considered confounding or biasing parameters tumour purity, length of CNA and CNA burden. Amongst the tested algorithms CGHcall provided the best performance with regard to prediction call accuracy. However, we observed that the accuracy of CGHcall drops once the CNA burden exceeds 50% of the genome. CGHcall*, an adjusted version of CGHcall was implemented and we could demonstrate its improved performance. The scripts of the workflow and conducted analyses is provided to the community in GitHub.

3.6.3 Methods, Results and Discussion

For performance assessment two data sets were used: A synthetic Affymetrix SNP 6.0 data set that was generated using the jointseg R package and the HapMap dataset on naturally occurring CNAs in the human populations for which comprehensive experimental validation data exist [71, 72]. For the evaluation of the performance on a realistic tumour sample derived data set the raw data of the TCGA HNSCC data were used. All SNP 6.0 data were preprocessed using Affymetrix Power Tools (APT) in order to receive LRR (probewise LogR-ratio) values and B-allele frequencies which were already available for the simulated data. All data were then subjected to analysis with the genomic copy number analysis tools OncoSNP, ASCAT, CGHcall, genoCNA and GISTIC. Performance for the simulated data was assessed using F score statistics which integrates precision

(positive predicted value) and recall (sensitivity). Good prediction is reflected by F-score values towards 1 and bad prediction by values towards 0. Differences between F-score distributions were assessed using Wilcoxon Cox test and resulting p-values were Bonferroni corrected.

When manually inspecting the copy number calling profiles we observed that in case of profiles with more than half of the genome changed in copy number reflected by the segmentation profiles, prediction of the CGHcall algorithm fails. This problem was caused by estimating the baseline copy number for which the median of all LRR values per profile is used and which fails in case of more than 50% of the profile being altered. We solved this by only considering LRRs which are covered by the interval $[-0.1, 0.1]$. This correction was implemented in the CGHcall workflow which we then called CGHcall*.

CGHcall* included, the prediction F-scores were determined for the 100 simulated genomic copy number data. When comparing the performance of the algorithms for profiles with different simulated tumour purity (i.e. cellularity) huge differences were obvious. GISTIC was not able to correctly predict CNAs in profiles with less than 100% tumour purity. The ability to correctly call losses was poor for all algorithms while CGHcall* and GenoCNA improved with higher tumour purity. The overall performance of the algorithms to call the normal state was not good for all algorithms but improved with higher purities. With purities greater than 70% CGHcall* and GenoCNA performed best. Prediction F-scores also for gains were not good for tumour purities below 70% for all algorithms but significantly improved for CGHcall* and GenoCNA for 70% purity and greater. For 100% purity CGHcall* and OncoSNP performed best. The results are summarised in Fig. 14

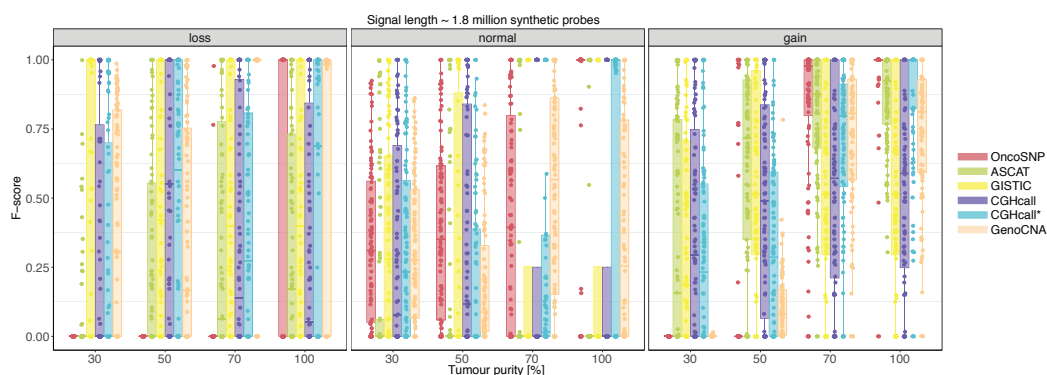


Figure 14: **Performance of CNA calling algorithms on synthetic data.** The y-axis represents the F-score and x-axis represents the tumour purity level in %. The three facets represent the different classes: loss, normal and gain. Each boxplot consists of F-scores for 100 synthetic samples. The total number of genetic markers covered by the synthetic signal was approximately 8 bp.

Another important factor in CNA analysis is the length of the underlying chromosomal segments. Optimally an algorithm performs equally well for short, medium and long segment lengths. We could show that for 100% tumour purity OncoSNP and CGHcall* perform best across different copy number segment lengths as outlined in Fig. 15 A. Finally, when comparing the performance of the algorithms dependent from the CNA burden OncoSNP and CGHcall* most accurately predicted the true copy number states when the CNA burden was greater 50% (Fig. 15 A).

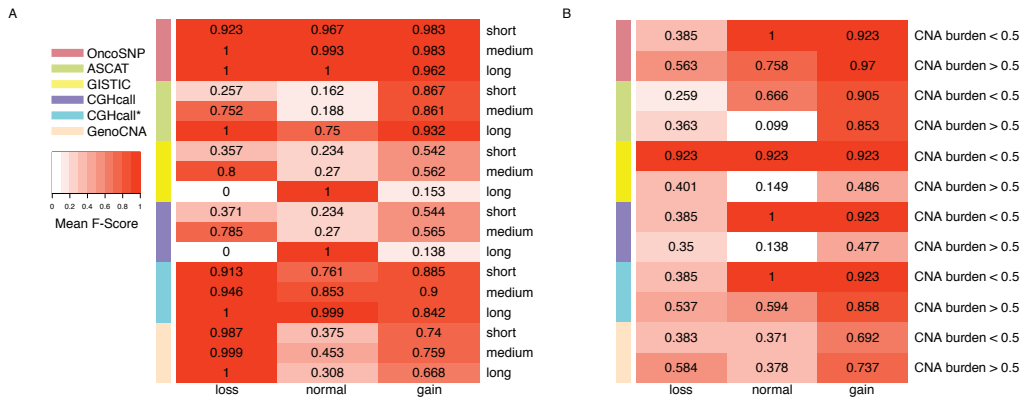


Figure 15: Heatmap of meanF-scores for different lengths of copynumber regions (A) and for CNA burdens smaller and greater 50%.

Further we tested performance of the algorithms on the non-tumour HapMap SNP 6.0 data (n = 81) for which comprehensive validation data were generated using array CGH and fluorescence *in situ* hybridisation and which we used as true state. Again, GenoCNA and CGHcall* performed almost similarly and best amongst all algorithms while the performance of GenoCNA was slightly better than that of CGHcall* (Fig. 16).

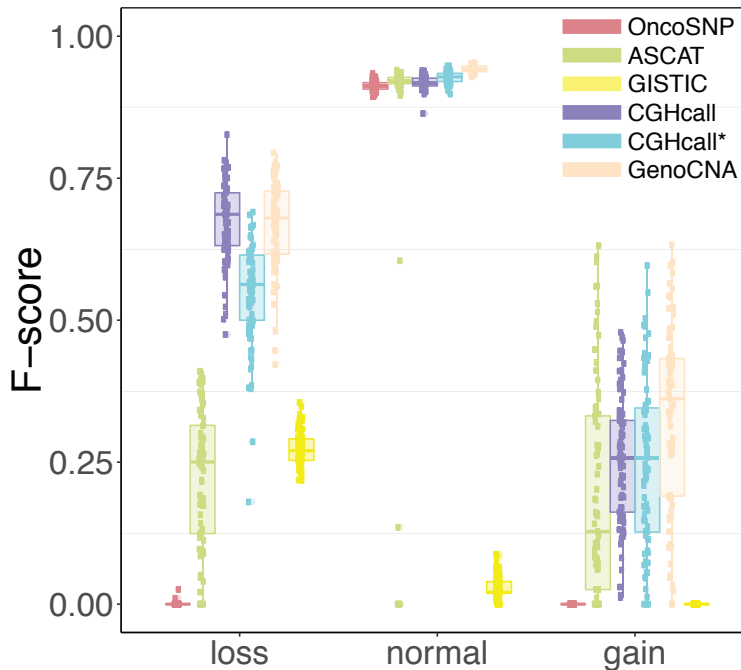


Figure 16: Heatmap of mean F-scores for different lengths of copynumber regions (A) and for CNA burdens smaller and greater 50%.

Finally, we typed the genomic copy number changes in the TCGA HNSCC data set. We com-

pared the frequencies of gene copy number alterations of genes CDKN2A and CCND1 which are frequently altered in HNSCC as reported by Gollin et al. (2014) with that determined in the TCGA data [73]. In this comparison CGHcall* and GISTIC showed the best comparable results while most profiles were from tumours with tumour purities greater 60%.

3.6.4 Conclusion

We evaluated the prediction performance of commonly used SNP 6.0 data genomic copy number calling algorithms and provide a pipeline for this purpose. Further, we could improve overall performance of CGHcall implemented in CGHcall*. Our study provides the basis for informed selection of the best suitable genomic copy number calling algorithm when using Affymetrix SNP 6.0 data.

3.6.5 Outlook

The provided pipeline could be used as blueprint for the performance assessment for copy number typing for other data types allowing for copy number typing such as genomic number from DNA methylation array data, whole exome or whole genome sequencing data.

Bibliography

- [1] Helleday T, Petermann E, Lundin C, Hodgson B, and Sharma RA. DNA repair pathways as targets for cancer therapy. *Nature reviews Cancer*, 2008. 8(3):193–204.
- [2] Begg AC, Stewart FA, and Vens C. Strategies to improve radiotherapy with targeted drugs. *Nature reviews Cancer*, 2011. 11(4):239–253.
- [3] Orth M, Lauber K, Niyazi M, et al. Current concepts in clinical radiation oncology. *Radiation and environmental biophysics*, 2014. 53(1):1–29.
- [4] Ogawa Y, Kubota K, Ue H, et al. Phase I study of a new radiosensitizer containing hydrogen peroxide and sodium hyaluronate for topical tumor injection: a new enzyme-targeting radiosensitization treatment, Kochi Oxydol-Radiation Therapy for Unresectable Carcinomas, Type II (KORTUC II). *International journal of oncology*, 2009. 34(3):609–618.
- [5] Ogawa K, Ishiuchi S, Inoue O, et al. Phase II trial of radiotherapy after hyperbaric oxygenation with multiagent chemotherapy (procarbazine, nimustine, and vincristine) for high-grade gliomas: long-term results. *International journal of radiation oncology, biology, physics*, 2012. 82(2):732–738.
- [6] Brown JM. Exploiting the hypoxic cancer cell: mechanisms and therapeutic strategies. *Molecular medicine today*, 2000. 6(4):157–162.
- [7] Seiwert TY, Salama JK, and Vokes EE. The concurrent chemoradiation paradigm—general principles. *Nature clinical practice Oncology*, 2007. 4(2):86–100.
- [8] Budach V, Stromberger C, Poettgen C, et al. Hyperfractionated accelerated radiation therapy (HART) of 70.6 Gy with concurrent 5-FU/Mitomycin C is superior to HART of 77.6 Gy alone in locally advanced head and neck cancer: long-term results of the ARO 95-06 randomized phase III trial. *International journal of radiation oncology, biology, physics*, 2015. 91(5):916–924.
- [9] Budach V, Stuschke M, Budach W, et al. Hyperfractionated accelerated chemoradiation with concurrent fluorouracil-mitomycin is more effective than dose-escalated hyperfractionated accelerated radiation therapy alone in locally advanced head and neck cancer: final results of the radiotherapy coo. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 2005. 23(6):1125–1135.
- [10] Seiwert TY, Salama JK, and Vokes EE. The chemoradiation paradigm in head and neck cancer. *Nature clinical practice Oncology*, 2007. 4(3):156–171.

- [11] Stupp R, Hegi ME, Mason WP, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet Oncology*, 2009. 10(5):459–466.
- [12] Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England journal of medicine*, 2005. 352(10):987–996.
- [13] Chalmers AJ, Ruff EM, Martindale C, Lovegrove N, and Short SC. Cytotoxic effects of temozolomide and radiation are additive- and schedule-dependent. *International journal of radiation oncology, biology, physics*, 2009. 75(5):1511–1519.
- [14] Dracham CB, Shankar A, and Madan R. Radiation induced secondary malignancies: a review article. *Radiation Oncology Journal*, 2018.
- [15] Armstrong GT, Liu Q, Yasui Y, et al. Late Mortality Among 5-Year Survivors of Childhood Cancer: A Summary From the Childhood Cancer Survivor Study. *Journal of Clinical Oncology*, 2009. 27(14):2328–38.
- [16] Mertens AC, Liu Q, Neglia JP, et al. Cause-specific late mortality among 5-year survivors of childhood cancer: The childhood cancer survivor study. *Journal of the National Cancer Institute*, 2008. 100(19):1368–79.
- [17] Hess J, Thomas G, Braselmann H, et al. Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proceedings of the National Academy of Sciences of the United States of America*, 2011. 108(23):9595–9600.
- [18] Selmansberger M, Kaiser JC, Hess J, et al. Dose-dependent expression of CLIP2 in post-Chernobyl papillary thyroid carcinomas. *Carcinogenesis*, 2015. 36(7):748–756.
- [19] Selmansberger M, Feuchtinger A, Zurnadzhy L, et al. CLIP2 as radiation biomarker in papillary thyroid carcinoma. *Oncogene*, 2015. 34(30):3917–3925.
- [20] Wilke CM, Hess J, Klymenko SV, et al. Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. *International journal of cancer*, 2018. 142(3):573–583.
- [21] Wilke CM, Braselmann H, Hess J, et al. A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer. *International journal of cancer*, 2018. 143(6):1505–1515.
- [22] Ma'ayan A. Introduction to network analysis in systems biology. *Science signaling*, 2011. 4(190):tr5.
- [23] Barker HE, Paget JTE, Khan AA, and Harrington KJ. The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. *Nature reviews Cancer*, 2015. 15(7):409–425.
- [24] Crick F. Central dogma of molecular biology. *Nature*, 1970. 227(5258):561–563.
- [25] Lazebnik Y. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer cell*, 2002. 2(3):179–182.

- [26] Bartocci E and Lió P. Computational Modeling, Formal Analysis, and Tools for Systems Biology. *PLoS computational biology*, 2016. 12(1):e1004591.
- [27] Dunkler D, Sánchez-Cabo F, and Heinze G. Statistical analysis principles for Omics data. *Methods in molecular biology (Clifton, NJ)*, 2011. 719:113–131.
- [28] Datta S, Datta S, Kim S, Chakraborty S, and Gill RS. Statistical Analyses of Next Generation Sequence Data: A Partial Overview. *Journal of proteomics & bioinformatics*, 2010. 3(6):183–190.
- [29] Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936. 8:3–62.
- [30] Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 1995. 57(1):289–300.
- [31] Hartmann C, Meyer J, Balss J, et al. Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: A study of 1,010 diffuse gliomas. *Acta Neuropathologica*, 2009.
- [32] Skinner HD, Sandulache VC, Ow TJ, et al. TP53 disruptive mutations lead to head and neck cancer treatment failure through inhibition of radiation-induced senescence. *Clinical Cancer Research*, 2012.
- [33] Dunn J, Baborie A, Alam F, et al. Extent of MGMT promoter methylation correlates with outcome in glioblastomas given temozolomide and radiotherapy. *British Journal of Cancer*, 2009.
- [34] King CR, Kraus MH, Williams LT, Merlino GT, Pastan IH, and Aaronson SA. Human tumor cell lines with EGF receptor gene amplification in the absence of aberrant sized mRNAs. *Nucleic Acids Research*, 1985.
- [35] Bar-Joseph Z, Gitter A, and Simon I. Studying and modelling dynamic biological processes using time-series gene expression data, 2012.
- [36] Opgen-Rhein R and Strimmer K. From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 2007.
- [37] Coates J, Souhami L, and El Naqa I. Big Data Analytics for Prostate Radiotherapy. *Frontiers in Oncology*, 2016.
- [38] Lindsey JK and Jones B. Choosing among generalized linear models applied to medical data, 1998.
- [39] Boulesteix AL and Hothorn T. Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics*, 2010.
- [40] Boulesteix AL and Sauerbrei W. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 2011.
- [41] Slawski M, Daumer M, and Boulesteix AL. CMA - A comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, 2008.

- [42] Kiiveri HT. Multivariate analysis of microarray data: Differential expression and differential connection. *BMC Bioinformatics*, 2011.
- [43] Charitou T, Bryan K, and Lynn DJ. Using biological networks to integrate, visualize and analyze genomics data, 2016.
- [44] Li Y, Pearl SA, and Jackson SA. Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis, 2015.
- [45] Hedberg ML, Goh G, Chiosea SI, et al. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. *The Journal of clinical investigation*, 2016. 126(4):1606.
- [46] O'Sullivan B, Huang SH, Siu LL, et al. Deintensification candidate subgroups in human papillomavirus-related oropharyngeal cancer according to minimal risk of distant metastasis. *J Clin Oncol*, 2013. 31(5):543–550.
- [47] O'Sullivan B, Huang SH, Su J, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol*, 2016. 17(4):440–451.
- [48] Vermorken JB and Specenier P. Optimal treatment for recurrent/metastatic head and neck cancer. *Ann Oncol*, 2010. 21 Suppl 7:vii252–61.
- [49] Dahiya K and Dhankhar R. Updated overview of current biomarkers in head and neck carcinoma. *World journal of methodology*, 2016. 6(1):77–86.
- [50] Lohaus F, Linge A, Tinhofer I, et al. HPV16 DNA status is a strong prognosticator of loco-regional control after postoperative radiochemotherapy of locally advanced oropharyngeal carcinoma: results from a multicentre explorative study of the German Cancer Consortium Radiation Oncology Group (. *Radiother Oncol*, 2014. 113(3):317–323.
- [51] Cho H, Yu A, Kim S, Kang J, and Hong SM. Robust Likelihood-Based Survival Modeling with Microarray Data. *2009*, 2009. 29(1):16.
- [52] Niyazi M, Jansen NL, Rottler M, Ganswindt U, and Belka C. Recurrence pattern analysis after re-irradiation with bevacizumab in recurrent malignant glioma patients. *Radiation oncology (London, England)*, 2014. 9:299.
- [53] Fleischmann DF, Unterrainer M, Bartenstein P, Belka C, Albert NL, and Niyazi M. javax.xml.bind.JAXBElement@4debf42a, F-FET PET prior to recurrent high-grade glioma re-irradiation-additional prognostic value of dynamic time-to-peak analysis and early static summation images? *Journal of neuro-oncology*, 2017. 132(2):277–286.
- [54] Ha M and Kim VN. Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, 2014. 15(8):509–524.
- [55] Xi Y, Nakajima G, Gavin E, et al. Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. *RNA (New York, NY)*, 2007. 13(10):1668–1674.
- [56] Cho H, Yu A, Kim S, Kang J, Hong SM, and Others. Robust likelihood-based survival modeling with microarray data. *J Stat Softw*, 2009. 29:1–16.

- [57] Ronckers CM, Erdmann CA, and Land CE. Radiation and breast cancer: a review of current evidence. *Breast cancer research : BCR*, 2005. 7(1):21–32.
- [58] Ibrahim EM, Abouelkhair KM, Kazkaz GA, Elmasri OA, and Al-Foheidi M. Risk of second breast cancer in female Hodgkin's lymphoma survivors: a meta-analysis. *BMC cancer*, 2012. 12:197.
- [59] McGregor H, Land CE, Choi K, et al. Breast cancer incidence among atomic bomb survivors, Hiroshima and Nagasaki, 1950-69. *Journal of the National Cancer Institute*, 1977. 59(3):799–811.
- [60] Prysyzhnyuk A, Gristchenko V, Fedorenko Z, et al. Twenty years after the Chernobyl accident: solid cancer incidence in various groups of the Ukrainian population. *Radiation and environmental biophysics*, 2007. 46(1):43–51.
- [61] Prysyzhnyuk AY, Bazyka DA, Romanenko AY, et al. Quarter of century since the Chornobyl accident: cancer risks in affected groups of population. *Problemy radiatsiinoi medytsyny ta radiobiologii*, 2014. 19:147–169.
- [62] Pukkala E, Kesminiene A, Poliakov S, et al. Breast cancer in Belarus and Ukraine after the Chernobyl accident. *International journal of cancer*, 2006. 119(3):651–658.
- [63] Neuvial P, Hupe P, Brito I, et al. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 2006. 7:264.
- [64] van de Wiel MA and van Wieringen WN. CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer informatics*, 2007. 3:55–63.
- [65] van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, and Ylstra B. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics (Oxford, England)*, 2007. 23(7):892–894.
- [66] Kryuchkov V, Chumak V, Maceika E, et al. Radrue method for reconstruction of external photon doses for Chernobyl liquidators in epidemiological studies. *Health physics*, 2009. 97(4):275–298.
- [67] Unger K, Wienberg J, Riches A, et al. Novel gene rearrangements in transformed breast cells identified by high-resolution breakpoint analysis of chromosomal aberrations. *Endocrine-related cancer*, 2010. 17(1):87–98.
- [68] Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, and Quackenbush J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC bioinformatics*, 2009. 10:409.
- [69] Fabregat A, Sidiropoulos K, Garapati P, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res*, 2016. 44(D1):D481–7.
- [70] Network CGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 2015. 517(7536):576–582.
- [71] Pierre-Jean M, Rigai G, and Neuvial P. Performance evaluation of DNA copy number segmentation methods. *Briefings in bioinformatics*, 2015. 16(4):600–615.
- [72] Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*, 2006. 444(7118):444–454.

-
- [73] Gollin SM. Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: a next generation window to the biology of disease. *Genes, chromosomes & cancer*, 2014. 53(12):972–990.

4 Publication list

4.1 Original works as first or last author

1. Pitea A, Kondofersky I, Sass S, Theis FJ, Mueller NS, **Unger K**. Copy number aberrations from Affymetrix SNP 6.0 genotyping data-how accurate are commonly used prediction approaches? *Briefings in Bioinformatics*. 2018;doi:10.1093/bib/bby096
2. Hess J*, **Unger K***, Maihoefer C, Schuttrumpf L, Wintergerst L, Heider T, Weber P, Marschner S, Braselmann H, Samaga D, Kuger S, Pflugradt U, Baumeister P, Walch A, Woischke C, Kirchner T, Werner M, Werner K, Baumann M, Budach V, Combs SE, Debus J, Grosu AL, Krause M, Linge A, Rodel C, Stuschke M, Zips D, Zitzelsberger HF, Ganswindt U, Henke M, Belka C. A Five-MicroRNA Signature Predicts Survival and Disease Control of Patients with Head and Neck Cancer Negative for HPV-infection. *Clinical cancer research*. epub ahead of print
3. Baumeister P, Hollmann A, Kitz J, Afthonidou A, Simon F, Shakhtour J, Mack B, Kranz G, Libl D, Leu M, Schirmer MA, Canis M, Belka C, Zitzelsberger H, Ganswindt U, Hess J, Jakob M, **Unger K***, Gires O*. High Expression of EpCAM and Sox2 is a Positive Prognosticator of Clinical Outcome for Head and Neck Carcinoma. *Sci Rep*. 2018;8(1):14582.
4. Wilke CM, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Feuchtinger A, Walch AK, Selmansberger M, Braselmann H, Schneider L, Pitea A, Steinhilber J, Fend F, Bosmuller HC, Zitzelsberger H, **Unger K**. Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. *International journal of cancer*. 2018;142(3):573-83.
5. Wilke CM, Braselmann H, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Walch AK, Selmansberger M, Samaga D, Weber P, Schneider L, Fend F, Bosmuller HC, Zitzelsberger H, **Unger K**. A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer. *International journal of cancer*. 2018;143(6):1505-15.
6. Niyazi M, Pitea A, Mittelbronn M, Steinbach J, Sticht C, Zehentmayr F, Piehlmaier D, Zitzelsberger H, Ganswindt U, Rodel C, Lauber K, Belka C, **Unger K**. A 4-miRNA signature predicts the therapeutic outcome of glioblastoma. *Oncotarget*. 2016;7(29):45764-75.
7. Michna A, Braselmann H, Selmansberger M, Dietz A, Hess J, Gomolka M, Hornhardt S, Bluthgen N, Zitzelsberger H, **Unger K**. Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data. *PloS one*. 2016;11(8):e0160791.
8. Selmansberger M, Braselmann H, Hess J, Bogdanova T, Abend M, Tronko M, Brenner A, Zitzelsberger H, **Unger K**. Genomic copy number analysis of Chernobyl papillary thyroid carcinoma in the Ukrainian-American Cohort. *Carcinogenesis*. 2015;36(11):1381-7.
9. Braselmann H, Michna A, Hess J, **Unger K**. CFAssay: statistical analysis of the colony formation assay. *Radiation oncology (London, England)*. 2015;10:223.
10. Hess J, Thomas G, Braselmann H, Bauer V, Bogdanova T, Wienberg J, Zitzelsberger H, **Unger K**. Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(23):9595-600.

11. **Unger K**, Wienberg J, Riches A, Hieber L, Walch A, Brown A, O'Brien PC, Briscoe C, Gray L, Rodriguez E, Jackl G, Knijnenburg J, Tallini G, Ferguson-Smith M, Zitzelsberger H. Novel gene rearrangements in transformed breast cells identified by high-resolution breakpoint analysis of chromosomal aberrations. *Endocrine-related cancer*. 2010;17(1):87-98.
12. **Unger K**, Malisch E, Thomas G, Braselmann H, Walch A, Jackl G, Lewis P, Lengfelder E, Bogdanova T, Wienberg J, Zitzelsberger H. Array CGH demonstrates characteristic aberration signatures in human papillary thyroid carcinomas governed by RET/PTC. *Oncogene*. 2008;27(33):4592-602.
13. **Unger K**, Zurnadzhy L, Walch A, Mall M, Bogdanova T, Braselmann H, Hieber L, Tronko N, Hutzler P, Jeremiah S, Thomas G, Zitzelsberger H. RET rearrangements in post-Chernobyl papillary thyroid carcinomas with a short latency analysed by interphase FISH. *British journal of cancer*. 2006;94(10):1472-7.
14. Rhoden KJ*, **Unger K***, Salvatore G, Yilmaz Y, Vovk V, Chiappetta G, Qumsiyeh MB, Rothstein JL, Fusco A, Santoro M, Zitzelsberger H, Tallini G. RET/papillary thyroid cancer rearrangement in nonneoplastic thyrocytes: follicular cells of Hashimoto's thyroiditis share low-level recombination events with a subset of papillary carcinoma. *The Journal of clinical endocrinology and metabolism*. 2006;91(6):2414-23.
15. **Unger K**, Zitzelsberger H, Salvatore G, Santoro M, Bogdanova T, Braselmann H, Kastner P, Zurnadzhy L, Tronko N, Hutzler P, Thomas G. Heterogeneity in the distribution of RET/PTC rearrangements within individual post-Chernobyl papillary thyroid carcinomas. *The Journal of clinical endocrinology and metabolism*. 2004;89(9):4272-9.

4.2 Original works as co-author

1. Kleemann M, Schneider H, **Unger K**, Bereuther J, Fischer S, Sander P, Marion Schneider E, Fischer-Posovszky P, Riedel CU, Handrick R, Otte K. Induction of apoptosis in ovarian cancer cells by miR-493-3p directly targeting AKT2, STK38L, HMGA2, ETS1 and E2F5. *Cellular and molecular life sciences : CMLS*. 2018;
2. Wintergerst L, Selmansberger M, Maihoefer C, Schuttrumpf L, Walch A, Wilke C, Pitea A, Woischke C, Baumeister P, Kirchner T, Belka C, Ganswindt U, Zitzelsberger H, **Unger K**, Hess J. A prognostic mRNA expression signature of four 16q24.3 genes in radio(chemo)therapy-treated head and neck squamous cell carcinoma (HNSCC). *Mol Oncol*. 2018.
3. Pan M, Schinke H, Luxenburger E, Kranz G, Shakhtour J, Libl D, Huang Y, Gaber A, Pavsic M, Lenarcic B, Kitz J, Jakob M, Schwenk-Zieger S, Canis M, Hess J, **Unger K**, Baumeister P, Gires O. EpCAM ectodomain EpEX is a ligand of EGFR that counteracts EGF-mediated epithelial-mesenchymal transition through modulation of phospho-ERK1/2 in head and neck cancers. *PLoS Biol*. 2018;16(9):e2006624.
4. Orth M, **Unger K**, Schoetz U, Belka C, Lauber K. Taxane-mediated radiosensitization derives from chromosomal missegregation on tripolar mitotic spindles orchestrated by AURKA and TPX2. *Oncogene*. 2018;37(1):52-62.
5. Mueller S, Engleitner T, Maresch R, Zukowska M, Lange S, Kaltenbacher T, Konukiewitz B, Ollinger R, Zwiebel M, Strong A, Yen HY, Banerjee R, Louzada S, Fu B, Seidler B, Gotzfried

- J, Schuck K, Hassan Z, Arbeiter A, Schonhuber N, Klein S, Veltkamp C, Friedrich M, Rad L, Barenboim M, Ziegenhain C, Hess J, Dovey OM, Eser S, Parekh S, Constantino-Casas F, de la Rosa J, Sierra MI, Fraga M, Mayerle J, Kloppel G, Cadinanos J, Liu P, Vassiliou G, Weichert W, Steiger K, Enard W, Schmid RM, Yang F, **Unger K**, Schneider G, Varela I, Bradley A, Saur D, Rad R. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature*. 2018;554(7690):62-8.
6. Kleemann M, Schneider H, **Unger K**, Sander P, Schneider EM, Fischer-Posovszky P, Handrick R, Otte K. MiR-744-5p inducing cell death by directly targeting HNRNPC and NFIX in ovarian cancer cells. *Sci Rep*. 2018;8(1):9020.
 7. Dalke C, Neff F, Bains SK, Bright S, Lord D, Reitmeir P, Rossler U, Samaga D, **Unger K**, Braselmann H, Wagner F, Greiter M, Gomolka M, Hornhardt S, Kunze S, Kempf SJ, Garrett L, Holter SM, Wurst W, Rosemann M, Azimzadeh O, Tapio S, Aubele M, Theis F, Hoeschen C, Slijepcevic P, Kadhim M, Atkinson M, Zitzelsberger H, Kulka U, Graw J. Lifetime study in mice after acute low-dose ionizing radiation: a multifactorial study with special focus on cataract risk. *Radiation and environmental biophysics*. 2018;57(2):99-113.
 8. Yuan D, Huang S, Berger E, Liu L, Gross N, Heinzmann F, Ringelhan M, Connor TO, Stadler M, Meister M, Weber J, Ollinger R, Simonavicius N, Reisinger F, Hartmann D, Meyer R, Reich M, Seehawer M, Leone V, Hochst B, Wohlleber D, Jors S, Prinz M, Spalding D, Protzer U, Luedde T, Terracciano L, Matter M, Longerich T, Knolle P, Ried T, Keitel V, Geisler F, **Unger K**, Cinnamon E, Pikarsky E, Huser N, Davis RJ, Tschaharganeh DF, Rad R, Weber A, Zender L, Haller D, Heikenwalder M. Kupffer Cell-Derived Tnf Triggers Cholangiocellular Tumorigenesis through JNK due to Chronic Mitochondrial Dysfunction and ROS. *Cancer cell*. 2017;31(6):771-89.e6.
 9. Wunderlich R, Ruehle PF, Deloch L, **Unger K**, Hess J, Zitzelsberger H, Lauber K, Frey B, Gaipl US. Interconnection between DNA damage, senescence, inflammation, and cancer. *Frontiers in bioscience (Landmark edition)*. 2017;22:348-69.
 10. Steens J, Zuk M, Benchellal M, Bornemann L, Teichweyde N, Hess J, **Unger K**, Gorgens A, Klump H, Klein D. In Vitro Generation of Vascular Wall-Resident Multipotent Stem Cells of Mesenchymal Nature from Murine Induced Pluripotent Stem Cells. *Stem cell reports*. 2017;8(4):919-32.
 11. Kleemann M, Bereuther J, Fischer S, Marquart K, Hanle S, **Unger K**, Jendrossek V, Riedel CU, Handrick R, Otte K. Investigation on tissue specific effects of pro-apoptotic micro RNAs revealed miR-147b as a potential biomarker in ovarian cancer prognosis. *Oncotarget*. 2017;8(12):18773-91.
 12. Hess J, **Unger K**, Orth M, Schotz U, Schuttrumpf L, Zangen V, Gimenez-Aznar I, Michna A, Schneider L, Stamp R, Selmansberger M, Braselmann H, Hieber L, Drexler GA, Kuger S, Klein D, Jendrossek V, Friedl AA, Belka C, Zitzelsberger H, Lauber K. Genomic amplification of Fanconi anemia complementation group A (FancA) in head and neck squamous cell carcinoma (HNSCC): Cellular mechanisms of radioresistance and clinical relevance. *Cancer letters*. 2017;386:87-99.
 13. Boege Y, Malehmir M, Healy ME, Bettermann K, Lorentzen A, Vucur M, Ahuja AK, Bohm F, Mertens JC, Shimizu Y, Frick L, Remouchamps C, Mutreja K, Kahne T, Sundaravinayagam D, Wolf MJ, Rehrauer H, Koppe C, Speicher T, Padriisa-Altes S, Maire R, Schattenberg JM, Jeong

- JS, Liu L, Zwirner S, Boger R, Huser N, Davis RJ, Mullhaupt B, Moch H, Schulze-Bergkamen H, Clavien PA, Werner S, Borsig L, Luther SA, Jost PJ, Weinlich R, **Unger K**, Behrens A, Hillert L, Dillon C, Di Virgilio M, Wallach D, Dejardin E, Zender L, Naumann M, Walczak H, Green DR, Lopes M, Lavrik I, Luedde T, Heikenwalder M, Weber A. A Dual Role of Caspase-8 in Triggering and Sensing Proliferation-Associated DNA Damage, a Key Determinant of Liver Cancer Development. *Cancer cell*. 2017;32(3):342-59.e10.
14. Penterling C, Drexler GA, Bohland C, Stamp R, Wilke C, Braselmann H, Caldwell RB, Reindl J, Girst S, Greubel C, Siebenwirth C, Mansour WY, Borgmann K, Dollinger G, **Unger K**, Friedl AA. Depletion of Histone Demethylase Jarid1A Resulting in Histone Hyperacetylation and Radiation Sensitivity Does Not Affect DNA Double-Strand Break Repair. *PLoS one*. 2016;11(6):e0156599.
 15. Michna A, Schotz U, Selmansberger M, Zitzelsberger H, Lauber K, **Unger K**, Hess J. Transcriptional 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. *Radiation oncology (London, England)*. 2016;11:94.
 16. Mathieson W, Marcon N, Antunes L, Ashford DA, Betsou F, Frاسquilho SG, Kofanova OA, McKay SC, Pericleous S, Smith C, **Unger KM**, Zeller C, Thomas GA. A Critical Evaluation of the PAX-gene Tissue Fixation System: Morphology, Immunohistochemistry, Molecular Biology, and Proteomics. *American journal of clinical pathology*. 2016;146(1):25-40.
 17. Maresch R, Mueller S, Veltkamp C, Ollinger R, Friedrich M, Heid I, Steiger K, Weber J, Engleitner T, Barenboim M, Klein S, Louzada S, Banerjee R, Strong A, Stauber T, Gross N, Geumann U, Lange S, Ringelhan M, Varela I, **Unger K**, Yang F, Schmid RM, Vassiliou GS, Braren R, Schneider G, Heikenwalder M, Bradley A, Saur D, Rad R. Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISPR/Cas9 delivery in mice. *Nature communications*. 2016;7:10770.
 18. Klein D, Schmetter A, Imsak R, Wirsdorfer F, **Unger K**, Jastrow H, Stuschke M, Jendrossek V. Therapy with Multipotent Mesenchymal Stromal Cells Protects Lungs from Radiation-Induced Injury and Reduces the Risk of Lung Metastasis. *Antioxidants & redox signaling*. 2016;24(2):53-69.
 19. Kaiser JC, Meckbach R, Eidemuller M, Selmansberger M, **Unger K**, Shpak V, Blettner M, Zitzelsberger H, Jacob P. Integration of a radiation biomarker into modeling of thyroid carcinogenesis and post-Chernobyl risk assessment. *Carcinogenesis*. 2016;37(12):1152-60.
 20. Hauptmann M, Haghdoost S, Gomolka M, Sarioglu H, Ueffing M, Dietz A, Kulka U, **Unger K**, Babini G, Harms-Ringdahl M, Ottolenghi A, Hornhardt S. Differential Response and Priming Dose Effect on the Proteome of Human Fibroblast and Stem Cells Induced by Exposure to Low Doses of Ionizing Radiation. *Radiation research*. 2016;185(3):299-312.
 21. Handkiewicz-Junak D, Swierniak M, Rusinek D, Oczko-Wojciechowska M, Dom G, Maenhaut C, **Unger K**, Detours V, Bogdanova T, Thomas G, Likhtarov I, Jaksik R, Kowalska M, Chmielik E, Jarzab M, Swierniak A, Jarzab B. Gene signature of the post-Chernobyl papillary thyroid cancer. *European journal of nuclear medicine and molecular imaging*. 2016;43(7):1267-77.
 22. Endig J, Buitrago-Molina LE, Marhenke S, Reisinger F, Saborowski A, Schutt J, Limbourg F, Konecke C, Schreder A, Michael A, Misslitz AC, Healy ME, Geffers R, Clavel T, Haller D, **Unger**

- K, Finegold M, Weber A, Manns MP, Longerich T, Heikenwalder M, Vogel A. Dual Role of the Adaptive Immune System in Liver Injury and Hepatocellular Carcinoma Development. *Cancer cell*. 2016;30(2):308-23.
23. Weber J, Ollinger R, Friedrich M, Ehmer U, Barenboim M, Steiger K, Heid I, Mueller S, Maresch R, Engleitner T, Gross N, Geumann U, Fu B, Segler A, Yuan D, Lange S, Strong A, de la Rosa J, Esposito I, Liu P, Cadinanos J, Vassiliou GS, Schmid RM, Schneider G, **Unger K**, Yang F, Braren R, Heikenwalder M, Varela I, Saur D, Bradley A, Rad R. CRISPR/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(45):13982-7.
24. Summerer I, **Unger K**, Braselmann H, Schuettrumpf L, Maihoefer C, Baumeister P, Kirchner T, Niyazi M, Sage E, Specht HM, Multhoff G, Moertl S, Belka C, Zitzelsberger H. Circulating microRNAs as prognostic therapy biomarkers in head and neck cancer patients. *British journal of cancer*. 2015;113(1):76-82.
25. Summerer I, Hess J, Pitea A, **Unger K**, Hieber L, Selmansberger M, Lauber K, Zitzelsberger H. Integrative analysis of the microRNA-mRNA response to radiochemotherapy in primary head and neck squamous cell carcinoma cells. *BMC genomics*. 2015;16:654.
26. Selmansberger M, Kaiser JC, Hess J, Guthlin D, Likhtarev I, Shpak V, Tronko M, Brenner A, Abend M, Blettner M, **Unger K**, Jacob P, Zitzelsberger H. Dose-dependent expression of CLIP2 in post-Chernobyl papillary thyroid carcinomas. *Carcinogenesis*. 2015;36(7):748-56.
27. Selmansberger M, Feuchtinger A, Zurnadzhy L, Michna A, Kaiser JC, Abend M, Brenner A, Bogdanova T, Walch A, **Unger K**, Zitzelsberger H, Hess J. CLIP2 as radiation biomarker in papillary thyroid carcinoma. *Oncogene*. 2015;34(30):3917-25.
28. Sass S, Pitea A, **Unger K**, Hess J, Mueller NS, Theis FJ. MicroRNA-Target Network Inference and Local Network Enrichment Analysis Identify Two microRNA Clusters with Distinct Functions in Head and Neck Squamous Cell Carcinoma. *International journal of molecular sciences*. 2015;16(12):30204-22.
29. Mu X, Espanol-Suner R, Mederacke I, Affo S, Manco R, Sempoux C, Lemaigre FP, Adili A, Yuan D, Weber A, **Unger K**, Heikenwalder M, Leclercq IA, Schwabe RF. Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. *The Journal of clinical investigation*. 2015;125(10):3891-903.
30. Gross C, Steiger K, Sayyed S, Heid I, Feuchtinger A, Walch A, Hess J, **Unger K**, Zitzelsberger H, Settles M, Schlitter AM, Dworniczak J, Altomonte J, Ebert O, Schwaiger M, Rummeny E, Steingotter A, Esposito I, Braren R. Model Matters: Differences in Orthotopic Rat Hepatocellular Carcinoma Physiology Determine Therapy Response to Sorafenib. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2015;21(19):4440-50.
31. Finkin S, Yuan D, Stein I, Taniguchi K, Weber A, **Unger K**, Browning JL, Goossens N, Nakagawa S, Gunasekaran G, Schwartz ME, Kobayashi M, Kumada H, Berger M, Pappo O, Rajewsky K, Hoshida Y, Karin M, Heikenwalder M, Ben-Neriah Y, Pikarsky E. Ectopic lymphoid structures function as microniches for tumor progenitor cells in hepatocellular carcinoma. *Nature immunology*. 2015;16(12):1235-44.

32. Eke I, Zscheppang K, Dickreuter E, Hickmann L, Mazzeo E, **Unger K**, Krause M, Cordes N. Simultaneous beta1 integrin-EGFR targeting and radiosensitization of human head and neck cancer. *Journal of the National Cancer Institute*. 2015;107(2).
33. Babini G, Bellinzona VE, Morini J, Baiocco G, Mariotti L, **Unger K**, Ottolenghi A. Mechanisms of the induction of apoptosis mediated by radiation-induced cytokine release. *Radiation protection dosimetry*. 2015;166(1-4):165-9.
34. Abend M, Azizova T, Muller K, Dorr H, Doucha-Senf S, Kreppel H, Rusinova G, Glazkova I, Vyazovskaya N, **Unger K**, Braselmann H, Meineke V. Association of radiation-induced genes with noncancer chronic diseases in Mayak workers occupationally exposed to prolonged radiation. *Radiation research*. 2015;183(3):249-61.
35. Wolf MJ, Adili A, Piotrowitz K, Abdullah Z, Boege Y, Stemmer K, Ringelhan M, Simonavicius N, Egger M, Wohlleber D, Lorentzen A, Einer C, Schulz S, Clavel T, Protzer U, Thiele C, Zischka H, Moch H, Tschop M, Tumanov AV, Haller D, **Unger K**, Karin M, Kopf M, Knolle P, Weber A, Heikenwalder M. Metabolic activation of intrahepatic CD8+ T cells and NKT cells causes nonalcoholic steatohepatitis and liver cancer via cross-talk with hepatocytes. *Cancer cell*. 2014;26(4):549-64.
36. Di Maro G, Salerno P, **Unger K**, Orlandella FM, Monaco M, Chiappetta G, Thomas G, Oczko-Wojciechowska M, Masullo M, Jarzab B, Santoro M, Salvatore G. Anterior gradient protein 2 promotes survival, migration and invasion of papillary thyroid carcinoma cells. *Molecular cancer*. 2014;13:160.
37. Abend M, Azizova T, Muller K, Dorr H, Senf S, Kreppel H, Rusinova G, Glazkova I, Vyazovskaya N, **Unger K**, Meineke V. Independent validation of candidate genes identified after a whole genome screening on Mayak workers exposed to prolonged occupational radiation. *Radiation research*. 2014;182(3):299-309.
38. Abend M, Azizova T, Muller K, Dorr H, Senf S, Kreppel H, Rusinova G, Glazkova I, Vyazovskaya N, Schmidl D, **Unger K**, Meineke V. Gene expression analysis in Mayak workers with prolonged occupational radiation exposure. *Health physics*. 2014;106(6):664-76.
39. Vucur M, Reisinger F, Gautheron J, Janssen J, Roderburg C, Cardenas DV, Kreggenwinkel K, Koppe C, Hammerich L, Hakem R, **Unger K**, Weber A, Gassler N, Luedde M, Frey N, Neumann UP, Tacke F, Trautwein C, Heikenwalder M, Luedde T. RIP3 inhibits inflammatory hepatocarcinogenesis but promotes cholestasis by controlling caspase-8- and JNK-dependent compensatory cell proliferation. *Cell reports*. 2013;4(4):776-90.
40. Summerer I, Niyazi M, **Unger K**, Pitea A, Zangen V, Hess J, Atkinson MJ, Belka C, Moertl S, Zitzelsberger H. Changes in circulating microRNAs after radiochemotherapy in head and neck cancer patients. *Radiation oncology (London, England)*. 2013;8:296.
41. van Wieringen WN, **Unger K**, Leday GG, Krijgsman O, de Menezes RX, Ylstra B, van de Wiel MA. Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. *BMC bioinformatics*. 2012;13:80.
42. Thomas G, **Unger K**, Krznaric M, Galpine A, Bethel J, Tomlinson C, Woodbridge M, Butcher S. The chernobyl tissue bank - a repository for biomaterial and data used in integrative and systems biology modeling the human response to radiation. *Genes*. 2012;3(2):278-90.

43. Nipp M, Elsner M, Balluff B, Meding S, Sarioglu H, Ueffing M, Rauser S, **Unger K**, Hofler H, Walch A, Zitzelsberger H. S100-A10, thioredoxin, and S100-A6 as biomarkers of papillary thyroid carcinoma with lymph node metastasis identified by MALDI imaging. *Journal of molecular medicine (Berlin, Germany)*. 2012;90(2):163-74.
44. Heiliger KJ, Hess J, Vitagliano D, Salerno P, Braselmann H, Salvatore G, Ugolini C, Summerer I, Bogdanova T, **Unger K**, Thomas G, Santoro M, Zitzelsberger H. Novel candidate genes of thyroid tumorigenesis identified in Trk-T1 transgenic mice. *Endocrine-related cancer*. 2012;19(3):409-21.
45. Dom G, Tarabichi M, **Unger K**, Thomas G, Oczko-Wojciechowska M, Bogdanova T, Jarzab B, Dumont JE, Detours V, Maenhaut C. A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas. *British journal of cancer*. 2012;107(6):994-1000.
46. Baumhoer D, Zillmer S, **Unger K**, Rosemann M, Atkinson MJ, Irmeler M, Beckers J, Siggelkow H, von Luettichau I, Jundt G, Smida J, Nathrath M. MicroRNA profiling with correlation to gene expression revealed the oncogenic miR-17-92 cluster to be up-regulated in osteosarcoma. *Cancer genetics*. 2012;205(5):212-9.
47. McKay SC, **Unger K**, Pericleous S, Stamp G, Thomas G, Hutchins RR, Spalding DR. Array comparative genomic hybridization identifies novel potential therapeutic targets in cholangiocarcinoma. *HPB: the official journal of the International Hepato Pancreato Biliary Association*. 2011;13(5):309-19.
48. Flavin R, Jackl G, Finn S, Smyth P, Ring M, O'Regan E, Cahill S, **Unger K**, Denning K, Jinghuan L, Aherne S, Tallini G, Gaffney E, O'Leary JJ, Zitzelsberger H, Sheils O. RET/PTC rearrangement occurring in primary peritoneal carcinoma. *International journal of surgical pathology*. 2009;17(3):187-97.
49. Bauer VL, Braselmann H, Henke M, Mattern D, Walch A, **Unger K**, Baudis M, Lassmann S, Huber R, Wienberg J, Werner M, Zitzelsberger HF. Chromosomal changes characterize head and neck cancer with poor prognosis. *Journal of molecular medicine (Berlin, Germany)*. 2008;86(12):1353-65.
50. Fiegler H, Geigl JB, Langer S, Rigler D, Porter K, **Unger K**, Carter NP, Speicher MR. High resolution array-CGH analysis of single cells. *Nucleic acids research*. 2007;35(3):e15.
51. Zitzelsberger H, Hieber L, Richter H, **Unger K**, Briscoe CV, Peddie C, Riches A. Gene amplification of atypical PKC-binding PARD3 in radiation-transformed neoplastic retinal pigment epithelial cell lines. *Genes, chromosomes & cancer*. 2004;40(1):55-9.

4.3 Review articles

1. **Unger K**, Heikenwalder M. Analysis of chromosomal aberrations in murine HCC. *Methods in molecular biology (Clifton, NJ)*. 2014;1193:213-26.
2. **Unger K**. Integrative radiation systems biology. *Radiation oncology (London, England)*. 2014;9:21.
3. Zitzelsberger H, **Unger K**. DNA copy number alterations in radiation-induced thyroid cancer. *Clinical oncology (Royal College of Radiologists (Great Britain))*. 2011;23(4):289-96.

4. Thomas GA, Bethel JA, Galpine A, Mathieson W, Krznaric M, **Unger K**. Integrating research on thyroid cancer after Chernobyl—the Chernobyl Tissue Bank. *Clinical oncology (Royal College of Radiologists (Great Britain))*. 2011;23(4):276-81.
5. Zitzelsberger H, Thomas G, **Unger K**. Chromosomal aberrations in thyroid follicular-cell neoplasia: in the search of novel oncogenes and tumour suppressor genes. *Molecular and cellular endocrinology*. 2010;321(1):57-66.
6. Zitzelsberger H, Bauer V, Thomas G, **Unger K**. Molecular rearrangements in papillary thyroid carcinomas. *Clinica chimica acta; international journal of clinical chemistry*. 2010;411(5-6):301-8.

5 Acknowledgements

I would like to say very special thanks to Prof. Dr. Horst Zitzelsberger for his scientific and professional support over so many years. Horst, many thanks for your trust and great support!

Special thanks also to the whole team at ZYTO and particularly to Dr. Julia Heß-Rieger with whom I had the pleasure to team up and to build up a great and fruitful cooperation and friendship.

Great thanks to the clinical partners at the LMU Radiation Oncology Department. In the first place Prof. Dr. Claus Belka as a great scientific and academic mentor. And PD Dr. Maximilian Niyazi for our fantastic collaboration.

Greatest thanks go to my family and my parents. Dear Carolin and Marlena, you are providing the uttermost important backbone of my professional life without nothing would be possible as it is!

6 Reprints of the underlying original works

- Hess J*, **Unger K***, Maihoefer C, Schüttrumpf L, Schneider L, Heider T, Weber P, Marschner S, Braselmann K, Kuger S, Pflugradt U, Baumeister P, Walch A, Woischke C, Kirchner T, Werner M, Werner K, Baumann M, Budach V, Combs SE, Debus J, Grosu AL, Krause M, Rödel C, Stuschke M, Zips D, Zitzelsberger H, Ganswindt U, Henke M, Belka C A 5-MicroRNA signature predicts Survival and Disease Control of Patients with Head and Neck Cancer negative for HPV-infection. (*co-first authorship)
- Niyazi M, Pitea A, Mittelbronn M, Steinbach J, Sticht C, Zehentmayr F, Piehlmaier D, Zitzelsberger H, Ganswindt U, Rödel C, Lauber K, Belka C, **Unger K** A 4-miRNA signature predicts the therapeutic outcome of glioblastoma. **Oncotarget**. 2016 Jul 19;7(29):45764-45775
- Wilke CM, Braselmann H, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Walch AK, Selmansberger M, Samaga D, Weber P, Schneider L, Fend F, Bösmüller HC, Zitzelsberger H, **Unger K** A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer. **Int J Cancer**. 2018 Sep 15;143(6):1505-1515.
- Wilke CM, Hess J, Klymenko SV, Chumak VV, Zakhartseva LM, Bakhanova EV, Feuchtinger A, Walch AK, Selmansberger M, Braselmann H, Schneider L, Pitea A, Steinhilber J, Fend F, Bösmüller HC, Zitzelsberger H, **Unger K** Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. **Int J Cancer**. 2018 Feb 1;142(3):573-583
- Michna A, Braselmann H, Selmansberger M, Dietz A, Hess J, Gomolka M, Hornhardt S, Blüthgen N, Zitzelsberger H, **Unger K**. Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data. **PLoS One**. 2016 Aug 9;11(8)
- Pitea A, Kondofersky I, Sass S, Theis F, Mueller N, **Unger K**. Copy number aberrations from Affymetrix SNP 6.0 genotyping data - how accurate are the commonly used prediction approaches? **Brief Bioinform**. 2018 accepted

Original article

A Five-MicroRNA Signature Predicts Survival and Disease Control of Patients with Head and Neck Cancer Negative for HPV-infection

Julia Hess^{1,2,*}, Kristian Unger^{1,2,*}, Cornelius Maihofer^{2,3}, Lars Schüttrumpf^{2,3}, Ludmila Wintergerst^{1,2}, Theresa Heider^{1,2}, Peter Weber^{1,2,3,4}, Sebastian Marschner^{2,3}, Herbert Braselmann^{1,2}, Daniel Samaga^{1,2}, Sebastian Kuger¹, Ulrike Pflugrad^{2,3}, Philipp Baumeister^{2,6}, Axel Walch⁷, Christine Woischke⁸, Thomas Kirchner^{8,9}, Martin Werner^{10,11,12}, Kristin Werner^{10,11,12}, Michael Baumann¹³, Volker Budach^{14,15}, Stephanie E. Combs^{9,16,17}, Jürgen Debus^{18,19}, Anca-Ligia Grosu^{12,20}, Mechthild Krause^{13,21,22,23}, Annett Linge^{13,21,22,23,24}, Claus Rödel^{25,26}, Martin Stuschke^{27,28}, Daniel Zips^{29,30}, Horst Zitzelsberger^{1,2,3}, Ute Ganswindt^{2,3,31}, Michael Henke^{12,20,+}, Claus Belka^{2,3,9+}

*These authors contributed equally as first authors

+These authors contributed equally as senior authors

Affiliations list:

1. Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
2. Clinical Cooperation Group “Personalized Radiotherapy in Head and Neck Cancer”, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
3. Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany
4. Institute for Diabetes and Cancer (IDC), Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
5. Joint Heidelberg-IDC Translational Diabetes Program, Heidelberg University Hospital, Heidelberg, Germany
6. Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital, Ludwig-Maximilians-University of Munich, Munich, Germany
7. Research Unit Analytical Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
8. Institute of Pathology, Faculty of Medicine, Ludwig-Maximilians-University of Munich, Munich, Germany
9. German Cancer Consortium (DKTK), Partner Site Munich, and German Cancer Research Center (DKFZ), Heidelberg, Germany
10. Institute for Surgical Pathology, Medical Center - University of Freiburg, Freiburg, Germany

11. Faculty of Medicine, University of Freiburg, Freiburg, Germany
12. German Cancer Consortium (DKTK), Partner Site Freiburg, and German Cancer Research Center (DKFZ), Heidelberg, Germany
13. German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany
14. Department of Radiooncology and Radiotherapy, Charité University Hospital Berlin, Berlin, Germany
15. German Cancer Consortium (DKTK), Partner Site Berlin, and German Cancer Research Center (DKFZ), Heidelberg, Germany
16. Department of Radiation Oncology, Klinikum rechts der Isar, Technische Universität München, Munich, Germany
17. Institute of Innovative Radiotherapy (iRT), Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
18. Department of Radiation Oncology, Heidelberg Ion Therapy Center (HIT), University of Heidelberg, Heidelberg, Germany
19. German Cancer Consortium (DKTK), Partner Site Heidelberg, and German Cancer Research Center (DKFZ), Heidelberg, Germany
20. Department of Radiation Oncology, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany
21. Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
22. OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany
23. National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany
24. Helmholtz-Zentrum Dresden – Rossendorf, Institute of Radiooncology – OncoRay Dresden, Germany
25. Department of Radiotherapy and Oncology, Goethe-University Frankfurt, Frankfurt, Germany
26. German Cancer Consortium (DKTK), Partner Site Frankfurt, and German Cancer Research Center (DKFZ), Heidelberg, Germany
27. Department of Radiotherapy, Medical Faculty, University of Duisburg-Essen, Essen,

Germany

28. German Cancer Consortium (DKTK), Partner Site Essen, and German Cancer Research Center (DKFZ), Heidelberg, Germany

29. Department of Radiation Oncology, Faculty of Medicine and University Hospital Tübingen, Eberhard Karls University Tübingen, Tübingen, Germany

30. German Cancer Consortium (DKTK), Partner Site Tübingen, and German Cancer Research Center (DKFZ), Heidelberg, Germany

31. Department of Therapeutic Radiology and Oncology, Innsbruck Medical University, Innsbruck, Austria

Running title:

A prognostic 5-miRNA-signature in HPV-negative HNSCC

Financial support:

This study was supported by the German Federal Ministry of Education and Research project ZiSstrans (02NUK047A, 02NUK047C, 02NUK047F) and the Joint Funding Grant within the German Consortium for Translational Cancer Research (DKTK) awarded to the DKTK-ROG (Radiation Oncology Group). The DKTK is funded as one of the National German Health Centers by the German Federal Ministry of Education and Research.

Corresponding author:

Dr. Julia Hess
Research Unit Radiation Cytogenetics
Helmholtz Zentrum München
Ingolstädter Landstrasse 1
85764 Neuherberg
Germany
+49-89-3187-3517
julia.hess@helmholtz-muenchen.de

Key words: Head and neck cancer, prognosis, recurrence, miRNA, risk groups

Disclosure Statement:

The authors have declared no conflicts of interest.

Statement of translational relevance (word count: 150 words)

HPV-negative HNSCC cancer is currently treated with a set of standard-of-care therapeutic approaches which in total result in approx. 50% overall survival for locally advanced HNSCC demonstrating that substantial subgroups are not likely to profit from state-of-the-art therapy. The most relevant clinical event limiting success of HNSCC therapy is recurrence of the disease after surgical tumor resection followed by radio(chemo)therapeutic treatment. The presented HNSCC HPV-negative five-miRNA-signature predicts the risk for recurrence in HNSCC and allows, in combination with the clinically established risk factors, the definition of four prognostically distinct groups. This provides the first prerequisite for the consideration of personalized treatment approaches in HPV-negative HNSCC. Possible personalized treatment options include consideration of adjusting therapy intensity according to the overall risk for therapy failure in the first line. Further, and most importantly, it represents the basis for a more focused search for molecular therapeutic targets improving therapy success for appropriate patients.

Abstract (word count: 249 words)

Purpose:

HPV-negative head and neck squamous cell carcinoma (HNSCC) associates with unfavorable prognosis while independent prognostic markers remain to be defined.

Experimental Design:

We retrospectively performed miRNA expression profiling. Patients were operated for locally advanced HPV-negative HNSCC and had received radiochemotherapy in eight different hospitals (DKTK-ROG; $n=85$). Selection fulfilled comparable demographic, treatment and follow-up characteristics. Findings were validated in an independent single-center patient sample (LMU-KKG; $n=77$). A prognostic miRNA-signature was developed for freedom from recurrence and tested for other endpoints. Recursive-partitioning analysis was performed on the miRNA-signature, tumor and nodal stage, and extracapsular nodal spread. Technical validation used qRT-PCR. A miRNA-mRNA target network was generated and analyzed.

Results:

For DKTK-ROG and LMU-KKG patients, the median follow-up was 5.1 and 5.3 years, the 5-year freedom from recurrence rate was 63.5% and 75.3%, respectively. A five-miRNA-signature (hsa-let-7g-3p, hsa-miR-6508-5p, hsa-miR-210-5p, hsa-miR-4306 and hsa-miR-7161-3p) predicted freedom from recurrence in DKTK-ROG (HR 4.42, 95% CI 1.98–9.88, $P<0.001$), which was confirmed in LMU-KKG (HR 4.24, 95% CI 1.40–12.81, $P=0.005$). The signature also predicted overall survival (HR 3.03, 95% CI 1.50–6.12, $P=0.001$), recurrence-free survival (HR 3.16, 95% CI 1.65–6.04, $P<0.001$) and disease-specific survival (HR 5.12, 95% CI 1.88–13.92, $P<0.001$), all confirmed in LMU-KKG data. Adjustment for relevant covariates maintained the miRNA-signature predicting all endpoints. Recursive-partitioning analysis of both samples combined classified patients into low ($n=17$), low-intermediate ($n=80$), high-intermediate ($n=48$) or high risk ($n=17$) for recurrence ($P<0.001$).

Conclusions:

The five-miRNA-signature is a strong and independent prognostic factor for disease recurrence and survival of patients with HPV-negative HNSCC.

Introduction

Prognosis of patients with locally advanced head and neck squamous cell carcinoma (HNSCC) generally remains poor. Whereas patients with high-risk human papillomavirus (HPV) associated HNSCC have a considerably more favorable outcome, HPV-negative patients still have to expect limited disease control and survival (1,2). From the biologic perspective, intrinsic resistance of tumor cells to radiochemotherapy or therapy failure caused by metastatic spread are possible underlying factors. Consequently, research aims at altering radiation dose and fractionation or - more recently - at the additional administration of targeted drugs and/or immune checkpoint inhibitors (3,4). However, biomarkers to predict which patients potentially would profit from these approaches are missing.

Complex and heterogeneous genomic aberrations and mutation patterns molecularly control initiation and progression of HNSCC (5-7). MicroRNAs (miRNAs), involved in posttranscriptional regulation, have been shown to be highly deregulated in most cancers and might well be of prognostic relevance (8,9). In HNSCC aberrantly expressed miRNAs were described (10-12). However, so far no study has investigated the prognostic role of miRNAs by comprehensive miRNA-profiling in well-characterized HPV-negative HNSCC cohorts.

Here we analyzed miRNA expression profiles in cancer tissue of locally advanced HNSCC ($n=162$). We hypothesized that we can develop a miRNA-based molecular signature, which allows to stratify HPV-negative HNSCC patients according to risk of recurrence following adjuvant radio(chemo)therapy.

Materials and Methods

Patient specimens and study design

In the present study, we analyzed two independent samples of HNSCC patients who had undergone surgical resection followed by adjuvant radio(chemo)therapy: the DKTK-ROG (German Consortium for Translational Cancer Research, Radiation Oncology Group) and the LMU-KKG (Ludwig-Maximilians-University of Munich, Clinical Cooperation Group “Personalized Radiotherapy in Head and Neck Cancer“) samples. For both of which, clinical data and treatment-naïve patient tissue specimens were collected retrospectively. All patients were diagnosed with histologically proven HNSCC of the hypopharynx, oropharynx or the oral cavity. Only HPV-negative HNSCC were included (Supplementary Methods). Ethical approval (EA) for this retrospective study, carried out in accordance with the Declaration of Helsinki, was obtained by the ethics committees of all DKTK-ROG partners including the LMU (EA 312-12, 448-13, 17-116). Tumor stage was assessed using the UICC TNM Classification of Malignant Tumors, 7th edition.

The multicentric study sample DKTK-ROG originally included 221 HNSCC patients treated at one of the eight different DKTK partner sites (13). This study reports on 85 out of 143 patients with HPV-negative tumors who were treated between 2005 and 2011. 58 patients had to be omitted due to insufficient tumor material. All patients received postoperative radiotherapy covering the previous tumor region and regional lymph nodes with concurrent cisplatin(CDDP)-based chemotherapy according to standard protocols. Inclusion criteria were positive microscopic resection margins and/or extracapsular extension (ECE) of lymph nodes and/or tumor stage pT4 and/or or more than three positive lymph nodes. The median overall treatment time was 44 days (interquartile range IQR 43-46 days). Adjuvant radiotherapy including elective irradiation of cervical lymph nodes was applied with a median dose of 50 Gy (median dose 2 Gy/fraction) and a boost to the former tumor region and to microscopic disease (if any) to a median dose of 66 Gy (median dose 2 Gy/fraction). Cisplatin was applied weekly with a median cumulative dose of 200 mg/m² body surface area (BSA) (range 100-300 mg/m² BSA).

The monocentric study sample LMU-KKG included originally all HNSCC patients with at least UICC TNM stage III or close/positive microscopic resection margins (resection margins were considered “close margin” when declared R0, but less than 5 mm by the local pathologist) who were treated with adjuvant radiotherapy between 06/2008 and 01/2013 at the LMU Department of Radiation Oncology (14). The median overall treatment time was 45 days (IQR 43-47 days)

with five fractions per week. A median radiation dose of 64 Gy (median dose 2 Gy/fraction) was applied to the former tumor bed or regions of ECE, elective lymph node regions have been covered according to tumor stage and localization with a median dose of 50 Gy (median dose 2 Gy/fraction), 56 Gy (median dose 2 Gy/fraction) were applied to involved lymph node regions. In the case of close/positive microscopic resection margins and/or ECE, patients received concurrent chemotherapy. The majority (76 %) of the patients received CDDP/5-fluorouracil (5-FU) (CDDP: 20 mg/m² BSA day 1–5/29–33; 5-FU: 600 mg/m² BSA day 1–5/29–33). In selected cases, Mitomycin C (MMC) or 5-FU/MMC replaced platin-based chemotherapy. This study reports on the HPV-negative tumor subset ($n=77$) of all patients with available tumor tissue specimens ($n=115$).

After histopathological review of haematoxylin and eosin stained tissue sections from available blocks with formalin-fixed and paraffin-embedded (FFPE) tumor tissue by a pathologist (DKTK-ROG: KW; LMU-KKG: CW/AW), the tumor area was annotated. If necessary, microdissection was performed prior nucleic acids extraction in order to ensure a tumor cellularity (i.e., the percentage of tumor cells in analyzed tissue) of at least 60% (DKTK-ROG: median 60%, IQR 60-70%; LMU-KKG: median 70%, IQR 70-80%).

Procedures

Total RNA, including the small RNA fraction, was extracted using the Qiagen miRNeasy FFPE- (DKTK-ROG) or the AllPrep DNA/RNA FFPE-Kit (LMU-KKG) according to the manufacturer's protocols (Qiagen, Hilden, Germany). Isolated RNA was quantified with the Qubit-Fluorometer and integrity of small RNAs was assessed (Supplementary Methods).

miRNA expression was profiled using SurePrint G3 8x60K Human miRNA Microarrays (AMADID 70156; Agilent Technologies, Santa Clara, CA, USA) representing 2,549 human miRNAs (content sourced from miRBase database, Release 21.0; Supplementary Methods). Microarray raw data were uploaded to the publicly available database ArrayExpress (accession no. E-MTAB-5793). miRNA expression microarray profiling resulted in a data set of 162 HNSCC samples (DKTK-ROG: $n=85$; LMU-KKG: $n=77$).

Data analysis was performed using the R statistical software (version 3.3.1) in combination with R-Bioconductor/CRAN packages (Supplementary Methods)(15).

For the purpose of building a Cox proportional hazards model predicting disease recurrence in combination with miRNA expression, we used a robust likelihood-based survival modelling approach deploying an iterative forward-selection algorithm implemented in the R package rbsurv (16). We recently built a miRNA-signature predicting outcome in glioblastoma using the same approach (Supplementary Methods)(17).

Experimentally validated miRNA-target genes of the signature miRNAs were obtained from the miRTarBase database (Release 6.0). The Cytoscape software (version 3.2.1) with the Reactome FI plugin (version 4.0.0) was used to generate a miRNA-mRNA target regulatory network and to conduct pathway enrichment analysis of the target genes. Pathways with P -values < 0.05 were considered as significantly enriched with target genes (18).

For technical validation of microarray data quantitative real-time RT-PCR (qRT-PCR) analysis was performed (Supplementary Methods).

Clinical endpoints and statistical analysis

As the main objective of the study was the identification of a miRNA-signature that allows separation of patients according to risk of recurrence, the primary endpoint was freedom from recurrence. Freedom from recurrence was defined as the time (days) from the start of radiotherapy treatment to the time of the first observation of confirmed locoregional or distant recurrence. Data for recurrence-free patients were right-censored either at the date of death or last follow-up visit. Additional endpoints included were recurrence-free survival, overall survival,

disease-specific survival, disease-unspecific survival, distant control, and locoregional control. We calculated recurrence-free survival (days) from the date of radiotherapy treatment start to the first observation of locoregional/distant recurrence or death due to any cause; overall survival from the date of radiotherapy treatment start to the date of death from any cause; disease-specific survival from the date of radiotherapy treatment start to the date of tumor related death; non-tumor related survival from the date of radiotherapy treatment start to the date of non-tumor related death; distant control from the date of radiotherapy treatment start to the date of distant recurrence; locoregional control from the date of radiotherapy treatment start to the date of local recurrence. In the absence of an event, patients were censored at the date of the last follow-up visit (or the date of death).

Kaplan-Meier curves were compared for statistical difference using the log-rank test using the R-package survival. Median time-to-event estimates and Hazard ratios (HR) with 95% confidence intervals (CI) were determined. Univariate Cox proportional hazard analysis was performed to evaluate the association of clinicopathological variables with outcome (Supplementary Methods). We used multivariate Cox proportional hazards analysis to assess the prognostic value of the identified miRNA-signature after adjustment for other prognostic clinical parameters as covariates.

The clinical endpoint prediction performance of the five-miRNA-signature and clinicopathological variables in terms of sensitivity and specificity, represented by the corresponding areas under the curve (AUCs), was determined for follow-up times from 1 to 5 years (Supplementary Methods).

Recursive partitioning analysis (RPA) for the generation of a decision tree considering the clinical parameters ECE status, TNM T stage, TNM N stage and resection margin status with or without the five-miRNA-signature defined risk groups was conducted using the R-package rpart (Supplementary Methods).

Results

The clinicopathological characteristics of the HNSCC patients included in our study (median follow-up: DTKK-ROG 5.1 years, IQR 3.7-5.6; LMU-KKG 5.3 years, IQR 4.4-6.4) are listed in Table 1. Compared to the DTKK-ROG sample, which exclusively contained patients treated by postoperative radiotherapy with concurrent cisplatin-based chemotherapy, only 63.6% of the LMU-KKG sample received concurrent radiochemotherapy. Accordingly, the LMU-KKG sample contained fewer patients with UICC TNM stage IV, advanced nodal stage, ECE or positive microscopic resection margins. 31.5% of all patients (51/162) developed disease recurrence within the observed follow-up time while the two samples did not differ with regard to the endpoints freedom from recurrence and recurrence-free survival (Figure S1). The 5-year freedom from recurrence rate was 63.5% and 75.3% for DTKK-ROG and LMU-KKG patients, respectively.

The miRNA expression profiling of 162 tumor specimens identified 1,031 expressed miRNAs. After univariate preselection 524 miRNAs remained for feature selection using a robust likelihood-based survival modeling forward-selection approach (Table S1). The best model according to the Akaike Information Criterion (AIC) contained the five miRNAs hsa-let-7g-3p, hsa-miR-6508-5p, hsa-miR-210-5p, hsa-miR-4306 and hsa-miR-7161-3p with the following Cox proportional hazard coefficients: -0.5214183, -0.5254865, 0.6461524, -0.3678727 and -0.8165854, respectively. The coefficients were subsequently used for individual risk score calculation after linear combination with appropriate expressions of the signature miRNAs. Using the median risk score as a cut-off, 43 patients of the DTKK-ROG sample (training set) were assigned to the low-risk group (median time to event not reached (NR), 95% CI 2047–not estimable (NE); eight events) and 42 to the high-risk group (median time to event 748 days, 95% CI 459–NE; 24 events). As expected, the groups differed significantly in their risk of recurrence (HR 4.42, 95% CI 1.98–9.88; log-rank $P < 0.001$; Figures 1A, S2).

We applied the five-miRNA-based signature prediction model to the miRNA expression data set of the LMU-KKG sample (validation set) using the cut-off as calculated from the training sample data (0.03629712) and assigned 38 patients to the low-risk (median NR, 95% CI NE–NE; four events) and 39 patients to the high-risk group (median NR, 95% CI 708–NE; 15 events). The risk for recurrence of the high-risk patients was significantly increased compared to that of the low-risk patients (HR 4.24, 95% CI 1.40–12.81; $P = 0.005$) confirming the prognostic value of the five-miRNA-signature (Figures 1A, S2). miRNA-based risk group classification was not

associated with simultaneous chemotherapy treatment (Table 1), which was further supported after stratification to LMU-KKG patients treated by concurrent radiochemotherapy ($n=49$; HR 3.85, 95% CI 1.09-13.58, $P=0.024$; Figure S3).

Moreover, high-risk patients of both samples showed significantly reduced recurrence-free survival, overall survival and disease-specific survival rates (Figure 1B). We could also demonstrate an impact of both failure sites (locoregional and distant) on the risk stratification, while low- and high-risk patients did not differ in non-tumor related death (Figure S4).

In order to assess whether the five-miRNA-signature was an independent prognosticator, associations of known clinicopathological factors with the miRNA-defined risk groups were tested. TNM T stage, ECE and tumor localization were associated with the miRNA risk groups (Table 1). In the subsequent univariate Cox proportional hazard analysis, TNM T stage and lymphovascular invasion (LVI) were significantly associated with freedom from recurrence in both samples, ECE was identified as a significant parameter in the DTKK-ROG sample only, whereas no differences between the three tumor localizations were observed (Table S2; Figures S5-S7). After adjustment for these parameters in multivariate Cox regression analysis, the five-miRNA-signature retained its independent and exclusive prognostic role in both samples (training set: HR 5.55, 95% CI 2.09-14.79, $P<0.001$; validation set: HR 3.94, 95% CI 1.23-12.59, $P=0.021$; Table 2).

We analyzed the sensitivity and specificity of the five-miRNA-signature in the prediction of different clinical endpoints in comparison to the clinical prognostic parameters TNM T stage, LVI and ECE. At 5 years follow-up, the five-miRNA-signature demonstrated a superior prediction of all endpoints analyzed (Figures 2A, S8). Furthermore, in time-dependent analysis (follow-up years 1 to 5), the five-miRNA-signature superiorly predicted all endpoints from 2 to 5 years compared to the clinicopathological parameters. After one year follow-up, higher AUCs for the miRNA-signature compared with the analyzed endpoints were only observed in the training set for the endpoints disease-specific survival and overall survival (Figures 2B, S9, S10). After combining the five-miRNA-signature with the clinicopathological parameters (TNM T stage, LVI, ECE) an even better prediction of all endpoints from 2 to 5 years was achieved for both HNSCC samples, also when compared to combinations of the clinicopathological risk factors (Figures 2C, S11). This was also the case after one year follow-up in the DTKK-ROG sample.

In order to obtain deeper insights into the biological regulatory function of the signature miRNAs, we generated a miRNA-mRNA target regulatory network comprising experimentally

validated miRNA-target interactions, whereby twelve target genes were found to be shared by the signature miRNAs (Table S3, Figure S12). Pathway enrichment analysis of the target genes revealed 36 pathways including *p53*, *ATM*, and *FoxO signalling*, *DNA double-strand break response*, *pre-NOTCH expression and processing*, *mitosis* and *senescence* associated pathways (Table S4).

For technical validation of the five-miRNA-signature and potential clinical diagnostic application, we measured the expression of the signature miRNAs in the validation set ($n=71$) by qRT-PCR confirming the microarray-derived results as the miRNA-classified risk groups significantly differed in freedom of recurrence (HR 5.07, 95% CI 1.17–21.94, $P=0.016$; Figure S13)

In a Kaplan-Meier analysis in which the samples were pooled ($n=162$) and stratified according to resection margin status, TNM T stage, TNM N stage, ECE and tumor localization the resulting five-miRNA-signature risk groups significantly differed in clinical outcome (Figures S14, S15). This motivated us to further combine the five-miRNA-signature with clinically relevant parameters. RPA identified four different risk groups for recurrence (“low-risk”, “low-intermediate-risk”, “high-intermediate-risk” and “high-risk”) including the five-miRNA-signature as strongest parameter together with TNM T stage, ECE and TNM N stage (Figure 3 and extended Figure version S16). The worst prognostic group included miRNA-signature-high-risk patients with ECE-positive T3/T4 tumors (median freedom from recurrence 438 days), while miRNA-signature-low-risk patients with T1/T2 N0/N1 HNSCC had the best prognosis (no event). The four risk groups also significantly differed with regard to locoregional and distant control, recurrence-free survival, overall survival and disease-specific survival (Figures S17, S18). RPA considering only the clinical parameters identified three risk groups for recurrence with T stage as the strongest parameter together with ECE and N stage (Figure S19A). Combining the three RPA derived risk groups with the risk factor of our five-miRNA-signature revealed patient subgroups significantly differing in clinical outcome (“RPA intermediate-risk”: HR 2.71, 95% CI 1.21-6.06, $P=0.012$; “RPA high-risk”: HR 12.20, 95% CI 1.54-96.90, $P=0.004$; Figure S19B).

Discussion

Here we report, for the first time a five-miRNA-signature in HPV-negative patients that predicts decreased cancer control following adjuvant radiochemotherapy. Freedom from recurrence was the chosen primary endpoint to better estimate treatment effects, as HNSCC patients often suffer from multiple comorbidities that affect overall survival (19). Overall, baseline and treatment characteristics of our patients were balanced and compare well to reports on HPV-negative HNSCC. Remarkably our identified five-miRNA-signature predicts survival as well. Of note: its prognostic significance is independent from known clinical parameters

A potential limitation of the study is the fact that clinical data for both samples were obtained retrospectively. We thus cannot fully exclude certain selection bias. Heterogeneity due to inclusion of a multicenter HNSCC patient sample minimized and potentially excluded selection bias. In addition, the signature's robustness and potential clinical applicability was underlined by identification in a multicenter sample and validation in an independent monocentric sample. Most other studies introducing prognostic miRNA-signatures (e.g. ovarian, nasopharyngeal and colon cancer) followed a comparable strategy (8,20,21).

The fact that the DKTK-ROG sample exclusively included HNSCC patients treated by post-operative radiochemotherapy, whereas the LMU-KKG sample comprised both adjuvant treatment groups – radiotherapy with simultaneous chemotherapy and radiotherapy alone – might be seen as another limitation of our study. However, from our point of view, the independence of the five-miRNA-signature from the addition of simultaneous chemotherapy even strengthens the potential of our five-miRNA-signature.

A further potential shortcoming of our study is that the final RPA was limited by small numbers of patients. In order to achieve the highest possible number of cases and the maximum statistical power, we pooled both HNSCC samples for this analysis ($n=162$). In all clinical endpoints a significant separation of risk groups defined by clinical risk factors combined with the five-miRNA-signature was achieved.

To substantiate our findings on patient stratification into risk groups, further validation of our five-miRNA-signature in independent retrospective and in particular prospective patient populations with fully annotated clinical data will be important future steps.

Previous studies have identified multiple deregulated miRNAs in HNSCC partly with prognostic relevance for patients (10-12,22-26). A meta-analysis revealed that in particular overexpression of miR-21, one of the most frequently studied cancer-related miRNAs, predicts poor prognosis in HNSCC (10). However, in general, the overlap of prognostic miRNAs across different HNSCC studies is small. This can be potentially explained by differences in demography, treatment parameters, composition of patient subgroups (e.g. subsite and HPV-status) as well as by methodological issues like the lack of independent validation, limitations due to small sample size, the analysis of different endpoints, the number of miRNAs screened and the non-availability of thorough clinical information including HPV-status (27). Our comprehensive miRNA profiling approach deliberately and exclusively focused on HPV-negative patients based on the fact that all current data indicate a completely distinct molecular pathogenesis of HPV-associated cancer, which, meanwhile, is regarded a distinct clinical entity (2,6).

Nevertheless, in our study we were able to confirm previously reported prognostic miRNAs in HNSCC such as hsa-miR-21-3p, hsa-let-7g-3p, hsa-miR-210-5p and hsa-miR-210-3p (Figure S20) underlining the validity of our miRNA analysis (10,22,26,28,29). In addition, hsa-mir-210-5p and hsa-let-7g-3p form part of our five-miRNA-signature. hsa-let-7g was shown to predict prognosis in oral cavity squamous cell carcinoma (29) and breast cancer patients (30) via inhibition of cell invasion and metastasis. Besides head and neck cancer (28), hsa-mir-210 was already reported as prognostic factor in breast cancer (31-34), soft-tissue sarcoma (35), osteosarcoma (36), pancreatic cancer (37), non-small cell lung cancer (38), renal cancer (39) and glioblastoma (40). Multiple functions of hsa-miR-210 are described including hypoxic response, regulation of mitochondrial metabolism, cell cycle, cell survival, differentiation DNA repair and immune response (41). To the best of our knowledge, the remaining three signature miRNAs (hsa-miR-6508-5p, hsa-miR-4306 and hsa-miR-7161-3p) have not yet been associated with HNSCC or cancer in general.

miRNAs are integrative regulator molecules with a highly promiscuous nature thereby interfering with multiple pathways. Thus, it is not possible to deduce a definitive functional role of a given miRNA within a signaling network. Nevertheless, studying the miRNA-mRNA-target network our five-miRNA-signature suggests enrichment of specific signaling pathways: *p53*, *ATM*, *FoxO signaling*, and *DNA double strand break response, pre-NOTCH expression and processing*, as well as *mitosis* and *senescence* associated pathways. Several of the pathways and miRNA target genes were already shown to be relevant for the pathogenesis and radiation response of HNSCC

(5-7,42-47). Mutations of IGF1R and ARID1A and the involvement of CADM1 and SOD2 in HNSCC have been reported (6,43,46,47).

Gene expression relates to prognosis of HNSCC (48) as does a seven-gene signature, recently also described in our patients (49); this signature, however, predicts freedom from recurrence independently from the above mentioned five-miRNA-signature (unpublished). Analogous to their prognostic independence the molecular impact of the Schmidt et al. seven-gene signature shows no obvious overlap with that of our five-miRNA-signature (49). However, to pin down mechanisms and pathogenic relevance of the five-miRNA-signature further studies are required.

At present, treatment decisions for patients with HNSCC are guided predominantly by clinical findings. The only relevant biological marker with yet limited influence on treatment decisions is HPV-status (1). A key prerequisite for the potential clinical application of a molecular signature is a robust, fast and easy to perform laboratory assay. Our qRT-PCR validation of the high-throughput omics data is a first step in this direction.

The five-miRNA-signature's potential is particularly exemplified by the fact that, when combined with the clinically relevant prognostic parameters TNM T stage, ECE and TNM N stage, it allowed the significant stratification of patients into four risk groups for recurrence. Strikingly, in this context, the five-miRNA-signature was the strongest factor for patient stratification. Furthermore, the integration of the molecular signature with clinical factors not only improved the prediction of outcome but also allowed a more detailed, clinically meaningful stratification of patients, which, in turn, could be used as a clinical patient stratification tool.

Possible personalized treatment options include consideration of adjusting therapy intensity according to the overall risk for therapy failure. In particular patients with the highest risk of recurrence, for whom the standard treatment is not sufficient, might be candidates for more personalized treatment options such as the addition of targeted drugs or immune checkpoint inhibitors to radio(chemo)therapy, dose escalation or further (neo)adjuvant chemotherapy. On the other hand, for patients with the lowest risk of recurrence de-escalation strategies for the reduction of therapy-associated toxicity could be considered. Here dose de-escalation and the omission of chemotherapy would be options, as the long-term benefit from the addition of simultaneous chemotherapy to radiotherapy is not given for all patients (50). Further, the five-miRNA-signature represents the basis for a more focused search for molecular therapeutic targets improving therapy success for appropriate patients.

In order to evaluate the predictive value of the five-miRNA-signature for the guidance of treatment decisions, prospective validation studies and clinical trials considering treatment stratification are required in future.

In summary, the herein identified prognostic five-miRNA-signature independently predicts disease control and survival of HPV-negative patients. The target gene network of the signature miRNAs is well in line with known mechanisms driving HNSCC pathogenesis. In combination with established prognostic clinical parameters the ability of the signature to predict disease control and survival even improves and allows the definition of four prognostically distinct groups. These may provide an important step towards personalized HNSCC treatment.

Acknowledgements

The study was supported by the DKTK-ROG and the Clinical Cooperation Group “Personalized Radiotherapy in Head and Neck Cancer”, Helmholtz Zentrum München.

The authors want to thank L. Dajka, S. Heuer, C. Innerlohinger, U Buchholz, and C.-M. Pflüger for their excellent technical assistance, R. Caldwell for editorial assistance, all co-workers of the Clinical Cooperation Group “Personalized Radiotherapy in Head and Neck Cancer” for scientific support, and all members of the DKTK-ROG for their valuable contribution to the DKTK-ROG data set.

References

1. O'Sullivan B, Huang SH, Siu LL, Waldron J, Zhao H, Perez-Ordonez B, *et al.* Deintensification candidate subgroups in human papillomavirus-related oropharyngeal cancer according to minimal risk of distant metastasis. *J Clin Oncol* **2013**;31:543-50
2. O'Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, *et al.* Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol* **2016**;17:440-51
3. Bossi P, Alfieri S. Investigational drugs for head and neck cancer. *Expert Opin Investig Drugs* **2016**;25:797-810
4. Argiris A, Harrington KJ, Tahara M, Schulten J, Chomette P, Ferreira Castro A, *et al.* Evidence-Based Treatment Options in Recurrent and/or Metastatic Squamous Cell Carcinoma of the Head and Neck. *Front Oncol* **2017**;7:72
5. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **2011**;333:1154-7
6. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **2015**;517:576-82
7. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **2011**;333:1157-60
8. Bagnoli M, Canevari S, Califano D, Losito S, Maio MD, Raspagliesi F, *et al.* Development and validation of a microRNA-based signature (MiROvaR) to predict early relapse or progression of epithelial ovarian cancer: a cohort study. *Lancet Oncol* **2016**;17:1137-46
9. Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med* **2012**;4:143-59
10. Jamali Z, Asl Aminabadi N, Attaran R, Pournagiazar F, Ghertasi Oskouei S, Ahmadpour F. MicroRNAs as prognostic molecular signatures in human head and neck squamous cell carcinoma: a systematic review and meta-analysis. *Oral Oncol* **2015**;51:321-31
11. Koshizuka K, Hanazawa T, Fukumoto I, Kikkawa N, Okamoto Y, Seki N. The microRNA signatures: aberrantly expressed microRNAs in head and neck squamous cell carcinoma. *J Hum Genet* **2017**;62:3-13
12. Sethi N, Wright A, Wood H, Rabbitts P. MicroRNAs and head and neck cancer: reviewing the first decade of research. *Eur J Cancer* **2014**;50:2619-35
13. Lohaus F, Linge A, Tinhofer I, Budach V, Gkika E, Stuschke M, *et al.* HPV16 DNA status is a strong prognosticator of loco-regional control after postoperative radiochemotherapy of locally advanced oropharyngeal carcinoma: results from a multicentre explorative study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). *Radiother Oncol* **2014**;113:317-23
14. Maihoefer C, Schüttrumpf L, Macht C, Pflugradt U, Hess J, Schneider L, *et al.* Postoperative (chemo) radiation in patients with squamous cell cancers of the head and neck – clinical results from the cohort of the clinical cooperation group “Personalized Radiotherapy in Head and Neck Cancer”. *Radiation Oncology* **2018**;13:123
15. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2016.

16. Cho H, Yu A, Kim S, Kang J, Hong S-M. Robust Likelihood-Based Survival Modeling with Microarray Data. 2009 **2009**;29:16
17. Niyazi M, Pitea A, Mittelbronn M, Steinbach J, Sticht C, Zehentmayr F, *et al.* A 4-miRNA signature predicts the therapeutic outcome of glioblastoma. *Oncotarget* **2016**
18. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **2016**;44:D481-7
19. Piccirillo JF, Vlahiotis A. Comorbidity in patients with cancer of the head and neck: prevalence and impact on treatment and prognosis. *Curr Oncol Rep* **2006**;8:123-9
20. Zhang JX, Song W, Chen ZH, Wei JH, Liao YJ, Lei J, *et al.* Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol* **2013**;14:1295-306
21. Liu N, Chen NY, Cui RX, Li WF, Li Y, Wei RR, *et al.* Prognostic value of a microRNA signature in nasopharyngeal carcinoma: a microRNA expression analysis. *Lancet Oncol* **2012**;13:633-41
22. Ganci F, Sacconi A, Manciocco V, Sperduti I, Battaglia P, Covello R, *et al.* MicroRNA expression as predictor of local recurrence risk in oral squamous cell carcinoma. *Head Neck* **2016**;38 Suppl 1:E189-97
23. Gao G, Gay HA, Chernock RD, Zhang TR, Luo J, Thorstad WL, *et al.* A microRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer* **2013**;119:72-80
24. Hess AK, Muer A, Mairinger FD, Weichert W, Stenzinger A, Hummel M, *et al.* MiR-200b and miR-155 as predictive biomarkers for the efficacy of chemoradiation in locally advanced head and neck squamous cell carcinoma. *Eur J Cancer* **2017**;77:3-12
25. Shi H, Chen J, Li Y, Li G, Zhong R, Du D, *et al.* Identification of a six microRNA signature as a novel potential prognostic biomarker in patients with head and neck squamous cell carcinoma. *Oncotarget* **2016**;7:21579-90
26. Wong N, Khwaja SS, Baker CM, Gay HA, Thorstad WL, Daly MD, *et al.* Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer Med* **2016**;5:1619-28
27. Shi X, Yi H, Ma S. Measures for the degree of overlap of gene signatures and applications to TCGA. *Brief Bioinform* **2015**;16:735-44
28. Gee HE, Camps C, Buffa FM, Patiar S, Winter SC, Betts G, *et al.* hsa-mir-210 is a marker of tumor hypoxia and a prognostic factor in head and neck cancer. *Cancer* **2010**;116:2148-58
29. Peng SC, Liao CT, Peng CH, Cheng AJ, Chen SJ, Huang CG, *et al.* MicroRNAs MiR-218, MiR-125b, and Let-7g predict prognosis in patients with oral cavity squamous cell carcinoma. *PLoS One* **2014**;9:e102403
30. Qian P, Zuo Z, Wu Z, Meng X, Li G, Wu Z, *et al.* Pivotal role of reduced let-7g expression in breast cancer invasion and metastasis. *Cancer Res* **2011**;71:6463-74
31. Camps C, Buffa FM, Colella S, Moore J, Sotiriou C, Sheldon H, *et al.* hsa-miR-210 Is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clin Cancer Res* **2008**;14:1340-8
32. Rothe F, Ignatiadis M, Chaboteaux C, Haibe-Kains B, Kheddoumi N, Majjaj S, *et al.* Global microRNA expression profiling identifies MiR-210 associated with tumor proliferation, invasion and poor clinical outcome in breast cancer. *PLoS One* **2011**;6:e20980

33. Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, *et al.* Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proc Natl Acad Sci U S A* **2012**;109:3024-9
34. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, *et al.* microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res* **2011**;71:5635-45
35. Greither T, Wurl P, Grochola L, Bond G, Bache M, Kappler M, *et al.* Expression of microRNA 210 associates with poor survival and age of tumor onset of soft-tissue sarcoma patients. *Int J Cancer* **2012**;130:1230-5
36. Cai H, Lin L, Cai H, Tang M, Wang Z. Prognostic evaluation of microRNA-210 expression in pediatric osteosarcoma. *Med Oncol* **2013**;30:499
37. Greither T, Grochola LF, Udelnow A, Lautenschlager C, Wurl P, Taubert H. Elevated expression of microRNAs 155, 203, 210 and 222 in pancreatic tumors is associated with poorer survival. *Int J Cancer* **2010**;126:73-80
38. Eilertsen M, Andersen S, Al-Saad S, Richardsen E, Stenvold H, Hald SM, *et al.* Positive prognostic impact of miR-210 in non-small cell lung cancer. *Lung Cancer* **2014**;83:272-8
39. McCormick RI, Blick C, Ragoussis J, Schoedel J, Mole DR, Young AC, *et al.* miR-210 is a target of hypoxia-inducible factors 1 and 2 in renal cancer, regulates ISCU and correlates with good prognosis. *Br J Cancer* **2013**;108:1133-42
40. Qiu S, Lin S, Hu D, Feng Y, Tan Y, Peng Y. Interactions of miR-323/miR-326/miR-329 and miR-130a/miR-155/miR-210 as prognostic indicators for clinical outcome of glioblastoma patients. *J Transl Med* **2013**;11:10
41. Qin Q, Furong W, Baosheng L. Multiple functions of hypoxia-regulated miR-210 in cancer. *J Exp Clin Cancer Res* **2014**;33:50
42. Hess J, Unger K, Orth M, Schotz U, Schuttrumpf L, Zangen V, *et al.* Genomic amplification of Fanconi anemia complementation group A (*FancA*) in head and neck squamous cell carcinoma (HNSCC): Cellular mechanisms of radioresistance and clinical relevance. *Cancer Lett* **2017**;386:87-99
43. Lui VW, Hedberg ML, Li H, Vangara BS, Pendleton K, Zeng Y, *et al.* Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discov* **2013**;3:761-9
44. Michna A, Schotz 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
45. Summerer I, Hess J, Pitea A, Unger K, Hieber L, Selmansberger M, *et al.* Integrative analysis of the microRNA-mRNA response to radiochemotherapy in primary head and neck squamous cell carcinoma cells. *BMC Genomics* **2015**;16:654
46. Vallath S, Sage EK, Kolluri KK, Lourenco SN, Teixeira VS, Chimalapati S, *et al.* *CADM1* inhibits squamous cell carcinoma progression by reducing *STAT3* activity. *Sci Rep* **2016**;6:24006
47. Ye H, Wang A, Lee BS, Yu T, Sheng S, Peng T, *et al.* Proteomic based identification of manganese superoxide dismutase 2 (*SOD2*) as a metastasis marker for oral squamous cell carcinoma. *Cancer Genomics Proteomics* **2008**;5:85-94
48. Tonella L, Giannoccaro M, Alfieri S, Canevari S, De Cecco L. Gene Expression Signatures for Head and Neck Cancer Patient Stratification: Are Results Ready for Clinical Application? *Curr Treat Options Oncol* **2017**;18:32

49. Schmidt S, Linge A, Zwanenburg A, Leger S, Lohaus F, Krenn C, *et al.* Development and Validation of a Gene Signature for Patients with Head and Neck Carcinomas Treated by Postoperative Radio(chemo)therapy. *Clin Cancer Res* **2018**
50. Cooper JS, Zhang Q, Pajak TF, Forastiere AA, Jacobs J, Saxman SB, *et al.* Long-term follow-up of the RTOG 9501/intergroup phase III trial: postoperative concurrent radiation therapy and chemotherapy in high-risk squamous cell carcinoma of the head and neck. *Int J Radiat Oncol Biol Phys* **2012**;84:1198-205

Tables

Table 1: Clinical and pathological characteristics of HNSCC patients included in the DKTK-ROG and the LMU-KKG sample and stratified according to the five-miRNA-signature

	Training set DKTK-ROG (n=85)				Validation set LMU-KKG (n=77)			
	Number of all patients	low-risk (n=43)	high-risk (n=42)	p-value*	Number of all patients	low-risk (n=38)	high-risk (n=39)	p-value*
Age (years)				0.77				0.86
<45	7 (8%)	2 (5%)	5 (12%)		3 (4%)	1 (3%)	2 (5%)	
45-54	26 (31%)	13 (30%)	13 (31%)		17 (22%)	7 (18%)	10 (26%)	
55-64	35 (41%)	18 (42%)	17 (40%)		28 (36%)	15 (39%)	13 (33%)	
65-74	17 (20%)	10 (23%)	7 (17%)		26 (34%)	13 (34%)	13 (33%)	
>75	0	0	0		3 (4%)	2 (5%)	1 (3%)	
Sex				1.0				1.0
Male	67 (79%)	34 (79%)	33 (79%)		52 (68%)	26 (68%)	26 (67%)	
Female	18 (21%)	9 (21%)	9 (21%)		25 (32%)	12 (32%)	13 (33%)	
Tumor Localization				0.12				0.022
Hypopharynx	13 (15%)	9 (21%)	4 (10%)		15 (19%)	4 (11%)	11 (28%)	
Oral cavity	32 (38%)	12 (28%)	20 (48%)		27 (35%)	11 (29%)	16 (41%)	
Oropharynx	40 (47%)	22 (51%)	18 (43%)		35 (45%)	23 (61%)	12 (31%)	
UICC TNM Stage				0.56				0.79
I	0	0	0		2 (3%)	1 (3%)	1 (3%)	
II	3 (4%)	2 (5%)	1 (2%)		6 (8%)	4 (11%)	2 (5%)	
III	13 (15%)	5 (12%)	8 (19%)		23 (30%)	12 (32%)	11 (28%)	
IV	69 (81%)	36 (84%)	33 (79%)		46 (60%)	21 (55%)	25 (64%)	
T stage				0.33				0.042
T1	12 (14%)	9 (21%)	3 (7%)		17 (22%)	9 (24%)	8 (21%)	
T2	35 (41%)	17 (40%)	18 (43%)		29 (38%)	18 (47%)	11 (28%)	
T3	22 (26%)	10 (23%)	12 (29%)		21 (27%)	10 (26%)	11 (28%)	
T4	16 (19%)	7 (16%)	9 (21%)		10 (13%)	1 (3%)	9 (23%)	
N stage				0.14				0.41
N0	10 (12%)	5 (12%)	5 (12%)		19 (25%)	8 (21%)	11 (28%)	
N1	10 (12%)	2 (5%)	8 (19%)		20 (26%)	10 (26%)	10 (26%)	
N2	57 (67%)	33 (77%)	24 (57%)		36 (47%)	20 (53%)	16 (41%)	
N3	8 (9%)	3 (7%)	5 (12%)		2 (3%)	0	2 (5%)	
Lymphovascular invasion (LVI)				0.46				1.0
0	42 (49%)	25 (58%)	17 (40%)		50 (65%)	26 (68%)	24 (62%)	
1	27 (32%)	13 (30%)	14 (33%)		17 (22%)	9 (24%)	8 (21%)	
Missing information	16 (19%)	5 (12%)	11 (26%)		10 (13%)	3 (8%)	7 (18%)	
Venous tumor invasion (VTI)				1.0				1.0
0	62 (73%)	33 (77%)	29 (69%)		66 (86%)	34 (89%)	32 (82%)	
1	7 (8%)	4 (9%)	3 (7%)		3 (4%)	2 (5%)	1 (3%)	
Missing information	16 (19%)	6 (14%)	10 (24%)		8 (10%)	0	6 (15%)	
Perineural invasion (PNI)				1.0				0.55
0	0	0	0		37 (48%)	19 (50%)	18 (46%)	
1	0	0	0		15 (19%)	6 (16%)	9 (23%)	
Missing information	85 (100%)	43 (100%)	42 (100%)		25 (32%)	13 (34%)	12 (31%)	
Resection margin status				0.52				0.49
0	45 (53%)	21 (49%)	24 (57%)		57 (74%)	28 (74%)	29 (74%)	
1	40 (47%)	22 (51%)	18 (43%)		17 (22%)	7 (18%)	10 (26%)	
2	0	0	0		1 (1%)	1 (3%)	0	
Missing information	0	0	0		2 (3%)	2 (5%)	0	
Extracapsular extension (ECE)				0.007				0.38
yes	41 (48%)	14 (33%)	27 (64%)		25 (32%)	11 (29%)	14 (36%)	
no	34 (40%)	24 (56%)	10 (24%)		32 (42%)	19 (50%)	13 (33%)	
not applicable (N0)	10 (12%)	5 (12%)	5 (12%)		19 (25%)	8 (21%)	11 (28%)	
Missing information	0	0	0		1 (1%)	0	1 (3%)	
Grading				0.56				0.29
1 (well differentiated)	3 (4%)	2 (5%)	1 (2%)		2 (3%)	2 (5%)	0	
2 (moderately differentiated)	50 (59%)	23 (53%)	27 (64%)		34 (44%)	15 (39%)	19 (49%)	
3 (poorly differentiated)	32 (38%)	18 (42%)	14 (33%)		41 (53%)	21 (55%)	20 (51%)	
ECOG performance status				0.64				0.20
0	18 (21%)	8 (19%)	10 (24%)		13 (17%)	4 (11%)	9 (23%)	
1	33 (39%)	17 (40%)	16 (38%)		40 (52%)	21 (55%)	19 (49%)	
2	6 (7%)	4 (9%)	2 (5%)		5 (6%)	1 (3%)	4 (10%)	
Missing information	28 (33%)	14 (33%)	14 (33%)		19 (25%)	12 (32%)	7 (18%)	
Smoking status				0.18				0.68
Non-smoker	5 (6%)	4 (9%)	1 (2%)		6 (8%)	2 (5%)	4 (10%)	
Smoker	52 (61%)	23 (53%)	29 (69%)		52 (68%)	25 (66%)	27 (69%)	
Missing information	28 (33%)	16 (37%)	12 (29%)		19 (25%)	11 (29%)	8 (21%)	
Smoking history – pack-years				0.20				0.67
≤10 (including non-smokers)	7 (8%)	5 (12%)	2 (5%)		6 (8%)	2 (5%)	4 (10%)	
>10	23 (27%)	9 (21%)	14 (33%)		48 (62%)	25 (66%)	23 (59%)	
Missing information	55 (65%)	29 (67%)	26 (62%)		23 (30%)	11 (29%)	12 (31%)	
Simultaneous Chemotherapy				1.0				0.16
Yes	85 (100%)	43 (100%)	42 (100%)		49 (64%)	21 (55%)	28 (72%)	
No	0	0	0		28 (36%)	17 (45%)	11 (28%)	

Data are numbers (%). *Chi-square test or Fisher's exact test.

Table 2: Multivariate Cox regression analysis of the five-miRNA-signature and clinicopathological parameters with freedom from recurrence (training and validation set)

Parameter	Training set DTK-ROG		Validation set LMU-KKG	
	HR (95% CI)	p-value	HR (95% CI)	p-value
Five-miRNA-signature (high-risk vs low-risk)	5.55 (2.09-14.79)	<0.001	3.94 (1.23-12.59)	0.021
TNM T stage (T3/T4 vs T1/T2)	2.19 (0.96-5.02)	0.064	2.71 (0.99-7.44)	0.052
Lymphovascular invasion (yes vs no)	2.22 (0.99-4.97)	0.053	2.50 (0.84-7.45)	0.099
Extracapsular extension (yes vs no [*])	1.45 (0.61-3.48)	0.40	2.29 (0.77-6.78)	0.13

^{*}N0 tumors were included in the group of extracapsular extension negative tumors

Figure Legends

Figure 1: Freedom from recurrence stratified by risk according to the five-miRNA-signature: miRNA expression and Kaplan-Meier curves in the DKTK-ROG (training set) and the LMU-KKG (validation set) sample

(A) Upper panel: Heat map colors indicate scaled miRNA log₂ expression values multiplied by the Cox proportional hazard coefficients (coxph) from low (blue) to high (red) on a scale from -3 to 3 for each of the five signature miRNAs in the DKTK-ROG (left panel) and the LMU-KKG sample (right panel). Lower panel: Kaplan-Meier curves for the endpoint freedom from recurrence for HNSCC patients of the training (DKTK-ROG sample; left panel) and validation set (LMU-KKG sample; right panel) stratified into low- and high-risk patients according to the five-miRNA-signature. *P*-values are derived by log-rank test. (B) Kaplan-Meier curves for recurrence-free survival (upper panel), overall survival (middle panel) and disease-specific survival (lower panel) in patients of the training (DKTK-ROG sample; left) and validation set (LMU-KKG sample; right) stratified according to their risk (low- and high-risk group) by the five-miRNA-signature.

Figure 2: Performance of the prediction of freedom from recurrence comparing the five-miRNA-signature with clinicopathological risk factors

(A) Sensitivity and specificity derived areas under the curve (AUCs) for the prediction of freedom from recurrence in the DKTK-ROG (training set; left panel) and the LMU-KKG sample (validation set; right panel) at five follow-up years. The AUCs and the 95% CI of the five-miRNA-signature derived risk factor (black dashed curve), TNM T stage, lymphovascular invasion (LVI) and extracapsular extension (ECE) are shown.

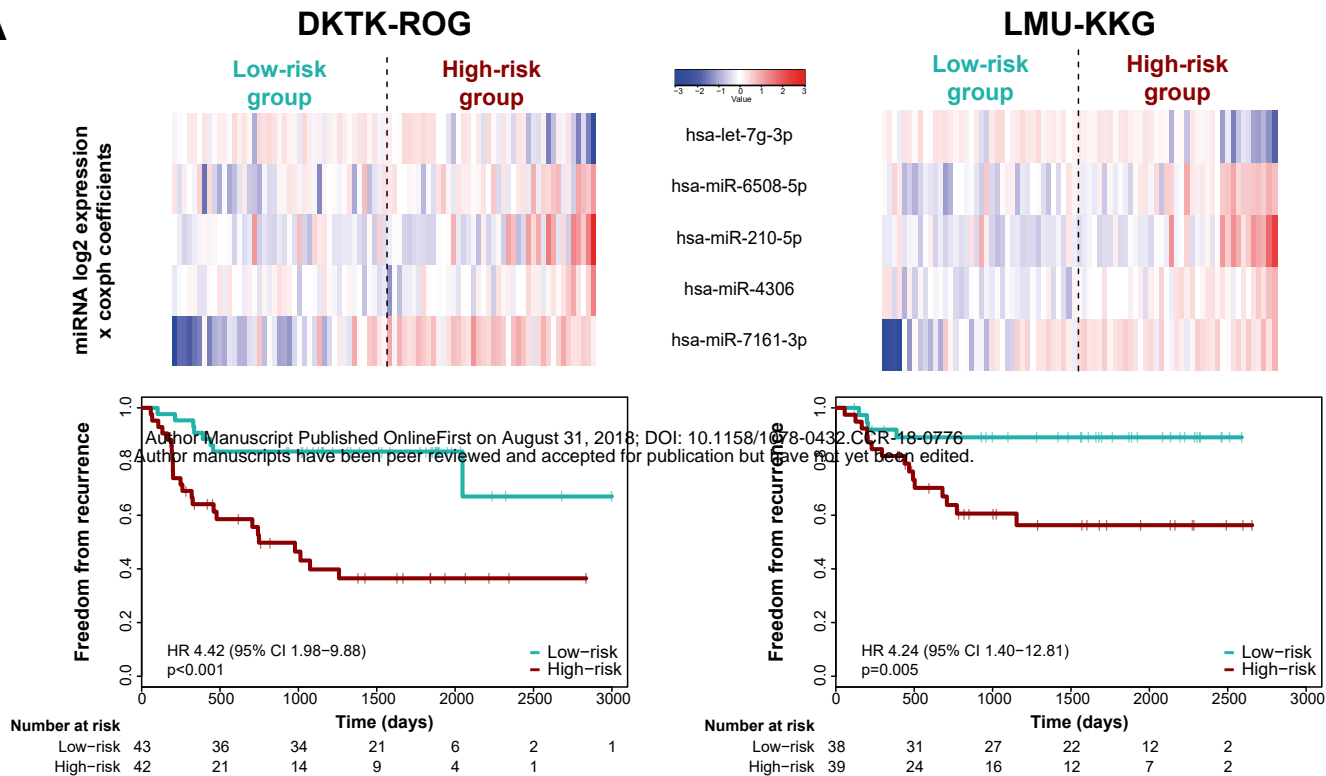
Time-dependent sensitivity and specificity derived AUCs for the prediction of freedom from recurrence in the DKTK-ROG (left panel) and the LMU-KKG sample (right panel) at follow-up years 1-5: (B) AUCs of the five-miRNA-signature derived risk factor (black dashed curve), TNM T stage, LVI and ECE. (C) AUCs for the five-miRNA-signature derived risk factor alone (black dashed curve), the five-miRNA-signature combined with TNM T stage, LVI and ECE (purple and greenish curves) and combinations of the clinicopathological risk factors TNM T stage, LVI and ECE (bluish curves).

Figure 3: Risk groups for recurrence identified by recursive partitioning analysis (RPA)

RPA tree and risk groups for recurrence combining the parameters five-miRNA-signature (high-risk, low-risk), ECE (negative - including N0 tumors, positive), T stage (T1/T2, T3/T4) and N stage (N0/N1, N2/N3) in the pooled HNSCC data set ($n=162$). Each node shows the predicted probability of recurrence (locoregional or distant failure; color code low to high: blue-red), the number of events for the total number of patients and the percentage of observations in the node. Kaplan-Meier curves for the endpoint freedom from recurrence for the four identified risk groups “low-risk”, “low-intermediate-risk”, “high-intermediate-risk” and “high-risk”. Multivariate and pairwise comparisons are shown. *P*-values are derived by log-rank test. See extended Figure version Supplementary Figure S16.

Figure 1

A



B

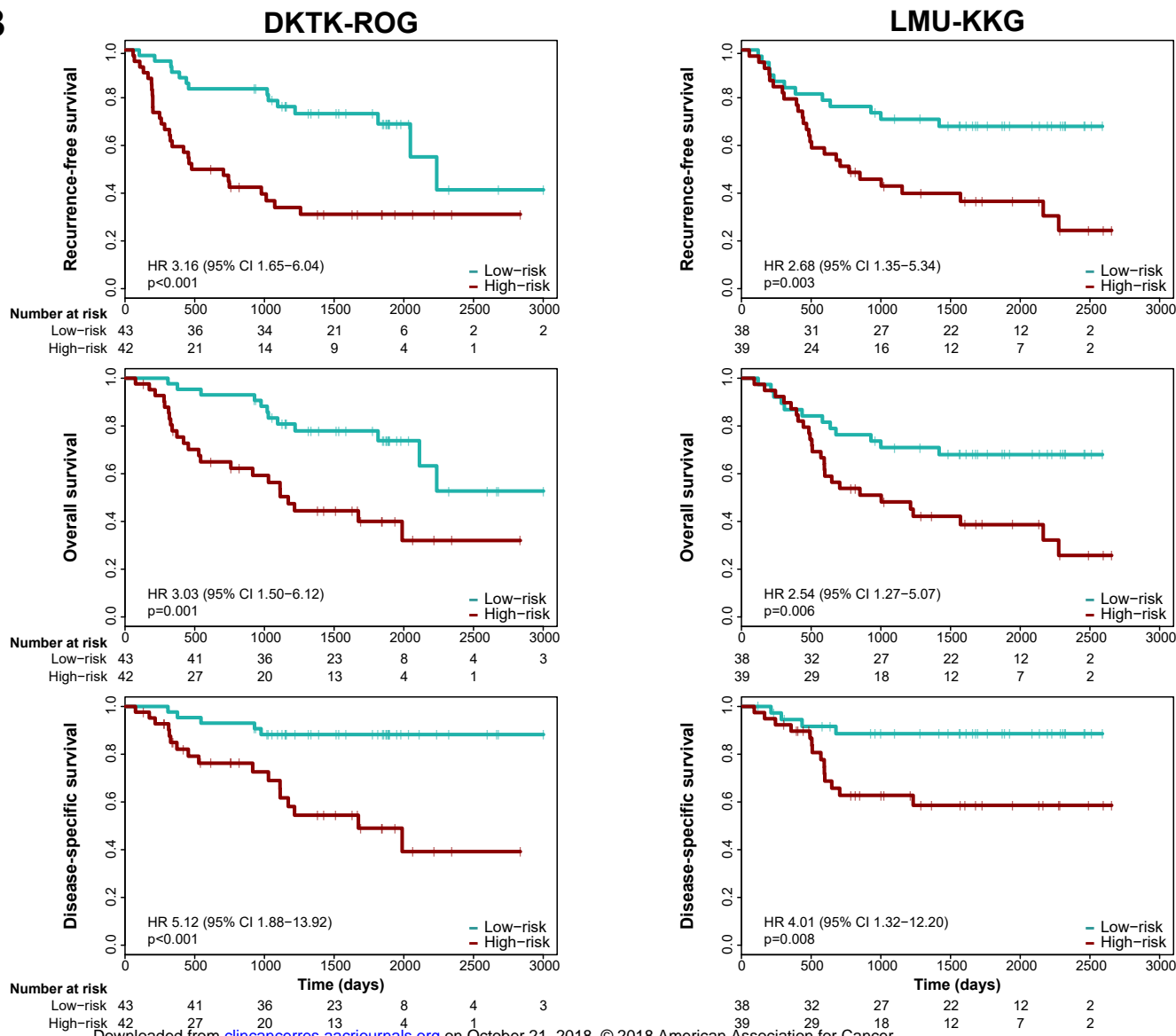
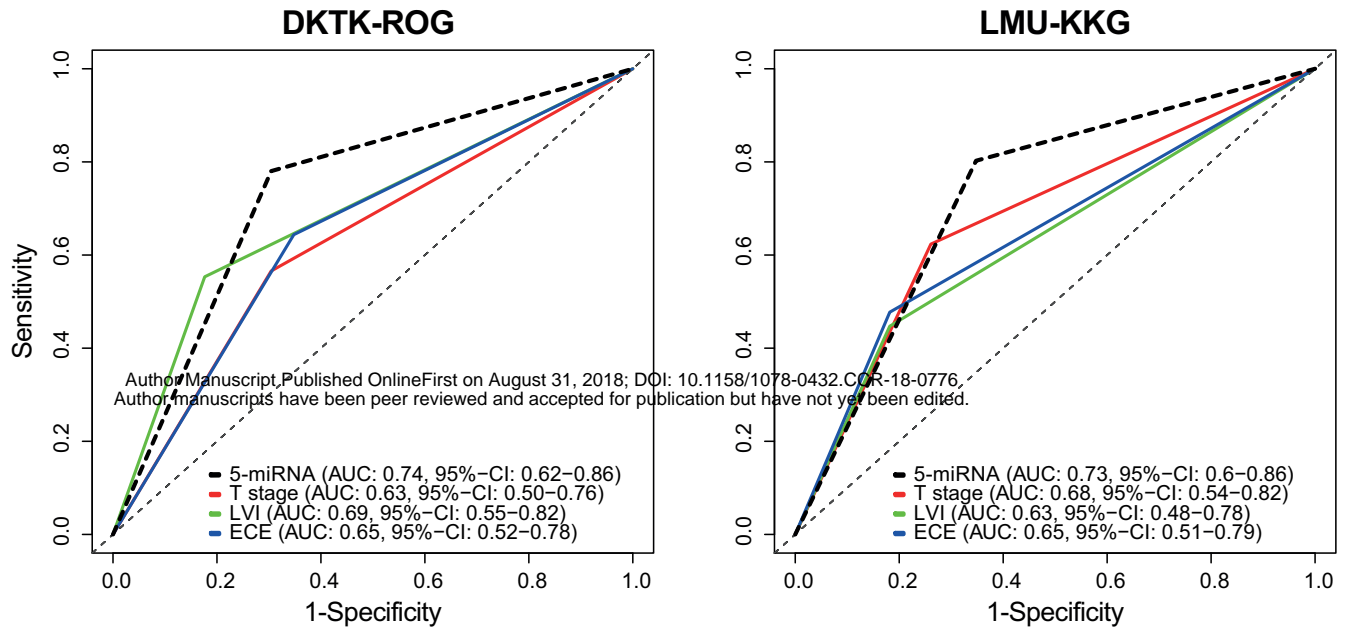
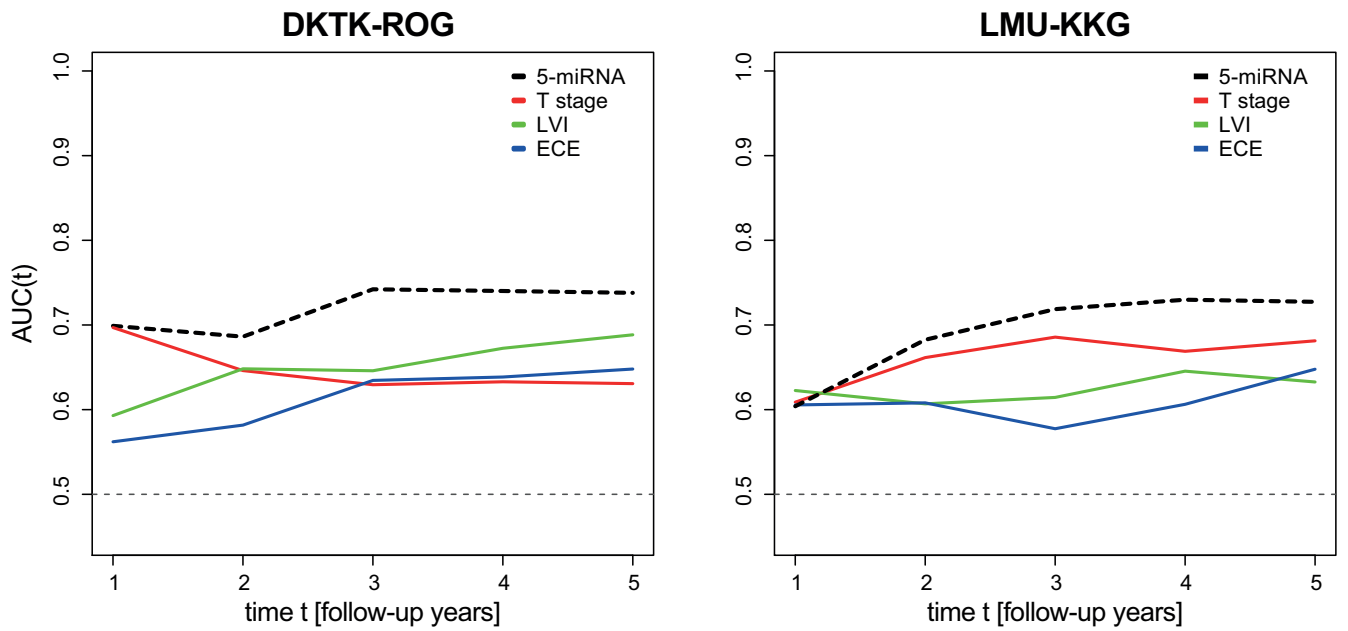


Figure 2

A



B



C

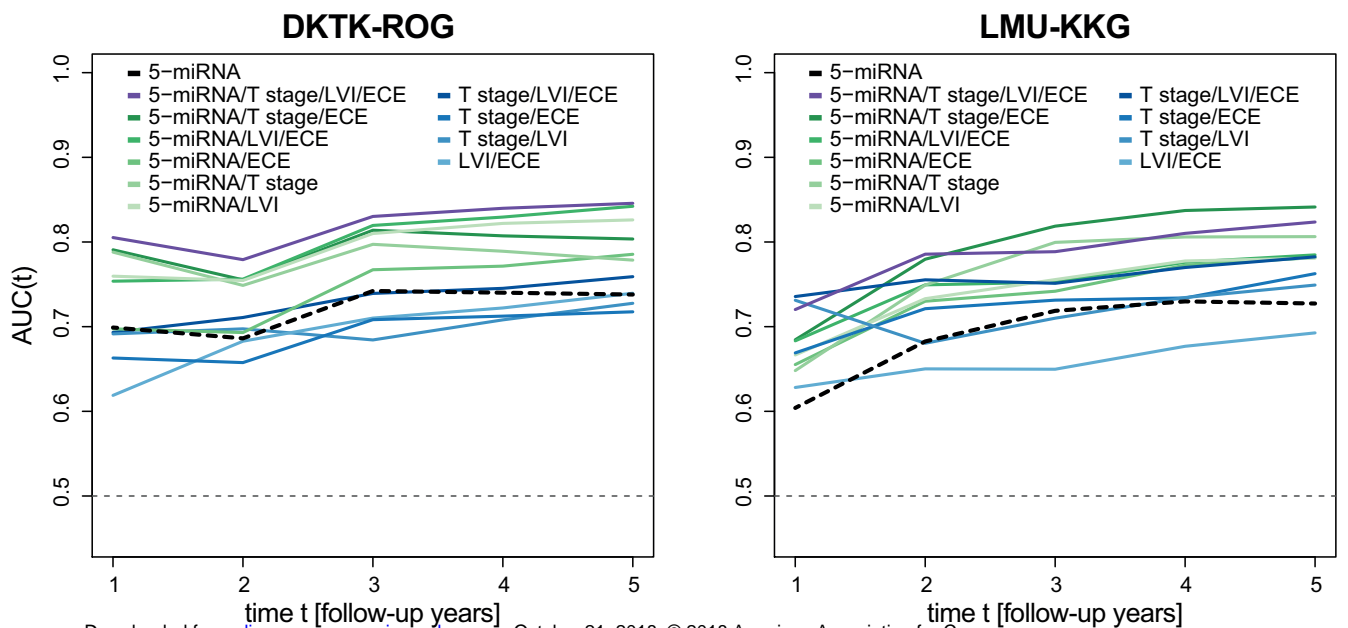
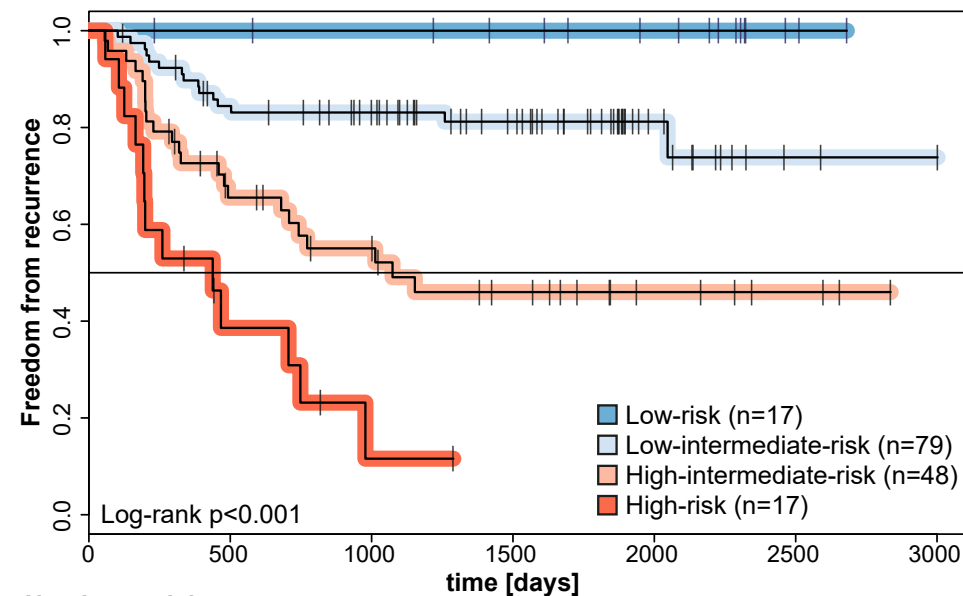
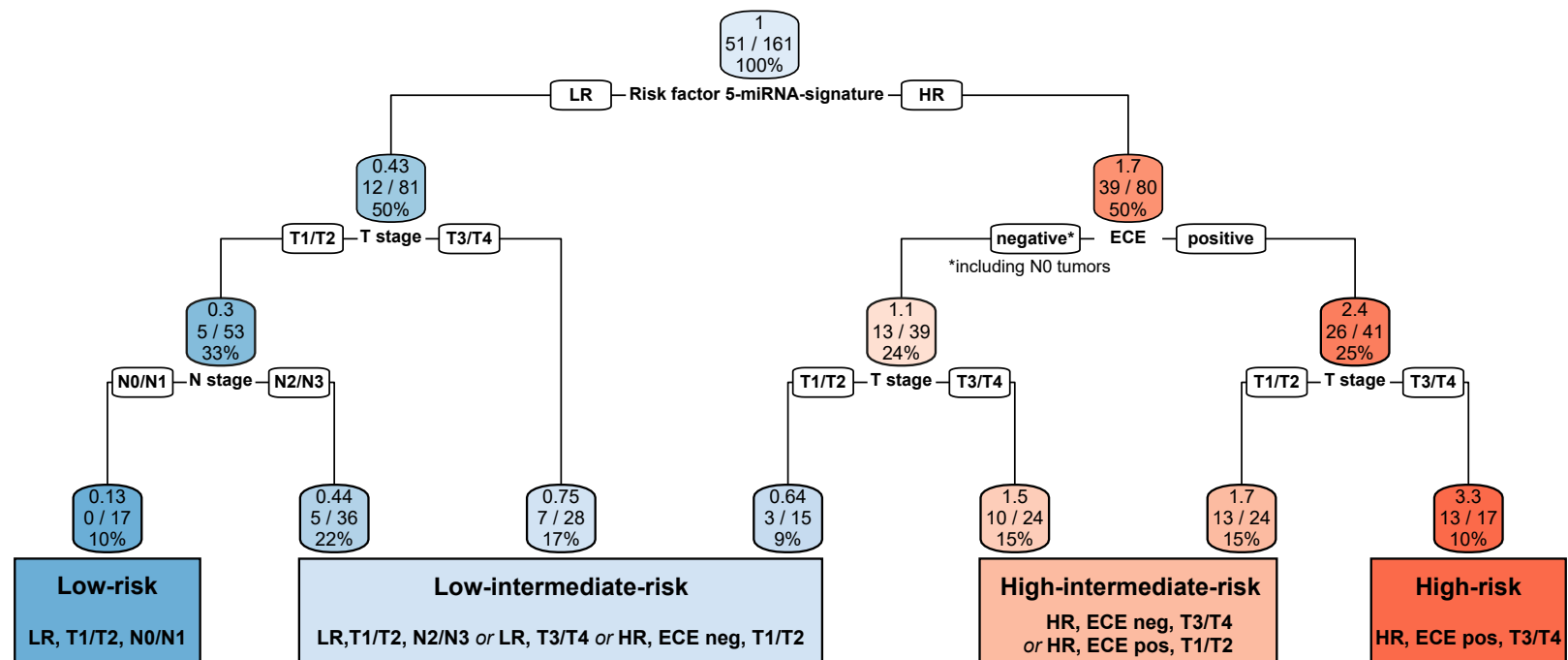


Figure 3



Low-risk vs. low-intermediate-risk

$p = 0.042$

HR NA*

*not estimable because no event in the low-risk group

Low-intermediate-risk vs. high-intermediate-risk

$p < 0.001$

HR 3.17 (95% CI 1.65-6.09)

High-intermediate-risk vs. high-risk

$p = 0.006$

HR 2.55 (95% CI 1.28-5.11)

Number at risk

	0	500	1000	1500	2000	2500	3000
Low-risk (n=17)	17	16	15	13	10	2	
Low-intermediate-risk (n=79)	79	63	54	37	12	2	1
High-intermediate-risk (n=48)	48	27	20	13	6	3	
High-risk (n=17)	17	16	15	13	10	2	

A 4-miRNA signature predicts the therapeutic outcome of glioblastoma

Maximilian Niyazi^{1,8,9}, Adriana Pitea^{2,8}, Michel Mittelbronn³, Joachim Steinbach⁴, Carsten Sticht⁵, Franz Zehentmayr^{1,6}, Daniel Piehlmaier^{2,8}, Horst Zitzelsberger^{2,8}, Ute Ganswindt^{1,8}, Claus Rödel⁷, Kirsten Lauber^{1,8}, Claus Belka^{1,8,9}, Kristian Unger^{2,8}

¹Ludwig-Maximilians-University of Munich, Department of Radiation Oncology, Munich, Germany

²Research Unit of Radiation Cytogenetics, Helmholtz Zentrum München, Neuherberg, Germany

³Institute of Neurology (Edinger Institute), Goethe-University Frankfurt, Frankfurt/Main, Germany

⁴Dr. Senckenbergisches Institut für Neuroonkologie, Klinikum der J.W. Goethe-Universität, Frankfurt/Main, Germany

⁵Zentrum für Medizinische Forschung, Medizinische Fakultät Mannheim, Mannheim, Germany

⁶Department of Radiation Oncology, Paracelsus Medical University, Salzburg, Austria

⁷Department of Radiation Oncology, University Hospital, Frankfurt, Germany

⁸Clinical Cooperation Group Personalized Radiotherapy in Head and Neck Cancer, Helmholtz Zentrum München, Neuherberg, Germany

⁹German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

Correspondence to: Kristian Unger, **email:** unger@helmholtz-muenchen.de

Keywords: glioblastoma, miRNA, signature

Received: February 10, 2016

Accepted: May 22, 2016

Published: June 11, 2016

ABSTRACT

Multimodal therapy of glioblastoma (GBM) reveals inter-individual variability in terms of treatment outcome. Here, we examined whether a miRNA signature can be defined for the *a priori* identification of patients with particularly poor prognosis.

FFPE sections from 36 GBM patients along with overall survival follow-up were collected retrospectively and subjected to miRNA signature identification from microarray data. A risk score based on the expression of the signature miRNAs and cox-proportional hazard coefficients was calculated for each patient followed by validation in a matched GBM subset of TCGA. Genes potentially regulated by the signature miRNAs were identified by a correlation approach followed by pathway analysis.

A prognostic 4-miRNA signature, independent of MGMT promoter methylation, age, and sex, was identified and a risk score was assigned to each patient that allowed defining two groups significantly differing in prognosis (p-value: 0.0001, median survival: 10.6 months and 15.1 months, hazard ratio = 3.8). The signature was technically validated by qRT-PCR and independently validated in an age- and sex-matched subset of standard-of-care treated patients of the TCGA GBM cohort (n=58). Pathway analysis suggested tumorigenesis-associated processes such as immune response, extracellular matrix organization, axon guidance, signalling by NGF, GPCR and Wnt. Here, we describe the identification and independent validation of a 4-miRNA signature that allows stratification of GBM patients into different prognostic groups in combination with one defined threshold and set of coefficients that could be utilized as diagnostic tool to identify GBM patients for improved and/or alternative treatment approaches.

INTRODUCTION

Malignant gliomas account for approximately 70% of primary brain tumors diagnosed in adults. Median age at diagnosis is 64 years with men being more frequently affected than women [1].

Amongst all gliomas, glioblastoma (GBM) is the most common and aggressive form [2]. State-of-the-art treatment of GBM comprises surgical resection and adjuvant radiochemotherapy followed by maintenance chemotherapy. Implementation of temozolomide (TMZ) into the radiochemotherapeutic regime improved 2-year survival rates of patients with newly diagnosed malignant glioma (mainly GBM) from 11% to 27%, 3-year survival rates from 4% to 16%, and 5-year survival rates from 2% to 10% [3]. Unfortunately several phase III trials employing targeted agents such as bevacizumab (AVAglio & RTOG 0825) or cilengitide failed to show an improvement in overall survival [4, 5]. Thus, TMZ-based radiochemotherapy remains standard treatment for GBM. However, prognosis of most GBM patients still remains dismal with a high rate of local recurrence, emphasizing the clear need for further optimization [6]. At present, several strategies are being followed in this regard: Firstly, more elaborate imaging techniques as well as improved image-guidance during radiotherapy are being tested [7, 8]. Secondly, various molecularly designed substances are undergoing pre-clinical and clinical testing for their therapeutic efficacy in combination with radio(chemo)therapy [9, 10]. These targeted treatment approaches require molecular stratification of patients in order to identify the subgroups that can benefit most from a given strategy. Classical radiochemotherapy also displays wide inter-individual differences in terms of response and survival rates [11]. Accordingly, numerous efforts are undertaken in order to characterize the molecular mechanisms orchestrating therapy sensitivity and resistance and to identify prognostic and predictive markers.

So far, only few prognostic factors have been defined for GBM, including age and Eastern Cooperative Oncology Group (ECOG) score. In addition, involvement of the subventricular zone and extent of resection are known to be of prognostic importance [12]. More recently, the first molecular markers have been established. In this regard, methylation of the O6-methylguanine DNA-methyltransferase (MGMT) promoter region was recognized to be of positive predictive value for the efficacy of TMZ-based radiochemotherapy, and molecular profiling of long-term survivors disclosed the positive prognostic value of a proneural-like expression pattern linked to mutations in the genes encoding for isocitrate dehydrogenases 1/2 (IDH1/2) [13].

During the last years, microRNAs (miRNAs) have increasingly received attention. With a high degree of promiscuity miRNAs target and regulate several mRNA species encoding for proteins involved in various signaling pathways [14]. Accumulating evidence indicates that miRNA expression signatures can serve as biomarkers for diagnosis

and risk assessment of diverse malignancies, including GBM [15–20]. Given that the available prognostic markers can segregate GBM patients only to a limited extent, additional markers and/or signatures have to be defined. We focused on miRNA profiles, because the characterization of epigenetic alterations in the field of GBM research has hitherto been underrepresented and miRNA expression is well accessible from clinical routine diagnostic tissue specimen such as formalin-fixed paraffin-embedded (FFPE) tissue sections [21].

We sought to delineate a miRNA expression signature that is of predictive/prognostic value for overall survival in a retrospective cohort of 36 primary GBM patients who underwent adjuvant radiochemotherapy. Applying an iterative forward selection approach on miRNA microarray expression data, we identified a distinct signature comprising 4 miRNAs that was technically confirmed by quantitative real-time PCR (qRT-PCR) and independently validated in an age- and sex-matched data subset of a cohort of GBM patients who received standard-of-care treatment, obtained from The Cancer Genome Atlas (<https://tcga-data.nci.nih.gov>) project [22, 23, 3, 24]. Multivariate analysis revealed this signature to be independent of the MGMT promoter methylation status and of any other prognostic parameters that were available for our dataset.

RESULTS

Characterization of the patient cohort: survival data and univariate analysis

The MGMT promoter methylation status had no statistically significant influence on overall survival (p -value=0.763), although in Kaplan-Meier analysis a trend towards better survival could be observed in MGMT methylated patients. (Supplementary Figure S1) We also did not find statistically significant associations of overall survival with age (p -value=0.053) and sex (p -value=0.222).

Extraction of a low complexity miRNA signature and evaluation of its prognostic significance for overall survival

We analyzed miRNA expression profiles in FFPE samples of our patient cohort and extracted a signature that consisted of the four miRNAs hsa-let-7a-5p, hsa-let-7b-5p, hsa-miR-125a-5p and hsa-miR-615-5p which was statistically significantly associated with overall survival (p -value=0.0048). The median risk score calculated from the expression levels of the signature miRNAs and the corresponding cox-proportional hazard coefficients (Table 1) separated the patients into a high- and a low-risk group. Cox regression analysis of the high- and the low-risk groups revealed a 3.79 fold increased risk of death (95% CI: 2.03-12.85) for the high-risk group compared to the low-risk group (p -value=0.000112). The median survival time was 13.5 months for patients of the high- risk group and 18.4 months

Table 1: Results of multivariate cox-proportional hazard analysis of 4-miRNA risk score, age, sex and MGMT promoter methylation status

Cohort	Model	Hazard ratios of parameters	Confidence intervals of hazard ratios	p-values of contributions of parameters to model	p-value of model
Discovery	4-miRNA risk-score	3.8	1.47-9.75	0.00574	0.00434
	MGMTmeth	0.9	0.39-2	0.7637	0.76298
	4-miRNA risk-score+MGMTmeth	3.8,0.9	1.48-9.81/0.38-1.93	0.00558,0.70124	0.0159
	Sex	1.7	0.72-3.86	0.22874	0.22151
	4-miRNA risk-score+Sex	3.6,1.3	1.36-9.32/0.57-3.16	0.00982,0.50574	0.01367
	Age	1	1-1.07	0.05469	0.05267
	4-miRNA risk-score+Age	3.5,1	1.35-9.12/0.99-1.07	0.00979,0.10002	0.00439
	4-miRNA risk-score+MGMTmeth+Sex+Age	3.3,0.9,1,1.4	1.23-8.62/0.38-2.15/0.99-1.07/0.59-3.44	0.01765,0.82414,0.11408,0.43099	0.02158
	4-miRNA risk-score	2.4	1.03-5.69	0.04207	0.04247
	MGMTmeth_1	0.5	0.18-1.27	0.13798	0.12236
MGMTmeth_2	0.5	0.21-1.32	0.16924	0.15812	
Validation	4-miRNA risk-score+MGMTmeth_1	3.1,0.4	1.24-7.73/0.12-1.01	0.0156,0.05204	0.01532
	4-miRNA risk-score+MGMTmeth_2	2.2,0.6	0.93-5.28/0.24-1.55	0.07096,0.29498	0.07214
	Sex	1.5	0.57-3.8	0.41996	0.41121
	4-miRNA risk-score+Sex	1	1.36-0.9-9.31-7.64	0.00947,0.07811	0.02402
	Age	1	0.94-1.04	0.63117	0.63333
	4-miRNA risk-score+Age	0	1.09-0.93-6.25-1.03	0.03098,0.37596	0.08763
	4-miRNA risk-score+MGMTmeth_1+Sex+Age	4.1,0.5,2.1,1	1.44-11.49/0.16-1.79	0.008,0.30791,0.24802,0.37623	0.03941
4-miRNA risk-score+MGMTmeth_2+Sex+Age	4.5,1,3,1	1.41-14.67/0.34-3.09	0.01139,0.95454,0.08428,0.22869	0.06184	

for patients of the low-risk group, respectively. These results were visualized by Kaplan-Meier overall survival curves (Figure 1B). Univariate testing of the individual miRNAs within the signature revealed p-values in the range between 0.0015 and 0.016, indicating that each single miRNA was able to statistically significantly predict overall survival. Expressions of miRNAs hsa-let-7a-5p, hsa-let-7b-5p and hsa-miR-125a-5p positively correlated with overall survival and that of hsa-miR-615-5p negatively correlated with overall survival. Figure 1A summarizes the survival data of the patients in relation to the calculated risk scores and expression levels. When including MGMT promoter methylation status in a multivariate cox-proportional hazard model, its contribution to the model was not statistically significant, thereby suggesting that the identified miRNA signature performs independently of the MGMT promoter

methylation status. Moreover, the other available clinical parameters, such as sex and age were not associated with the calculated risk-score and also did not statistically significantly contribute to the multivariate model when included. A detailed representation of the results can be found in Table 1. Further, patients in the high-risk group were older compared to that in the low-risk group. Concerning distribution of sex there were no differences between the high- and the low-risk groups (Figure 1C).

Independent in silico validation of the detected miRNA signature

For the purpose of independent validation the miRNA signature was tested in an age-matched miRNA data subset of standard-of-care treated patients (see Supplementary

Table S1) of an independent GBM cohort downloaded from the TCGA database [25]. The high- and the low-risk groups were defined by using the median risk score of the discovery set (0.07811832) to dichotomize the patients of the validation set. The resulting cox-proportional hazard model revealed a hazard ratio of 2.11 (95% CI 1.13-3.91) and a p-value of 0.02 (Figure 2C). Figure 2B summarizes the survival data of the patients of the validation cohort in relation to calculated risk scores and expression levels. Also for the validation cohort no statistically significant association was found between high- and low-risk groups and the parameters age and sex (Figure 2D). Univariate testing of the MGMT promoter methylation status derived from two DNA methylation array

probes that have been shown previously to reliably measure MGMT promoter methylation status for association with overall survival was conducted. No statistically significant association was observed (cg12434587: p-value: 0.122/hazard-ratio: 0.48, cg12981137: p-value: 0.16/hazard-ratio: 0.48) although in Kaplan-Meier analysis a trend towards better survival in MGMT promoter methylation positive was also observable here (Supplementary Figure S2). No differences in the distribution of age and sex were observed in the high- and low-risk groups identified in the validation cohort. Also, including in the multivariate cox model MGMT promoter methylation status did not show statistical significant influence on survival (Table 1).

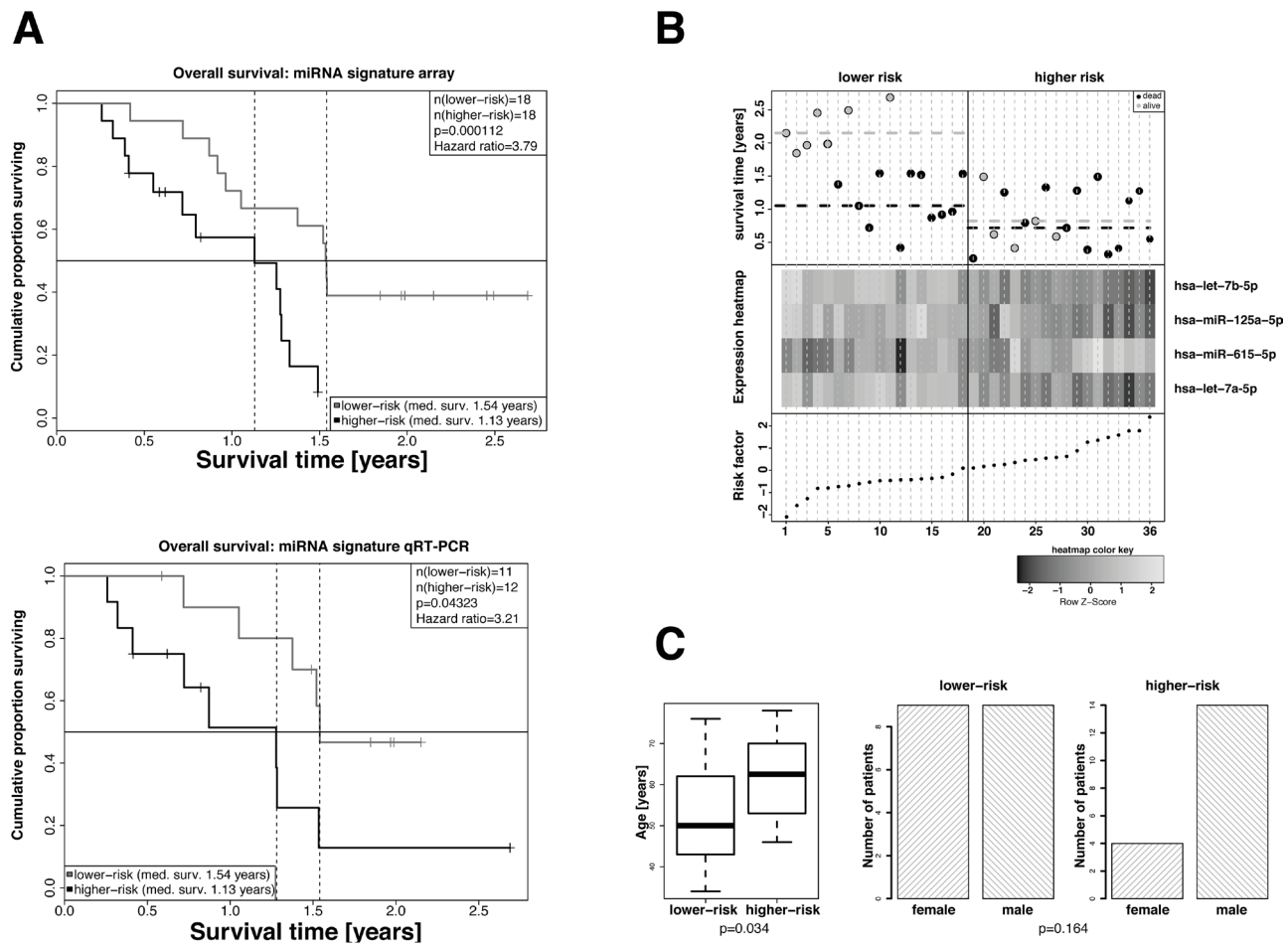


Figure 1: Extraction of a 4-miRNA signature as independent predictive marker for the overall survival of GBM patients in the exploratory cohort. **A.** Kaplan-Meier overall survival analyses of high-risk and low-risk GBM patients. High-risk and low-risk patients were stratified based on the risk factors calculated from the cox-proportional hazard coefficients and the miRNA expression levels as measured in the microarray (left panel, 35 patients) or by qRT-PCR analyses (right panel, 19 patients). Hazard ratios and p-values were calculated by log-rank test. **B.** Overall survival (top panel), hierarchical cluster heatmap of miRNA array expression levels (middle panel), and risk factors calculated on the basis of miRNA expression values and cox-proportional hazard coefficients (bottom panel) for all patients. miRNAs hsa-let-7a-5p, hsa-let-7b-5p and hsa-miR-125a-5p in patients of the higher-risk group show a tendency towards lower expression and that of hsa-miR-615-5p a tendency towards higher expression. The median risk factor value was used to classify high-risk and low-risk patients. **C.** Distribution of age (left panel) and sex (right panel) in high-risk and low-risk GBM patients. Statistical comparison was performed by Student's t-test or Fisher's exact test. The patients of the lower-risk group were statistically significantly older compared with that of the lower-risk group. The differences in the numbers of male and female patients of the lower- and higher-risk groups were not statistically significant.

Technical validation of signature by qRT-PCR

In order to technically validate the 4-miRNA signature and to support potential applicability in clinical routine diagnostics, we measured the expression of the four miRNAs in a subset of samples (n=23), for which residual material was available by qRT-PCR. Analogous cox-proportional hazard analysis with the qRT-PCR data confirmed the results obtained with the miRNA array data. Patients of the high-risk group revealed significantly impaired overall survival (p-value=0.043) and a hazard-ratio of 3.21 (95% CI 1.02-10.16) as compared to patients of the low-risk group. (Figure 1A).

miRNA-mRNA correlation and pathway enrichment analysis

For hsa-let-7b-5p we identified 104 significantly correlating genes (53 negative and 51 positive correlations), for hsa-miR-125a-5p 112 genes (35 negative and 77 positive correlations), for hsa-miR-615-5p 26 genes (10 negative and 16 positive correlations) and for hsa-let-7a-5p 412 genes (245 negative and 167 positive correlations). The overlap between genes correlating with expression of the signature miRNAs was sparse (Supplementary Figure S1). Heatmaps of the top 25 miRNA-mRNA correlations with regard to absolute correlation coefficients are depicted in Figure 3.

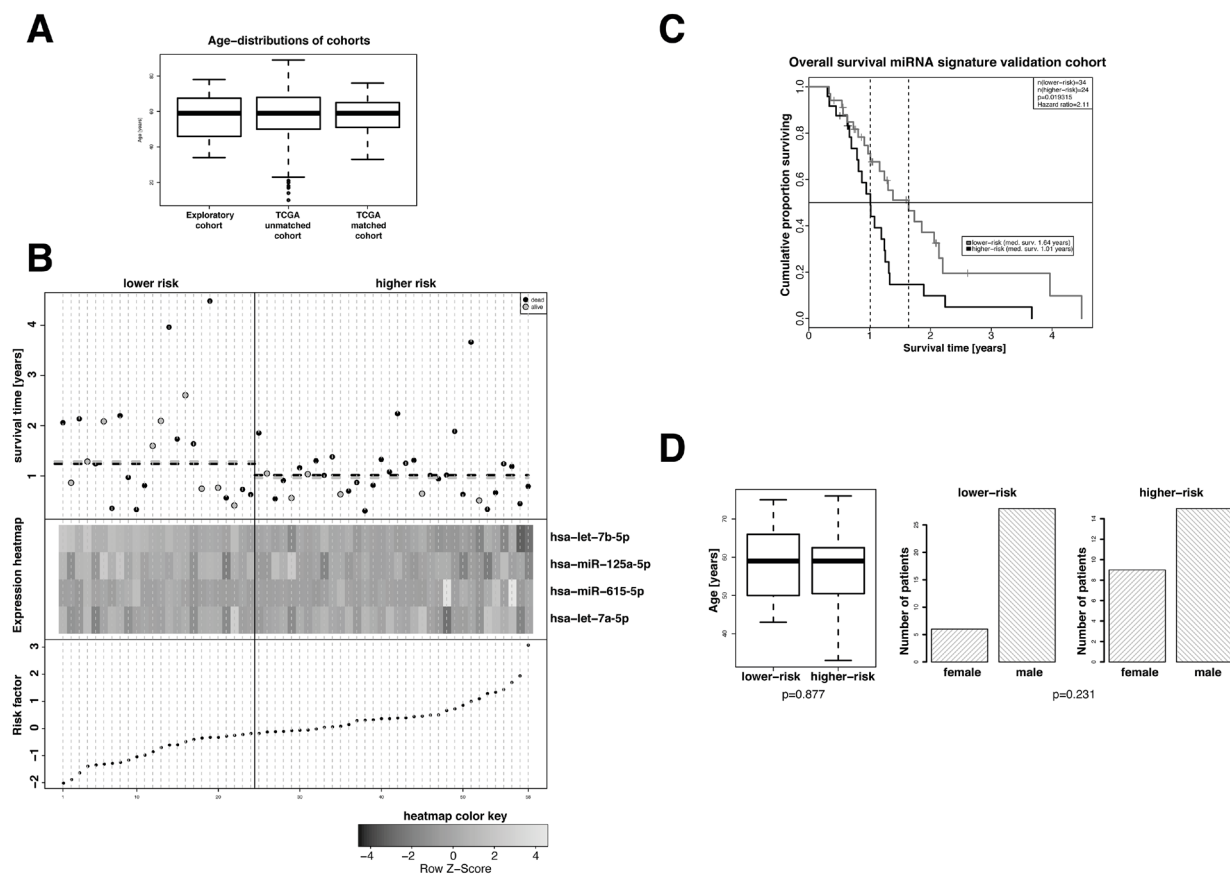


Figure 2: Evaluation of the prognostic value of the extracted 4-miRNA signature in an age- and sex-matched subgroup of standard-of-care treated patients of the TCGA GBM dataset. **A.** Age distribution in the exploratory cohort and the TCGA GBM cohort before and after age matching. **B.** Overall survival (top panel), hierarchical cluster heat map of miRNA expression levels (middle panel), and risk factors for patients of the age- and sex-matched TCGA GBM cohort. The median risk factor value was used to classify high-risk and low-risk patients. miRNAs hsa-let-7a-5p, hsa-let-7b-5p and hsa-miR-125a-5p in patients of the higher-risk group show a slight tendency towards lower expression and that of hsa-miR-615-5p a slight tendency towards higher expression. **C.** Kaplan-Meier overall survival analyses of high-risk and low-risk standard-of-care treated patients of the age- and sex-matched TCGA GBM cohort. Classification of high-risk and low-risk patients was performed on the basis of the risk factors calculated from the cox-proportional hazard coefficients (Table 2) and the miRNA expression levels. Hazard ratios and p-values were calculated by log-rank test. **D.** Distribution of age (left panel) and sex (right panel) in high-risk and low-risk patients of the age- and sex-matched TCGA GBM cohort. Student's t-test and Fisher's exact test were employed for statistical comparison as depicted.

Interestingly, whereas hsa-let-7b-5p, hsa-miR-125a-5p, and hsa-let-7a-5p displayed predominantly negative correlations among the top correlations as to be expected, hsa-miR-615-5p also showed positive correlations. All genes with significant correlations (Pearson correlation test, p -value < 0.01) were combined into one list of genes ($n=654$; Supplementary Table S2) and subjected to pathway enrichment analysis. In total, 28 statistically significant pathways were identified (Supplementary Table S3), and the top ten of these (i.e. Transmembrane transport of small molecules, Innate Immune System, Extracellular matrix organization, Axon guidance, Signalling by NGF, Developmental Biology, Neuronal System, GPCR downstream signaling, Signaling by

GPCR and Signaling by Wnt) were considered for interpretation of results.

DISCUSSION

GBM patients, who receive surgical resection and postoperative radio(chemo)therapy, reveal profound differences in terms of overall survival which motivated us to search for a miRNA signature that allows the identification of patients with specifically poor prognosis independently of any other outcome-associated parameters. Moreover, we intended to investigate such a signature with regard to the molecular mechanisms potentially underlying the poor outcome of GBM patients.

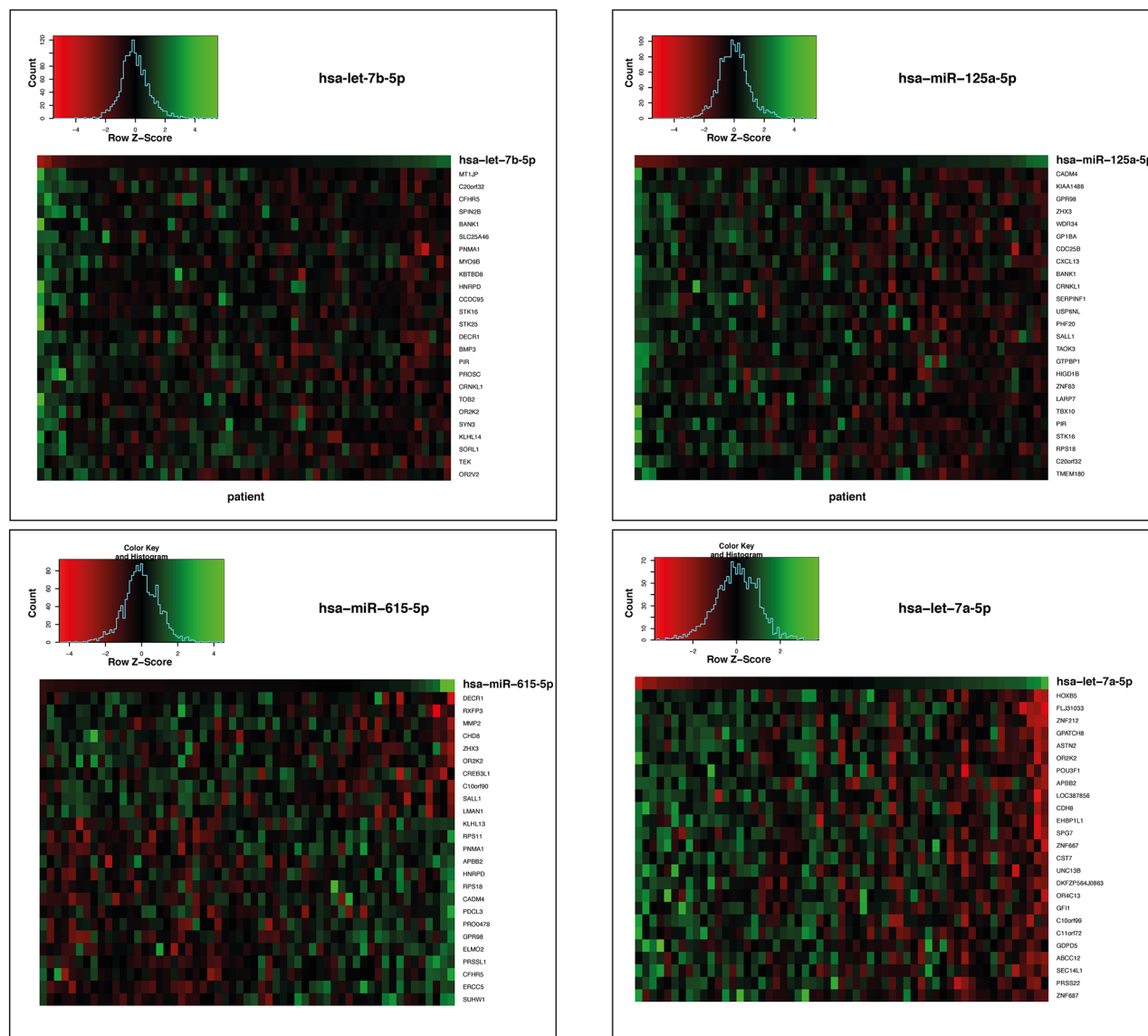


Figure 3: Heatmaps of the gene expressions correlating with the 4 miRNAs hsa-let-7b-5p, hsa-miR-125a-5p, hsa-miR-615-5p and hsa-let-7a-5p the age- and sex-matched TCGA GBM cohort of standard-of-care treated patients. Genes whose expression levels statistically significantly correlated ($p < 0.01$) with the respective miRNA expression levels are shown. Every column represents an individual patient. Data are ordered from left to right by increasing miRNA expression.

To this end, we performed miRNA microarray analysis followed by low-complexity miRNA signature identification. We could extract a 4-miRNA signature which, with high statistical significance, allowed differentiating between high- and low-risk GBM patients independently of the MGMT methylation status. Technical validation by qRT-PCR confirmed the microarray data results. Most importantly, the prognostic value of the signature could be confirmed by independent validation in a large subset of the TCGA study on GBM [25].

At present, the role of individual miRNAs in GBM is poorly understood. miRNAs are small non-coding regulatory RNAs that reduce stability and/or inhibit translation of target mRNAs with full or partial sequence-complementarity [14]. In this sense, they are important post-transcriptional regulators and play essential roles in the pathogenesis, development, and progression of cancer as well as in the response to therapy [26–28].

It has been shown that GBMs display distinct miRNA expression signatures, and several studies have linked these miRNA alterations to hallmarks of GBM, including proliferation, survival, invasion, angiogenesis, and stem cell-like behavior [29]. Moreover, resistance to TMZ might be associated with miRNA deregulation [30]. In this regard, Ciafre et al. studied the expression of 245 microRNAs in GBM in comparison to normal brain tissue using a microarray technique [31] in comparison to normal brain tissue. This approach enabled the identification of miRNAs whose expression levels were significantly altered in tumor tissue compared to peripheral brain tissue of the same patient, including miR-221, which was strongly up-regulated in GBM, and a set of brain-enriched miRNAs (miR-128, miR-181a, miR-181b, and miR-181c), which were down-regulated in GBM [32]. Very recently, a number of prognostic miRNA signatures have been reported for GBM [33–35, 15, 36, 37, 18–20, 38, 39]. We compared these signatures with our signature in terms of complexity, independent validation, and the approach used for identification of the signature. One important feature of molecular signatures is their level of complexity (i.e. the number of miRNAs), which should be most optimal with regard to prediction performance but at the same time should not overfit the data. For data sets with moderate dimensionality such as miRNA microarray data sets with typically a few hundreds of miRNAs expressed, the number of features contained in a signature should be of low complexity and in the range of smaller than 10. From the above cited studies, only five extracted a miRNA signature with low complexity that have been subsequently validated in an independent cohort [15, 37, 18–20]. Cheng et al. [15] focused on MGMT promoter methylation positive tumors only and defined a 5-miRNA signature that was validated in an appropriate subset of the so-called Chinese Glioma Genome Atlas. In contrast, our signature was developed using data from both MGMT promoter methylation-positive and -negative tumors. The signatures

described by Li et al. were developed for each of the five molecular GBM subtypes as defined by transcriptomic profiling of TCGA GBM cases [37, 40]. Only the signature for the ‘mesenchymal’ subtype consisting of five miRNAs was independently validated in a set of GBM tissues [37]. Our signature, in contrast, is not limited to this molecular subtype only. The small-noncoding RNAs described in Manterola et al. [18] allow molecular diagnosis of GBM using blood serum samples but the authors did not show usability with regard to outcome prediction whilst our signature was particularly developed for the purpose of predicting survival outcome. In a study by Shou et al. [20] three miRNAs were presented that statistically significantly allow differentiation into groups of patients with favorable and unfavorable prognosis. However, this approach was limited to separate and univariate analysis for each of the three miRNAs and the study did not include an independent validation of the results. The report by Sana et al. is most comparable to our study and introduced a 6-miRNA-signature which was also validated in the TCGA GBM data set [19]. However, contrary to our study Sana et al. used different thresholds for the calculated risk scores of the discovery and the validation set. Most importantly, the thresholds were chosen in such a way that they statistically significantly separated the two resulting groups of patients [19]. This can be interpreted as a technical drawback, and the approach has to be considered as a biased one. In strong contrast, we applied the same cox-proportional hazard coefficients and the same risk score thresholds to all of our three datasets (discovery Febit microarray, discovery qRT-PCR and validation Agilent microarray). This renders our 4-miRNA-signature, in principle, applicable to any other data set regardless of the platform the data were generated with.

Comparing our signature with that of the above mentioned studies did show no overlapping miRNAs of our signature with of the published ones. This can be explained by the fact that signature identification is very much dependent on the methodology used, the dataset with regard to the specific selection criteria of patients and the platform that is used for measurements. Since all mentioned studies vary with regard to these parameters one could not expect overlap of our signature with the published ones or overlap between the published ones.

Besides its prognostic value, it is of major interest to understand the impact of the 4-miRNA-signature on the biological characteristics of GBM. Our panel of miRNAs consists of hsa-let-7a-5p, hsa-let-7b-5p, hsa-miR-125a-5p and hsa-miR-615-5p. Two miRNAs of the signature belong to the let-7 family, which is very well known for its tumor suppressor function in various cancer entities [41]. The two let-7 miRNAs hsa-let-7a-5p and hsa-let-7b-5p showed a tendency towards higher expression levels in the low-risk compared to the high-risk group of patients, which is in line with the concept of their involvement in tumor suppression. miR-125a-

5p was described as a tumor suppressor in GBM only recently. It is engaged in the repression of target genes of the TAZ (transcriptional co-activator with PDZ-binding motif) transcription factor, including connective tissue growth factor (CTGF) and survivin [42]. Our analysis revealed a tendency towards higher miR-125a-5p expression levels in the low-risk compared to the high-risk group of patients, again supporting its role in tumor suppression. hsa-miR-615-5p was described to act as tumor suppressor in pancreatic ductal adenocarcinoma [17]. In our analyses, a clear tendency towards higher or lower expression levels of hsa-miR-615-5p in the low- and high-risk group of patients was not observable. Therefore, conclusions concerning its tumorsuppressive role in GBM cannot be drawn. Overall, our 4-miRNA-signature reflects trends of higher expression levels of tumor suppressive miRNAs in low-risk GBMs, supporting the notion that these GBMs exhibit a lower degree of malignancy due to operational tumor suppressive mechanisms. In order to gain insights into the putative functional role of the four signature miRNAs, we conducted miRNA-transcriptome correlation analyses and obtained 654 genes, whose expression levels were positively or negatively correlated with that of the miRNAs. We deliberately followed this approach to identify direct and indirect regulatory effects of the signature miRNAs on the transcriptome. An alternative approach would have been to utilize miRNA target prediction. This, however, relies strongly on the prediction algorithm and the database that are used, and databases providing information on in vitro validated miRNA targets are still limited with regard to the number of miRNAs they provide information on [43]. The genes that were identified in our correlation approach were subjected to pathway enrichment analyses, which disclosed the top ten pathways Transmembrane transport of small molecules, Innate Immune System, Extracellular matrix organization, Axon guidance, Signalling by NGF, Developmental Biology, Neuronal System, GPCR downstream signaling, Signaling by GPCR and Signaling by Wnt all of which very well known in the context of glioblastoma tumorigenesis. These results suggest that our 4-miRNA signature regulates genes that are well known to be involved in the tumorigenesis, progression and migration of GBM and may potentially act as druggable targets in an alternative treatment approach.

CONCLUSIONS

In the present study, we extracted and validated a 4-miRNA-signature, which allows to differentiate GBM patients undergoing surgical resection and subsequent radio(chemo)therapy with favorable and poor prognosis. This signature may serve as a potential new marker for patient stratification independent of the MGMT methylation status. It may furthermore pave the way for personalized treatment approaches based on measurements that are

well feasible in GBM biopsies in the clinical routine. Patients with a high-risk score are likely not profiting from standard-of-care treatment and therefore the 4-miRNA signature could be used to identify patients who require therapy intensification. Compared to existing GBM miRNA signatures the herein presented signature is of lower complexity, was independently validated, and appears to be in principle applicable to any data set containing expression values of the four signature miRNAs regardless of the platform they were generated with.

MATERIALS AND METHODS

For a detailed description of Material and Methods sections 'Patient characteristics', 'miRNA array analysis', 'Technical validation of the 4-miRNA signature by qRT-PCR' and 'miRNA-mRNA correlation and gene set enrichment analysis' see Supplementary Data.

Patient characteristics

We examined FFPE tissue samples of a non-selected, retrospective cohort of patients who were consecutively treated at the University hospital Frankfurt between 1/2009 and 12/2010. Ethics approval (4/09) was given by the ethics committee of the medical faculty of the Johann-Goethe University (Frankfurt am Main, Germany). Only patients who underwent surgical resection and post-operative radio(chemo)therapy were included into the analyses. Patients underwent resection and adjuvant radiotherapy, regularly combined with TMZ according to the EORTC/NCIC26981/22981-NCIC CE3 protocol if no contraindications were present (for details see Table 2) [3, 24]. The median overall survival time of this patient cohort was 1.28 years with a median follow-up of 1.99 years (95%-CI, 634 - 816 days). MGMT promoter methylation status was available for all 36 cases (see Table 1). Karnofsky performance status (KPS) score and associated recursive partitioning analysis (RPA) class had not been collected systematically, and no data on the extent of resection was available. For independent validation, the miRNA expression dataset from an age- and sex-matched subset (n=58) of the TCGA GBM cohort (n=357) was used. The subset resulted after adjusting the distribution of age of

Table 2: Cox-proportional hazard coefficients used in risk score calculation

miRNA	coefficient
hsa-let-7b-5p	-0.9669152
hsa-miR-125a-5p	-0.2821517
hsa-miR-615-5p	0.3254795
hsa-let-7a-5p	0.5059587

Table 3: Clinical characteristics of discovery cohort

Characteristic	Patients (N=36)
Sex	
Male	23 (63.9 %)
Female	13 (36.1 %)
Median Age [y]	59 (34-78)
Age Category	
< 50 y	13 (36.1 %)
≥ 50 y	23 (63.9 %)
MGMT promoter methylation status	
Methylated	18 (50.0 %)
Unmethylated	18 (50.0 %)
Secondary Malignisation	
Yes	12 (33.3 %)
No	24 (66.7 %)
Concomitant Temozolomide	
Yes	32 (88.9 %)
No	1 (2.8 %)
Unknown	3 (8.3 %)
Median adjuvant TMZ cycles	6 (0–20)

the whole TCGA GBM dataset to that of our discovery cohort (Table 3) and only selecting patients that were treated according to standard-of-care.

miRNA array analysis

miRNA analysis was carried out using the Geniom Biochip MPEA homo sapiens biochips containing 1223 miRNA probes (CBC, Heidelberg, Germany). FFPE sample preparation, hybridization, washing and scanning of arrays was performed as described previously [44]. We applied 'winsorized mean' scaling on normalized data with exclusion of 30% of the top and bottom values.

TCGA glioblastoma miRNA data set

The validation data set was constructed from miRNA microarray profiles of the matched subset of patients from the TCGA GBM cohort. Data were generated by the University of North Carolina Cancer Genomic Characterization Center (CGCC) using the Agilent 8x15K Human miRNA-specific microarray platform [22]. For the analysis level 3 data were used and in order to allow comparability of the data set with the discovery data set scaling with 'winsorized mean' was applied as described above.

miRNA signature robust selection

In order to search for a miRNA signature in miRNA expression data set of the discovery cohort associated with patient survival, the R package rbsurv was used [45]. The forward-selection algorithm implemented in the package computed the partial likelihood of the Cox model for a sequential selection of miRNAs. The best performing model was chosen based on the Akaike Information Criterion (AIC), which allowed to determine the best trade-off between the complexity of a model and its goodness of fit.

Calculation of risk scores

The Cox model coefficients (Table 2) were multiplied with the scaled expression values of appropriate miRNAs and the products were summed up resulting in an individual risk score for each patient. The median risk score of all patients (0.07811832) was used as a cut-off for defining a high-risk (> median risk score) and a low-risk group (< median risk score). Subsequently, the log-rank test was used to test whether the differences in overall survival times between the resulting two groups were statistically significant (p-value threshold: 0.05). Further, Kaplan-Meier survival curves were plotted for the two groups and the hazard ratio was calculated. The influence

of the available known risk factors age, sex, and MGMT promoter methylation status was assessed univariately and by inclusion into the multivariate cox-proportional hazard model.

Independent *in silico* validation of the 4-miRNA signature

For each of the 58 included TCGA GBM patients (Supplementary Table S1) we calculated a risk score by building the sum of the products of the expressions of the four miRNAs of the signature and the coxproportional hazard coefficients obtained from the initial dataset for each of the miRNAs (Table 2).

The patients were assigned to high- and low-risk groups by using the same threshold (0.07811832) that was defined for the discovery data set. The resulting two groups were tested for differential survival outcome using log-rank test.

Technical validation of the 4-miRNA signature by qRT-PCR

MiScript primer assays (QIAGEN, MD, USA) for the four miRNAs of the signature were used for relative quantification along with a reference assay for the small RNA SNORD61. The relative expression values in combination with the cox-proportional hazard coefficients were used to calculate a risk score for every patient. The patients were dichotomized into a low- and high-risk group using the risk score threshold from the discovery cohort (0.07811832) and the resulting two groups were tested for differences in overall survival using log-rank test of the resulting cox-proportional hazard model.

miRNA-mRNA correlation and gene set enrichment analysis

In order to investigate the impact of the four signature miRNAs on the transcriptome level we downloaded the transcriptome data (level 3) of the cases matching the miRNAs data set (n=132) from the TCGA database and calculated correlations between the four signature miRNAs and expression levels of all genes. Genes that statistically significantly correlated with the signature miRNAs were subjected to pathway enrichment analysis.

MGMT promoter methylation typing

For the discovery cohort determination of MGMT promoter methylation was performed using both methylation-specific PCR and sequencing analysis as described previously [46, 47].

For the TCGA validation cohort no systematic assessment of the MGMT promoter methylation status was available. However, in order to determine methylation of

the MGMT promoter we followed an approach published by Bady et al. from methylation array data [48]. The resulting MGMT promoter-positive and -negative groups were then subsequently tested for association with survival univariately and multivariately.

ACKNOWLEDGMENTS

We thank all co-workers of the Clinical Cooperation Group Personalized Radiotherapy in Head and Neck Cancer, a collaborative effort between the Research Unit Radiation Cytogenetics at the Helmholtz Zentrum München and the Department of Radiation Oncology of the University Munich for technical, scientific and conceptual support.

CONFLICTS OF INTEREST

The authors have no potential conflicts of interest to disclose.

REFERENCES

1. Fisher JL, Schwartzbaum JA, Wrensch M and Wiemels JL. Epidemiology of brain tumors. *Neurol Clin.* 2007; 25:867-890, vii.
2. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW and Kleihues P. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 2007; 114:97-109.
3. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 2009; 10:459-466.
4. Chinot OL, de La Motte Rouge T, Moore N, Zeaiter A, Das A, Phillips H, Modrusan Z and Cloughesy T. AVAglio: Phase 3 trial of bevacizumab plus temozolomide and radiotherapy in newly diagnosed glioblastoma multiforme. *Adv Ther.* 2011; 28:334-340.
5. Gilbert MR, Dignam JJ, Armstrong TS, Wefel JS, Blumenthal DT, Vogelbaum MA, Colman H, Chakravarti A, Pugh S, Won M, Jeraj R, Brown PD, Jaeckle KA, et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. *N Engl J Med.* 2014; 370:699-708.
6. Niyazi M, Jansen NL, Rottler M, Ganswindt U and Belka C. Recurrence pattern analysis after re-irradiation with bevacizumab in recurrent malignant glioma patients. *Radiat Oncol.* 2014; 9:299.
7. Cyran CC, Paprottka PM, Eisenblatter M, Clevert DA, Rist C, Nikolaou K, Lauber K, Wenz F, Hausmann D, Reiser MF, Belka C and Niyazi M. Visualization, imaging and new

- preclinical diagnostics in radiation oncology. *Radiat Oncol.* 2014; 9:3.
8. Niyazi M, Siefert A, Schwarz SB, Ganswindt U, Kreth FW, Tonn JC and Belka C. Therapeutic options for recurrent malignant glioma. *Radiother Oncol.* 2011; 98:1-14.
 9. Beal K, Abrey LE and Gutin PH. Antiangiogenic agents in the treatment of recurrent or newly diagnosed glioblastoma: analysis of single-agent and combined modality approaches. *Radiat Oncol.* 2011; 6:2.
 10. Flieger M, Ganswindt U, Schwarz SB, Kreth FW, Tonn JC, la Fougere C, Ertl L, Linn J, Herrlinger U, Belka C and Niyazi M. Re-irradiation and bevacizumab in recurrent high-grade glioma: an effective treatment option. *J Neurooncol.* 2014; 117:337-345.
 11. Frenel JS, Botti M, Loussouarn D and Campone M. Prognostic and predictive factors for gliomas in adults. *Bull Cancer.* 2009; 96:357-367.
 12. Adeberg S, Bostel T, Konig L, Welzel T, Debus J and Combs SE. A comparison of long-term survivors and short-term survivors with glioblastoma, subventricular zone involvement: a predictive factor for survival? *Radiat Oncol.* 2014; 9:95.
 13. Kim HJ, Kim JH, Chie EK, Young PD, Kim IA and Kim IH. DNMT (DNA methyltransferase) inhibitors radiosensitize human cancer cells by suppressing DNA repair activity. *Radiat Oncol.* 2012; 7:39.
 14. Ha M and Kim VN. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol.* 2014; 15:509-524.
 15. Cheng W, Ren X, Cai J, Zhang C, Li M, Wang K, Liu Y, Han S and Wu A. A five-miRNA signature with prognostic and predictive value for MGMT promoter-methylated glioblastoma patients. *Oncotarget.* 2015; 6:29285-29295. doi: 10.18632/oncotarget.4978.
 16. De Smaele E, Ferretti E and Gulino A. MicroRNAs as biomarkers for CNS cancer and other disorders. *Brain Res.* 2010; 1338:100-111.
 17. Gao W, Gu Y, Li Z, Cai H, Peng Q, Tu M, Kondo Y, Shinjo K, Zhu Y, Zhang J, Sekido Y, Han B, Qian Z and Miao Y. miR-615-5p is epigenetically inactivated and functions as a tumor suppressor in pancreatic ductal adenocarcinoma. *Oncogene.* 2015; 34:1629-1640.
 18. Manterola L, Guruceaga E, Gallego Perez-Larraya J, Gonzalez-Huarriz M, Jauregui P, Tejada S, Diez-Valle R, Segura V, Sampron N, Barrera C, Ruiz I, Agirre A, Ayuso A, et al. A small noncoding RNA signature found in exosomes of GBM patient serum as a diagnostic tool. *Neuro Oncol.* 2014; 16:520-527.
 19. Sana J, Radova L, Lakomy R, Kren L, Fadrus P, Smrcka M, Besse A, Nekvindova J, Hermanova M, Jancalck R, Svoboda M, Hajduch M, Slampa P, Vyzula R and Slaby O. Risk Score based on microRNA expression signature is independent prognostic classifier of glioblastoma patients. *Carcinogenesis.* 2014; 35:2756-2762.
 20. Shou J, Gu S and Gu W. Identification of dysregulated miRNAs and their regulatory signature in glioma patients using the partial least squares method. *Exp Ther Med.* 2015; 9:167-171.
 21. Xi Y, Nakajima G, Gavin E, Morris CG, Kudo K, Hayashi K and Ju J. Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. *RNA.* 2007; 13:1668-1674.
 22. Brennan CW, Verhaak RG, McKenna A, Campos B, Nounshmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, et al. The somatic genomic landscape of glioblastoma. *Cell.* 2013; 155:462-477.
 23. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 2012; 487:330-337.
 24. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J, Janzer RC, Ludwin SK, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med.* 2005; 352:987-996.
 25. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008; 455:1061-1068.
 26. Chen G, Zhu W, Shi D, Lv L, Zhang C, Liu P and Hu W. MicroRNA-181a sensitizes human malignant glioma U87MG cells to radiation by targeting Bcl-2. *Oncol Rep.* 2010; 23:997-1003.
 27. Malzkorn B, Wolter M and Reifenberger G. (2009). MicroRNA: Biogenesis, Regulation, and Role in Primary Brain Tumors. In: Erdmann AV, Reifenberger G and Barciszewski J, eds. *Therapeutic Ribonucleic Acids in Brain Tumors.* (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 327-354.
 28. Malzkorn B, Wolter M, Liesenberg F, Grzendowski M, Stuhler K, Meyer HE and Reifenberger G. Identification and functional characterization of microRNAs involved in the malignant progression of gliomas. *Brain Pathol.* 2010; 20:539-550.
 29. Asadi-Moghaddam K, Chiocca EA and Lawler SE. Potential role of miRNAs and their inhibitors in glioma treatment. *Expert Rev Anticancer Ther.* 2010; 10:1753-1762.
 30. Ujifuku K, Mitsutake N, Takakura S, Matsuse M, Saenko V, Suzuki K, Hayashi K, Matsuo T, Kamada K, Nagata I and Yamashita S. miR-195, miR-455-3p and miR-10a(*) are implicated in acquired temozolomide resistance in glioblastoma multiforme cells. *Cancer Lett.* 2010; 296:241-248.
 31. Ciafre SA, Galardi S, Mangiola A, Ferracin M, Liu CG, Sabatino G, Negrini M, Maira G, Croce CM and Farace MG. Extensive modulation of a set of microRNAs in primary glioblastoma. *Biochem Biophys Res Commun.* 2005; 334:1351-1358.

32. Conti A, Aguenouz M, La Torre D, Tomasello C, Cardali S, Angileri FF, Maio F, Cama A, Germano A, Vita G and Tomasello F. miR-21 and 221 upregulation and miR-181b downregulation in human grade II-IV astrocytic tumors. *J Neurooncol.* 2009; 93:325-332.
33. Agrawal R, Pandey P, Jha P, Dwivedi V, Sarkar C and Kulshreshtha R. Hypoxic signature of microRNAs in glioblastoma: insights from small RNA deep sequencing. *BMC Genomics.* 2014; 15:686.
34. Barbano R, Palumbo O, Pasculli B, Galasso M, Volinia S, D'Angelo V, Icolaro N, Coco M, Dimitri L, Graziano P, Copetti M, Valori VM, Maiello E, Carella M, Fazio VM and Parrella P. A miRNA signature for defining aggressive phenotype and prognosis in gliomas. *PLoS One.* 2014; 9:e108950.
35. Chen T, Wang XY, Li C and Xu SJ. Downregulation of microRNA-124 predicts poor prognosis in glioma patients. *Neurol Sci.* 2015; 36:131-135.
36. Huang YT, Hsu T, Kelsey KT and Lin CL. Integrative analysis of micro-RNA, gene expression, and survival of glioblastoma multiforme. *Genet Epidemiol.* 2015; 39:134-143.
37. Li R, Gao K, Luo H, Wang X, Shi Y, Dong Q, Luan W and You Y. Identification of intrinsic subtype-specific prognostic microRNAs in primary glioblastoma. *J Exp Clin Cancer Res.* 2014; 33:9.
38. Wang Z, Bao Z, Yan W, You G, Wang Y, Li X and Zhang W. Isocitrate dehydrogenase 1 (IDH1) mutation-specific microRNA signature predicts favorable prognosis in glioblastoma patients with IDH1 wild type. *J Exp Clin Cancer Res.* 2013; 32:59.
39. Xiong J, Bing Z, Su Y, Deng D and Peng X. An integrated mRNA and microRNA expression signature for glioblastoma multiforme prognosis. *PLoS One.* 2014; 9:e98419.
40. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, 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.
41. Boyerinas B, Park SM, Hau A, Murmann AE and Peter ME. The role of let-7 in cell differentiation and cancer. *Endocr Relat Cancer.* 2010; 17:F19-36.
42. Yuan J, Xiao G, Peng G, Liu D, Wang Z, Liao Y, Liu Q, Wu M and Yuan X. MiRNA-125a-5p inhibits glioblastoma cell proliferation and promotes cell differentiation by targeting TAZ. *Biochem Biophys Res Commun.* 2015; 457:171-176.
43. Witkos TM, Koscianska E and Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. *Curr Mol Med.* 2011; 11:93-109.
44. Niyazi M, Zehentmayr F, Niemoller OM, Eigenbrod S, Kretzschmar H, Schulze-Osthoff K, Tonn JC, Atkinson M, Mortl S and Belka C. MiRNA expression patterns predict survival in glioblastoma. *Radiat Oncol.* 2011; 6:153.
45. Cho H, Yu A, Kim S, Kang J and Hong S-M. Robust Likelihood-Based Survival Modeling with Microarray Data. *Genomics Proteomics Bioinformatics.* 2009. 2009; 29:16.
46. Grasbon-Frodl EM, Kreth FW, Ruiter M, Schnell O, Bise K, Felsberg J, Reifenberger G, Tonn JC and Kretzschmar HA. Intratumoral homogeneity of MGMT promoter hypermethylation as demonstrated in serial stereotactic specimens from anaplastic astrocytomas and glioblastomas. *Int J Cancer.* 2007; 121:2458-2464.
47. Thon N, Eigenbrod S, Grasbon-Frodl EM, Ruiter M, Mehrkens JH, Kreth S, Tonn JC, Kretzschmar HA and Kreth FW. Novel molecular stereotactic biopsy procedures reveal intratumoral homogeneity of loss of heterozygosity of 1p/19q and TP53 mutations in World Health Organization grade II gliomas. *J Neuropathol Exp Neurol.* 2009; 68:1219-1228.
48. Bady P, Sciuscio D, Diserens AC, Bloch J, van den Bent MJ, Marosi C, Dietrich PY, Weller M, Mariani L, Heppner FL, McDonald DR, Lacombe D, Stupp R, Delorenzi M and Hegi ME. MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. *Acta Neuropathol.* 2012; 124:547-560.

A genomic copy number signature predicts radiation exposure in post-Chernobyl breast cancer

Christina M. Wilke^{1†}, Herbert Braselmann^{1,2†}, Julia Hess^{1,2}, Sergiy V. Klymenko⁴, Vadim V. Chumak⁴, Liubov M. Zakhartseva⁵, Elena V. Bakhanova⁴, Axel K. Walch⁶, Martin Selmansberger¹, Daniel Samaga¹, Peter Weber¹, Ludmila Schneider^{1,2}, Falko Fend⁷, Hans C. Bösmüller⁷, Horst Zitzelsberger^{1,2,3} and Kristian Unger^{1,2}

¹Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany

²Clinical Cooperation Group 'Personalized Radiotherapy of Head and Neck Cancer', Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg 85764, Germany

³Department of Radiation Oncology, University Hospital, LMU Munich, München, Germany

⁴National Research Center for Radiation Medicine of National Academy of Medical Sciences of Ukraine, Kyiv, Ukraine

⁵Bogomolets National Medical University, Kyiv, Ukraine

⁶Research Unit Analytical Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany

⁷Institute of Pathology and Neuropathology, Tübingen, Germany

Breast cancer is the second leading cause of cancer death among women worldwide and besides life style, age and genetic risk factors, exposure to ionizing radiation is known to increase the risk for breast cancer. Further, DNA copy number alterations (CNAs), which can result from radiation-induced double-strand breaks, are frequently occurring in breast cancer cells. We set out to identify a signature of CNAs discriminating breast cancers from radiation-exposed and non-exposed female patients. We analyzed resected breast cancer tissues from 68 exposed female Chernobyl clean-up workers and evacuees and 68 matched non-exposed control patients for CNAs by array comparative genomic hybridization analysis (aCGH). Using a stepwise forward-backward selection approach a non-complex CNA signature, that is, less than ten features, was identified in the training data set, which could be subsequently validated in the validation data set (p value < 0.05). The signature consisted of nine copy number regions located on chromosomal bands 7q11.22-11.23, 7q21.3, 16q24.3, 17q21.31, 20p11.23-11.21, 1p21.1, 2q35, 2q35, 6p22.2. The signature was independent of any clinical characteristics of the patients. In all, we identified a CNA signature that has the potential to allow identification of radiation-associated breast cancer at the individual level.

Ionizing radiation is a known risk factor for the development of breast cancer.¹ An association with increased breast cancer risk has been reported after exposure to ionizing radiation in the course of medical treatment, after nuclear reactor accidents or by the Japan atomic bombings.^{2,3} In particular, for female breast cancer in Chernobyl clean-up workers, who participated in recovery operation works in 1986–1987 after the Chernobyl reactor accident, an almost doubled standardized incidence ratio has been reported when compared to the

national sporadic breast cancer incidence.^{4,5} Furthermore an increased breast cancer rate could also be detected among the population of the most contaminated regions of Ukraine and Belarus.⁶ So far only associations with genomic instability, Her2 and c-myc amplification and higher histological grade have been described for breast cancers that developed in atomic bomb survivors in Japan.^{7,8} Results of breast cancers that developed in women previously irradiated for Hodgkin Lymphoma are conflicting with some studies suggesting a

Key words: copy number signature, Chernobyl, breast cancer, ionizing radiation

Abbreviations: AIC: Akaike Information Criterion; Array CGH: array comparative genomic hybridization analysis; AUC: area under the curve; CNAs: genomic copy number alterations; FFPE: formalin-fixed paraffin-embedded; HNSCC: head and neck squamous cell carcinoma; IHC: immunohistochemistry; NHEJ1: non-homologous end-joining factor 1; NPV: negative predictive value; NST: invasive ductal carcinomas of no special type; PPV: positive predictive value; PTC: papillary thyroid cancer; qPCR: quantitative real-time polymerase chain reaction; SVM: support vector machine; TNM: primary tumor, lymph node metastases, distant metastases

Additional Supporting Information may be found in the online version of this article.

[†]C.M.W. and H.B. contributed equally to this project and should be considered co-first authors

Grant sponsor: Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit (BMUB); **Grant numbers:** 3615S32454, 3611S30019

DOI: 10.1002/ijc.31533

History: Received 18 Oct 2017; Accepted 23 Mar 2018; Online 16 Apr 2018

Correspondence to: Kristian Unger, Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany, Tel.: +49-893-1870-3516, E-mail: unger@helmholtz-muenchen.de

What's new?

Exposure to ionizing radiation during medical procedures or following nuclear accidents can increase breast cancer risk by inducing DNA double-strand breaks that potentially lead to DNA copy number alterations. In this study, the authors identified a genomic copy number signature associated with radiation exposure in breast cancers in women who were exposed to ionizing radiation as Chernobyl clean-up workers or accident evacuees. The signature, composed of nine genomic copy number regions, enabled the calculation of a breast cancer radiation-exposure risk score, which was independent of clinical characteristics. The findings cast light on a new approach to radiation-induced breast cancer detection.

higher rate of the basal-like subtype in irradiated women and others showing a higher rate of Her2 amplification.^{9,10} However, no histological or molecular marker has been reported so far that allows identification of radiation-associated breast cancers after low-dose exposure. In this study, we aimed to identify genomic copy number alterations that specifically allow detection of radiation-associated breast cancers. CNAs account for 85% of the variation in gene expression and define key genetic events driving tumorigenesis.^{11,12} Knowledge of radiation-exposure specific CNAs should therefore also provide mechanistic insights into radiation-associated breast carcinogenesis. Breast cancer is a heterogeneous disease with distinct biological features and clinical behaviour.¹³ Copy number and gene expression profiling of sporadic breast cancer has led to the identification of different molecular subtypes (luminal, Her2, basal-like breast cancer).¹⁴ Hence, CNAs represent an important molecular layer in breast cancer that also bears the potential providing prognostic markers.¹⁵ The thyroid is another radiation-sensitive organ and it has been shown that in papillary thyroid carcinomas that developed in patients who were exposed to ionizing radiation at young age, chromosomal band 7q11.22-11.23 was specifically amplified.¹⁶ In this study, a combined forward-backward selection approach was applied on CNA data in order to identify a CNA-signature with low complexity that allows the identification of radiation-associated breast cancers. The approach was applied to a whole genome array CGH data set on breast cancers from a cohort of female clean-up workers who were exposed to ionizing radiation from the Chernobyl reactor accident and non-exposed controls matched for residence, tumor type, age at diagnosis, TNM classification and histological grading.

Material and Methods**Clinical samples and data**

We analyzed formalin-fixed paraffin-embedded (FFPE) breast cancer tissue samples from 68 female Ukrainian patients that were exposed to ionizing radiation after the Chernobyl reactor accident in 1986. For comparison, a matched set of 68 breast cancer samples from non-exposed patients from Ukraine was investigated. The exposed and non-exposed patients included in this study were matched for residence, tumor type, age at diagnosis, TNM classification and histological grading. All tumors were diagnosed as invasive

carcinomas of no special type (NST) and were from female patients younger than 60 years at the time of diagnosis. The 136 breast cancer cases were randomly split into a training set ($n = 68$) and validation set ($n = 68$), while for each of the sets half of the cases were exposed and the other half were non-exposed controls. A genomic copy number signature was developed from the training set data with subsequent performance assessment in the validation set.

Out of the 34 patients from the training set, 27 were registered as clean-up workers, five patients as evacuees and two patients were registered as both evacuee and clean-up worker. Seven out of 68 patients of the training set received neoadjuvant radiotherapy (1–3 days before surgery). The majority (29 out of 34) of patients from the validation set were registered as clean-up workers. Three patients were registered as evacuees and two were registered as both evacuee and clean-up worker. Seven out of 68 patients of the validation set received neoadjuvant radiotherapy (1–3 days before surgery). The absorbed doses of the exposed breast cancer patients were reconstructed by the RADRUE method, which was adapted specifically for estimation of breast doses.¹⁷ The doses showed a large inter-individual variability ranging from 0.06 to 582.96 mGy (median 13.07 mGy) in the clean-up workers and from 5.72 to 36.68 mGy (median 18.40 mGy) in the evacuees.¹⁸

HER2 genomic copy number status was detected by fluorescence *in situ* hybridization as published by Wilke *et al.* Progesterone and estrogen receptors, C-kit, cytokeratin 5/6, p53 and Ki67 antigen expression detection was performed by immunohistochemical staining according to the previously described protocol.¹⁹

An overview of the clinicopathologic characteristics of the training and validation sets as well as information about age at time of exposure, age at time of diagnosis and latency is shown in Table 1. The patient's individual data are listed in Supporting Information, Tables S1 and S2. For testing associations of exposure status with clinical characteristics of the patients such as estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status (positive/negative), C-kit-expression status (positive/negative), Ki67-expression status (positive/negative), Her2/neu-status, p53-mutation status, BRCA1/2-mutation status, pT-status, pN-status and histological grading, Fisher's exact test was used. For testing associations of exposure status with the age at time of diagnosis, *t* test was used. Significance was accepted for *p* values < 0.05.

Table 1. Patient characteristics of the Chernobyl training and validation set

Characteristics	Training set			Validation set			
	Exposed	Not exposed	<i>p</i> value ¹	Exposed	Not exposed	<i>p</i> value ¹	
Number of patients	34	34		34	34		
Tumor type, no. (%)	Invasive carcinoma of no special type	34 (100)	34 (100)	1 ¹	34 (100)	34 (100)	1 ¹
Age at diagnosis, median (years), (range (years))	51.50 (37.58–59.67)	49.83 (34.67–59.25)	0.47 ²	48.04 (35.33–59.17)	50.96 (35.58–58.50)	0.55 ²	
Age at exposure, median (years), (range (years))	33.92 (24.17–45.50)	NA		30.58 (18.50–42.58)	NA		
Latency, median (years), (range (years))	18.83 (10.00–23.83)	NA		19.92 (9.00–29.58)	NA		
Estrogen-receptor status, no. (%)	Positive	21 (62)	20 (59)	1 ¹	26 (76)	28 (82)	0.77 ¹
	Negative	13 (38)	14 (41)		8 (24)	6 (18)	
Progesterone-receptor status, no. (%)	Positive	18 (53)	21 (62)	0.62 ¹	25 (74)	25 (74)	1 ¹
	Negative	16 (47)	13 (38)		9 (26)	9 (26)	
C-kit status, no. (%)	Positive	4 (12)	2 (6)	0.67 ¹	4 (12)	5 (12)	1 ¹
	Negative	30 (88)	32 (94)		30 (88)	29 (88)	
Cytokeratin 5/6 status, no. (%)	Positive	6 (18)	3 (9)	0.48 ¹	6 (18)	4 (12)	0.73 ¹
	Negative	28 (82)	31 (91)		28 (82)	30 (88)	
P53 status, no. (%)	Positive	18 (53)	14 (41)	0.47 ¹	13 (38)	20 (59)	0.15 ¹
	Negative	16 (47)	20 (59)		21 (62)	14 (41)	
Ki-67 status, no. (%)	Positive	31 (91)	34 (100)	0.24 ¹	30 (88)	30 (88)	1 ¹
	Negative	3 (9)	0 (0)		4 (12)	4 (12)	
BRCA1/2 status, no. (%)	Positive	4 (12)	4 (12)	1 ¹	0 (0)	1 (3)	1 ¹
	Negative	30 (88)	29 (85)		34 (100)	33 (97)	
	Not evaluable	0 (0)	1 (3)		0 (0)	0 (0)	
Her2 status, no. (%)	Positive	4 (12)	7 (21)	0.52 ¹	4 (12)	2 (6)	0.43 ¹
	Negative	27 (79)	27 (79)		29 (85)	32 (94)	
	Not evaluable	3 (9)	0 (0)		1 (3)	0 (0)	
pT stage, no. (%)	pT1	13 (38)	15 (44)	0.9 ¹	13 (38)	12 (35)	1 ¹
	pT2	20 (59)	18 (53)		19 (56)	20 (59)	
	pT3	1 (3)	1 (3)		2 (6)	2 (6)	
pN stage, no. (%)	pN0	18 (53)	19 (56)	1 ¹	17 (50)	17 (50)	1 ¹
	PN1	14 (41)	15 (44)		17 (50)	17 (50)	
	pN2	1 (3)	0 (0)		0 (0)	0 (0)	
	pNx	1 (3)	0 (0)		0 (0)	0 (0)	
pM stage, no. (%)	M0	34 (100)	34 (100)	1 ¹	34 (100)	34 (100)	1 ¹
Grade, no. (%)	G1	1 (3)	1 (3)	1 ¹	3 (9)	3 (9)	1 ¹
	G2	20 (59)	20 (59)		24 (71)	24 (71)	
	G3	13 (38)	13 (38)		7 (21)	7 (21)	

¹The *p* value was calculated by Fisher's-exact test.²The *p* value was calculated by *t* test.

Genomic copy number analysis by array CGH

To characterize genomic copy number alterations in the post-Chernobyl breast cancer cohorts, array CGH was performed using high-resolution oligonucleotide-based SurePrint G3 Human 60k CGH microarrays (AMADID 21924, Agilent Technologies, USA). The workflow is described in the Supporting Information, material and methods part.

Hierarchical cluster analysis of DNA copy number profiles was performed using correlation distance and method “Ward.” For testing associations of clusters with exposure status, estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status, C-kit-expression status, Ki67-expression status, Her2/neu-status, p53-mutation status, BRCA1/2-mutation status, triple negative status, tumor size, lymph-node status, histological grading, age at exposure, Fisher’s exact test was used. ANOVA F-test was used for calculating associations of clusters with age at diagnosis, age at exposure and latency. Significance was accepted for p values < 0.05 .

Generation of CNA signature

To identify a genomic copy number signature that allows the prediction of radiation exposure we followed a multivariate logistic regression approach. Logistic regression models the probabilities P of class membership for each patient (exposed or non-exposed) directly according to the formula $P = P(h) = \exp(h)/(1 + \exp(h))$, where $h = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n = \log(P/(1 - P))$ is the logit or logarithmic odds value, with predictor variables X_i , coefficients β_i and n the number of variables in the model. The calculated probability P serves then as risk score for radiation exposure. Tumors with a prediction probability $P > 0.5$ were classified as radiation associated. For more details, see James *et al.*²⁰

Binary copy number alteration states of all altered copy number regions have been used as variables whilst gains and losses were treated separately. Thus, for every region gain/no gain (0/1) and loss/no loss (0/1) were reported. Hence, for each copy number region gain status and loss status were treated as independent variables. For the purpose of model fit and validation, the described training and validation sets were used. Feature selection was performed by stepwise combined forward–backward selection, using the functions *glm* (for generalized linear modelling) and *step* for Akaike Information Criterion (AIC) based selection of the best models from the R package *stats*.²¹ The algorithm of function *step* computes the likelihoods of each model fit for a sequential selection of features, whilst the best performing model was determined using AIC for the sake of the best trade-off between bias and variance of the model.²⁰ The negative likelihood, which is a positive value, decreases with increasing number of features in the model. AIC simply adds twice the number of features to the negative likelihood, so that it reaches a minimum, which determines the optimal number of features. Only CNAs (gains or losses) that occurred at

least 5 times in the training set and with univariate p values up to 0.25 between exposed and non-exposed tumors (Fisher’s exact test) were admitted for the selection algorithm. The number 5 roughly reflects a standard deviation $\sqrt{5}$ (Poisson rule) corresponding to a CV $< 50\%$, which makes calculations more stable. 0.25 is also used as a default entry value for example in variable selection the SAS procedure PROC PHREG. Subsequently, the afore-defined risk score, based on the coefficients defined using the training set, was calculated for every tumor in the validation set. Finally, a confusion table was built for the comparison of the true and predicted exposure states and a p value using one-tailed Fisher’s exact test was determined.

Fisher’s exact test was also used to test the binary associations of the risk score with any clinical characteristics of the patients such as estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status (positive/negative), C-kit-expression status (positive/negative), Ki67-expression status (positive/negative), Her2/neu-status, p53-mutation status, BRCA1/2-mutation status and intrinsic subtypes. Significance was accepted for p values < 0.05 .

Quantitative PCR (qPCR)

For technical validation of the CNAs detected by aCGH, the copy number status of genes representative for the copy number regions included in the CNA-signature, was determined by genomic copy qPCR. The workflow of the genomic copy number qPCR is described in the Supporting Information, material and methods part.

The calculated copy number state was used as the basis for further calculations in R. Values smaller than 1.5 were considered as losses and values > 2.5 were considered as gains. The thresholds were taken from the CopyCaller software. As reference assay Life Technologies recommend to use a gene that is known to exist in two copies in a diploid genome and is being unaffected in all of the experimental samples. It was not possible to extract a gene showing no CNA in the whole data set. From the most commonly used reference genes, the *RNaseP* gene showed the lowest number of CNAs over all experimental samples. Therefore, we decided to use copy number reference assay for this gene as reference. To make results comparable between qPCR and aCGH, we also corrected the aCGH copy number states with that of the appropriate locus covering the *RNaseP* gene. The copy number state as determined by array CGH and qPCR were summarized in a confusion table and subjected to Fisher’s exact test. p values < 0.05 indicated confirmation of the array CGH results by qPCR.

Dose–response analysis

Logistic-regression analysis was performed in order to test for relation between radiation dose and the occurrence of signature CNAs. The workflow is described in the Supporting Information, material and methods part.

Results

This study aimed at the identification of radiation-associated DNA copy number changes in a cohort of breast cancers from post-Chernobyl clean-up workers and evacuees from highly contaminated territories. For this purpose, copy number profiles of exposed and non-exposed control cases were generated and a radiation-exposure CNA-signature was established.

Hierarchical clusters reveal association with radiation exposure

High-resolution aCGH profiles of 136 breast cancer samples were generated in order to characterize genomic copy number patterns of radiation-associated breast cancer. Supporting Information, Figure S1 shows all genomic copy number profiles after unsupervised hierarchical clustering with annotated parameters exposure status, estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status, C-kit-expression status, Ki67-expression status, Her2/neu-status, p53-mutation status, BRCA1/2-mutation status, triple negative status, tumor size, lymph-node status and histological grading. The two main clusters C1 and C2 of the hierarchical cluster analysis consisted of 33 and 103 cases, respectively, the subclusters of C2 consisted of 36 cases (C2.1) and 67 cases (C2.2), respectively, and the sub-sub clusters of C2.2 consisted of 22 cases (C2.2.1) and 45 cases (C2.2.2), respectively. In general DNA losses and gains occurred more frequently in cluster C1 compared to clusters C2.1, C2.2.1 and C2.2.2. Furthermore, C2.2.1 in general showed a lower number of aberrations compared to clusters C1, C2.1 and C2.2.2. From all tested parameters exposure status ($p = 0.019$), histological grading ($p = 0.03$), estrogen-receptor status ($p = 0.04$), cytokeratin-expression status ($p = 0.04$), Her2/neu-status ($p = 0.01$), BRCA1/2-mutation status ($p = 0.04$), age at diagnosis (F-test, degrees of numerator $dn = 3$, degrees of denominator $dd = 132$, $p = 0.03$) and tumor size ($p = 0.02$) were differentially distributed across C1, C2.1, C2.2.1 and C2.2.2 (Supporting Information, Table S3). With regard to exposure status all clusters showed equal distributions except cluster C2.1, which contained significantly more non-exposed than exposed cases (26 out of 36, 72%). Further, no association of exposure status with age at diagnosis or other clinical characteristics of the patients was detected (Table 1). Large tumors (pT2 and pT3) were associated with clusters C1, C2.1 and C2.2.2 (76 out of 83, 92%). Within clusters C2.2.1 and C2.2.2 significantly less G3 tumors (12 out of 40, 30%) were included. In addition aCGH profiles from estrogen-receptor negative cases were underrepresented in clusters C2.2.1 and C2.2.2 (13 out of 41, 32%). Cases with Her2/neu-status positive and Cytokeratin 5/6-expression positive were associated with clusters C1, C2.1 and C2.2.2 (Cytokeratin 5/6-expression positive: 19 out of 19, 100%, Her2/neu-status positive 17 out of 17, 100%). Cases

with a BRCA1/2-mutation were enriched in cluster C2.1 (6 out of 9, 67%).

Moreover, patients of cluster C2.1 were significantly younger at age of diagnosis (mean: 47.08 years) compared to cases of cluster C1 (mean: 50.79 years), cluster C2.2.1 (mean: 51.66 years) and cluster C2.2.2 (mean: 50.04 years).

Identification of a nine-genomic CNA-signature predicting radiation exposure

In the first step, univariate testing was used as a preselection step for selection of highly discriminating copy number changes. Admitted for the selection algorithm were only gains or losses that occurred at least five times in the training set and that showed univariate p values < 0.25 (see Material and Methods and Supporting Information, Table S4). This resulted in 144 out of 910 CNA regions. In a next step, the most discriminating features (i.e., CNA regions) were selected by stepwise combined forward and backward selection and the optimal number of features was determined by Akaike Information Criterion (AIC, see Material and Methods) to avoid overfitting. This approach revealed a CNA-signature composed of nine altered genomic copy number regions located on chromosomal bands 7q11.22–11.23 (7:70899666–72726548), 7q21.3 (7:97597612–97749420), 16q24.3 (16:89472538–90111178), 17q21.31 (17:44210733–44231916), 20p11.23–11.21 (20:20226791–24223097), 1p21.1 (1:105300245–105546898), 2q35 (2:220499593–220503940), 2q35 (2:219083470–220474362), 6p22.2 (6:26033303–26234636) in the Chernobyl training set. The parameter values of the features are shown in Table 2. Further, as explained in Material and Methods, the model, defined by the calculated parameters, was evaluated in the validation set. For every tumor, the probability P was calculated as a risk score according to the model formula. The score values P appeared to be strongly clustered. 22 values were $< 1.0 \times 10^{-7}$, 11 times 0.833, 33 times $> (1 - 10^{-7})$ and two values 0.355 and 0.667. After rounding to a few decimal digits, 5 uniquely different values remained. Tumors were then predicted as exposed if $P > 0.5$ or as non-exposed if $P < 0.5$. The results of the prediction performance assessment of the CNA-signature on the validation set are shown in Figure 1. Of the 68 cases, 45 were predicted to be exposed and 23 non-exposed (predicted positive and predicted negative, right and left side in the three panels of Figure 1, respectively). From the lower panel in Figure 1 performance parameters can be read. The 45 positive predicted split into 27 true and 18 false positives, the 23 negative predicted into 16 true and 7 false negatives. We found a significant binary association of the risk score with radiation exposure status, which means that among the positive predicted cases we found an enrichment of exposed cases (PPV = $27/43 = 0.60$, lower panel, right side) compared to exposed cases on the left side ($1 - NPV = 7/23 = 0.304$, lower panel, right side, one-tailed Fisher's-exact test, p value = 0.02). Under the given conditions (34 exposed, 34 non-exposed), this is equivalent to say that the true positive rate

Table 2. Stepwise forward selection of CNA regions in the training set and technical validation by qPCR of the nine-genomic CNA signature

Forward selection step	Region identifier ¹	Number of clones ²	Start of region ^{2,3}	End of region ^{2,3}	Chromosomal location	Residual degree of freedom	Residual deviance (Log-Likelihood ratio)	Akaike Information Criterion	Coefficient of linear risk score ⁴	Gene/region for qPCR validation	Fisher's-exact test <i>p</i> value of qPCR validation
Intercept						67	94.3	96.3	1.61		
1	G142	3	97597612	97749420	7q21.3	66	80.4	84.4	162.14***	chr7:97654486	0.00001
2	L133	17	26033303	26234636	6p22.2	65	68.0	74.0	-141.51***	HIST1H1E	0.00001
3	G365	2	44210733	44231916	17q21.31	64	55.5	63.5	162.07***	KANSL1	0.00960
4	L55	68	219083470	220474362	2q35	63	47.7	57.7	-25.21**	NHEJ1	0.00090
5	L15	2	105300245	105546898	1p21.1	62	37.8	49.8	-22.76**	chr1:105368724	0.02540
6	G347	41	89472538	90111178	16q24.3	61	29.3	43.3	79.84**	FANCA	0.01300
7	G417	61	20226791	24223097	20p11.23-11.21	60	25.0	41.0	40.14*	FOXA2	0.04390
8	G132	26	70899666	72726548	7q11.22-11.23	59	17.9	35.9	21.85**	CALN1	0.01560
9	L56	2	220499593	220503940	2q35	58	14.6	34.6	-36.14 ^x	SLC4A3	0.00250

¹G for gain, L for loss, number according to CGH regions.

²Number of clones determined by CGH regions start = position of first, end = position of last clone region identifier according to CGH regions.

³According to annotation GRCh37.

⁴Likelihood-ratio tests. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ^x $p < 0.10$ ($df = 1$), significance of model $p < 0.0001$ ($df = 9$).

(sensitivity = $27/34 = 0.794$) is higher than false positive rate ($1 - \text{specificity} = 18/34 = 0.529$). The overall prediction error is 0.368. The foregoing analysis could be done with any other cutoff level of the probability score, yielding for each cutoff a pair of specificity and sensitivity values. These are shown in the ROC curve, Figure 2. Due to the discrete distribution of the rounded scores, the ROC contains only 4 points. One of these points, corresponding to a level of about $P = 0.70$ (between scores to avoid boundary ambiguities) shows a slightly better specificity (0.50) and prediction error (0.353), PPV = 0.614. However, this is in good agreement with the level of $P = 0.50$ which corresponds to the smallest expected prediction error bases on theoretical probabilistic considerations. The AUC (area under the curve) amounted to 0.617.

Technical validation of the nine-CNA-signature by qPCR

The copy number status of the nine signature CNAs, which was initially determined by array CGH, was technically validated by qPCR ($p < 0.05$) (Table 2 and Supporting Information, Figure S2). For this purpose, aliquots of the same genomic DNA samples that were used in array CGH analysis were analyzed by qPCR. All nine representative genes/regions from the copy number regions of the CNA-signature showed similar copy number changes compared to array CGH, confirming the initial finding ($p < 0.05$).

Association of the nine-CNA-signature with clinical and histological data

The risk score derived from the CNA-signature (7q11.22-11.23, 7q21.3, 16q24.3, 17q21.31, 20p11.23-11.21, 1p21.1, 2q35, 2q35, 6p22.2) was not associated with any clinical characteristics of the patients such as estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status (positive/negative), C-kit-expression status (positive/negative), Ki67-expression status (positive/negative), Her2/neu-status, p53-mutation status, BRCA1/2-mutation status and intrinsic subtypes in the Chernobyl training or the Chernobyl validation set. This suggests an independent association of the discovered nine-CNA-signature with radiation exposure of patients.

Dose-response analysis

No statistically significant association of the occurrence of each of the nine signature CNAs with reconstructed radiation dose was detected. Moreover, no significant influence of radiation-dose on the occurrence of each of the nine signature CNAs could be found in logistic-regression analysis.

Discussion

In this study, we identified a genomic copy number signature that predicts radiation exposure in post-Chernobyl breast cancer. Previous studies reported that even at low doses, ionising radiation alters gene expression as a result of induced CNAs and thus is capable of driving the process of carcinogenesis.²² In young patients who were exposed to radiation at

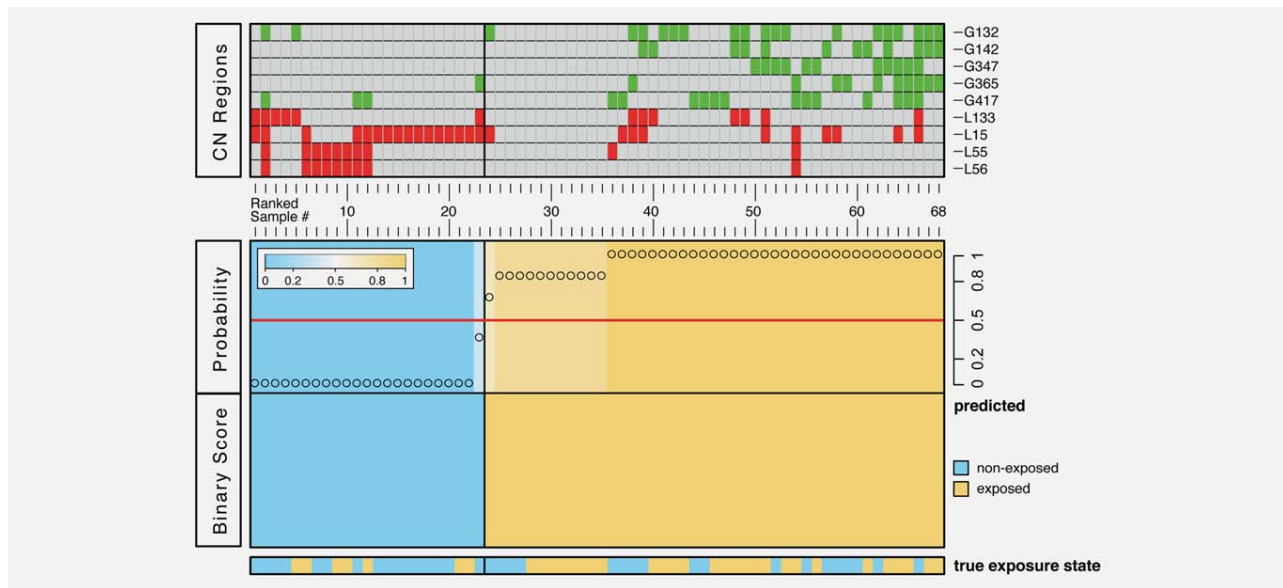


Figure 1. Heatmap of the 9-CNA-signature of 68 breast cancer patients of the validation set composed of 34 exposed and 34 non-exposed cases. Copy number gains are represented by green color, losses by red color (top panel). The middle panel shows the risk score on the probability scale calculated according to the formula described in Material and Methods. Samples (columns) are sorted in ascending order of the risk score. Cases with probabilities ≥ 0.5 are predicted as exposed, otherwise as non-exposed (middle panel, right and left side, respectively). Given exposure status is shown in the lower panel, thus on the right orange cases mark true positives, blue cases mark false positives. On the left side orange cases mark false negatives, blue cases mark true negatives.

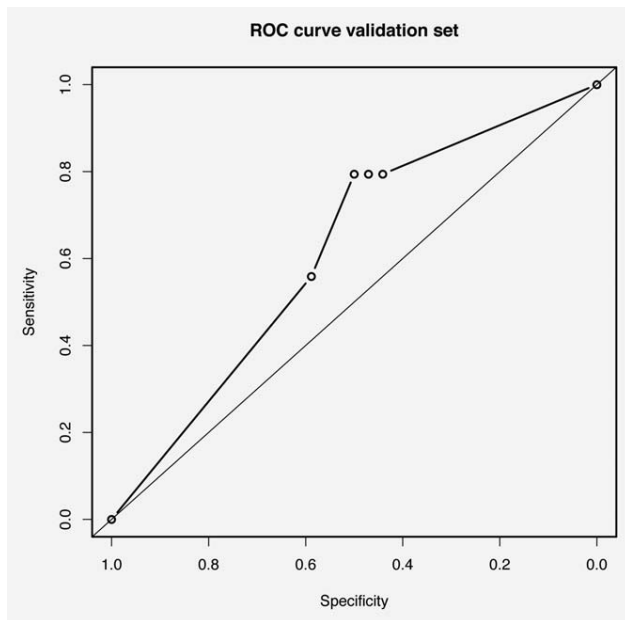


Figure 2. ROC curve calculated by applying a logistic regression model fitted on the training set and evaluated on the validation set. Each point (circles) corresponds to a probability cutoff level decreasing from left to right, given by the steps visualized in Figure 1. Points are connected by straight lines.

very young age, copy number gain of the chromosomal band 7q11.22–11.23 has been identified as a marker of radiation exposure in papillary thyroid carcinomas.¹⁶ As for thyroid

cancer, ionizing radiation is also known to be a risk factor for the development of breast cancer; however, radiation-specific markers in these tumors are yet undiscovered.^{1,4–6} Initial studies on gene alterations in breast cancers from the Atomic bomb survivors in Japan revealed a higher frequency of Her2 and c-myc oncogene amplifications as well as a higher histological grading in these radiation-associated tumors.^{7,8} However, we did not detect an association of Her2 and c-myc amplification and high histological grade with breast cancer of patients from the exposed group in our study (Table 1). This could be due to the fact that patients in our study were exposed to different radiation conditions compared to those the Atomic bomb survivors were exposed to. Clean-up workers of the Chernobyl accident were exposed to more heterogeneous conditions in contrast to the rather homogeneous conditions the Atomic bomb survivors were exposed to. In addition, women in our study were younger at time of diagnosis (under 60 years old). Furthermore, exposed and non-exposed samples were matched for histological grading in the present study. For the identification of radiation-specific copy number changes, we used an exploratory approach on whole genome profiling of genomic copy number alterations of resected breast cancer tissues from exposed and matched non-exposed patients.

So far, CNAs are very well described in sporadic breast cancer while frequently observed CNAs include gain of chromosomal bands 1q, 3q, 4p, 8q, 11q, 17q and 20q and losses of chromosomal bands 1p, 8p, 11p, 13q, 16q, 17p, 19p and 22q.^{15,23–25}

Table 3. Cancer-related candidate genes and miRNAs located in the chromosomal regions of the nine-CNA-signature predicting radiation exposure in breast cancer

Chromosomal location	Start of region ^{1,2}	End of region ^{1,2}	Cancer-related candidate genes and miRNAs	Type of aberration
7q21.3	97597612	97749420	OCM2, LMTK2	Gain
6p22.2	26033303	26234636	HIST1H1C, HIST1H1T, HIST1H1E, HIST1H1D, HIST1H2AB, HIST1H2AC, HIST1H2AD, HIST1H2BB, HIST1H2BC, HIST1H2BD, HIST1H2BE, HIST1H2BF, HIST1H2BG, HIST1H4C, HIST1H4D, HIST1H4E HFE	Loss
17q21.31	44210733	44231916	KANSL1	Gain
2q35	219083470	220474362	ARPC2, TMBIM1, GPBAR1, AAMP, PNKD, SLC11A1, USP37, TTL4 RQCD1, CYP27A1, WNT6, WNT10A, IHH, NHEJ1, ATG9A, PTPRN, STK36, hsa-miR-26b-5p, hsa-miR-375	Loss
1p21.1	105300245	105546898	No tumor-related candidate gene	Loss
16q24.3	89472538	90111178	ANKRD11, SPG7, RPL13, CPNE7, DPEP1, CHMP1A, CDK10, FANCA, MC1R, TUBB3, C16orf3	Gain
20p11.23-11.21	20226791	24223097	INSM1, RALGAPA2, PAX1, XRN2, NKX2-2, FOXA2, SSTR4, CD93	Gain
7q11.22-11.23	70899666	72726548	CALN1, STAG3L3, SBDSP1	Gain
2q35	220499593	220503940	SLC4A3	Loss

¹Number of clones determined by CGH regions start = position of first, end = position of last clone region identifier according to CGH regions.

²According to annotation GRCh37.

All these CNAs are in good agreement with CNA-profiles of this study, which substantiates the plausibility of our results. Similar findings have been observed in breast cancers associated with exposure to ionizing radiation in the course of medical treatment.¹⁰ Other cytogenetic studies on breast cancer have identified CNAs that are associated with clinical parameters and overall survival.^{15,24–26} Of special interest is an association of histological grading and estrogen-receptor status with specific DNA copy number patterns derived from primary breast cancers.²⁴ These estrogen-receptor and histological grading specific patterns, such as gain of 1q and loss of 16q which are associated with lower histological grading and estrogen-positive tumors, could also be confirmed in our study after unsupervised clustering of array-CGH profiles (Supporting Information, Figure S1). Overall, unsupervised hierarchical clustering separated the breast cancer CNA profiles into four main clusters that correlate with histological grading, estrogen-receptor status, Her2/neu-status, BRCA1/2-mutation status, cytokeratin-expression status, age at diagnosis and tumor size (Supporting Information, Figure S1 and Table S3). In addition, the profiles of exposed and non-exposed cases were differentially distributed between observed clusters suggesting a radiation-exposure-specific signal within the genomic copy number profiling data. However,

delineation of copy number alterations determining the clustering is not trivial and might not result in radiation-exposure specific copy number alterations since an influence of the other cluster-associated parameters is likely. However, these findings from the unsupervised cluster analysis motivated us to develop a low-complex CNA-signature predicting radiation exposure. From mRNA and miRNA expression data, signatures have been already generated predicting clinical outcome or estrogen-, progesteron-receptor-status and Her2-status in sporadic breast cancer but there is no such prediction rule at the genomic copy number level.^{27,28} Compared to results from association testing, prediction models come with the advantage that they provide both biological mechanistic insights and, moreover, bare the potential of being used as diagnostic or prognostic tools. In the context of radiation-associated breast cancer a prediction rule could allow identification of breast cancer tissues that developed after exposure of patients to ionizing radiation. In order to generate such a prediction rule we deployed stepwise combined forward-backward selection in combination with multivariate logistic regression. Signature modeling approaches using copy number alterations were applied earlier by Pronold *et al.* and by Sung *et al.* who applied other statistical approaches.^{29,30} Pronold *et al.* used nearest shrunken

centroids applied to sums of log₂-ratios within common copy number variation segments to predict human ancestry of healthy individuals.²⁹ Sung *et al.* applied a 1-norm support vector machine (SVM) to binary copy number alteration data for a binary classification of histological subtypes of endometrial cancer.³⁰ In our study, logistic regression for a binary classification of radiation exposure status was chosen for two reasons: First, called copy number data should preferentially represent raw or segmented log₂-ratios because of the reduction of noise, interpretability and downstream analysis according to Van Wieringen *et al.*³¹ Second, logistic regression allows to provide a risk score on the individual level which is directly associated to the class probabilities.³² Our approach resulted in a CNA-signature predicting radiation exposure in breast cancer that is composed of nine genomic copy number regions located on chromosomal bands 7q11.22–11.23, 7q21.3, 16q24.3, 17q21.31, 20p11.23–11.21, 1p21.1, 2q35, 2q35 and 6p22.2 (Figure 1 and Table 2). The signature allowed calculating a breast cancer radiation exposure risk score on the probability scale (Figure 1), which was statistically not associated with any clinical characteristics. This suggests the signature being an independent prognosticator of radiation exposure of patients. At this point one limitation factor is, that we do not have data on lifestyle factors such as obesity (in postmenopausal women) and alcohol consumption, which are known to increase the risk for developing a breast cancer.³³ Therefore, we cannot address any potential influence of these in our analysis. Moreover, although having information on the smoking status of patients, we considered working out potential influence of smoking as not meaningful since most of the patients were non-smokers.³⁴

Furthermore, no dose–response or statistical association of the occurrence of CNAs of the signature regions could be detected. This might be due to another limitation, which is that dose estimates by RADRUE were only available for a subset of patients. In addition, an important fact is the uncertainty of dose estimation. The intrinsic uncertainty is mostly influenced by the uncertainty of dose rates. Another important component is the ‘human factor uncertainty,’ which includes intentional or unintentional mistakes of recollection and description of the clean-up activities.³⁵ In case of the female clean-up workers included in this study, this factor is less pronounced due to the relative simplicity of individual histories and their operation away from highly heterogeneous dose rate fields. Furthermore, a small proportion of patients received very small irradiation doses (0.06 mGy) according to the RADRUE dose estimation. Although it is possible that such low doses have no biological effects the samples were not excluded since we aimed at the identification of a robust CNA signature for which we preferred a heterogeneous data set over a homogeneous one. A further limitation point of this study is, that some of the patients received neoadjuvant radiotherapy one to three days prior surgery. However, it is unlikely that over this short period clonal expansion of cells

harboring the same CNAs occurs. Therefore, we would not expect detectable CNAs that developed in the course of the neoadjuvant radiotherapy treatment.

However, like many statistical methods, the application of the signature as a classifier has its own limitations. The best performance values calculated on the validation set were a sensitivity of about 80% (0.794) and an NPV (negative predictive value) of 70% (0.70, given a prevalence of 0.50, that is, 34 exposed and 34 non-exposed). The PPV (positive predictive value) was 61.7% (0.617). Often, in diagnostic practice, one tries to improve the PPV by increasing the cutoff level of the risk score at the cost of sensitivity. This assumes a continuous relationship between the score and the PPV. Using the highest discriminating probability cutoff level in the data ($P \sim 0.9$) yields a PPV of $19/33 = 0.576$ (Figure 1). Modeled probabilities higher than 0.9 were clustered close to 1.0. They correspond to linear score values h larger than 20.0 up to 300.0. From a *post hoc* logistic regression of exposure status (lower panel in Figure 1) with the linear score values h as independent variable, a smoothed estimate of the PPV could be achieved, approaching values up to 0.74; however, this continuous dependency was not significant (results not shown). Fisher’s exact test showed a significant binary association between exposure status and the risk score, using a probability of 0.5 as decision cutoff. The optimal cutoff (0.7) determined by ROC analysis (Figure 2) appeared to be slightly better (one case different); however, from Bayesian decision theoretic considerations 0.5 is the cutoff with the smallest expected prediction error. A continuous association between a risk score given by a signature of CNA and exposure status can also not be expected, because CNA are binary features. This is one reason for the discrete appearing probability scores (middle panel in Figure 1 and ROC curve Figure 2). Many of the signature patterns (heatmap, Figure 1) have frequency 1 and one cannot interpolate between different combinations of CNA. On the other hand, dosimetric uncertainties may add to the noise seen in the lower panel of Figure 1. Also and most importantly, it cannot be expected to predict a complex biological process such as tumorigenesis with only one parameter such as the signature risk score. The ability to partly explain the variance of tumorigenesis with a prediction model is scientifically important.

To get insights into the potential functional impact of the nine-CNA-signature we extracted all tumor-associated genes and miRNAs that are mapped to the signature regions (Table 3 and Supporting Information, Table S5). Interestingly, one region of the CNA-signature overlaps largely with the chromosomal band 7q11.22–11.23 which was gained in the majority of patients that have been classified as exposed. 7q11.22–11.23 has been reported to be exclusively gained in papillary thyroid carcinomas of patients who were exposed to ionizing radiation at very young age in aftermath of the Chernobyl reactor accident.¹⁶ This finding suggests that gain of the chromosomal band 7q11.22–11.23 could be a radiation marker of low doses of ionizing radiation, independent of the

tumor type. Another region of the signature, which is located on chromosomal band 16q24.3 and overexpression of the gene FANCA, which is located in this region, predicts reduced clinical outcome of radiotherapy-treated patients with head and neck squamous cell carcinoma (HNSCC).^{36,37} FANCA is a key regulator of the Fanconi anemia (FA)/breast cancer (BRCA) pathway and controls homology-directed DNA repair.³⁸ Besides FANCA, many of the genes located within the copy number regions of the signature are known to be involved in DNA-damage response and repair (Supporting Information, Table S5). A very prominent gene in this context is the non-homologous end-joining factor 1 gene (NHEJ1), which is located on chromosomal band 2q35. NHEJ1 is required for the non-homologous end-joining pathway of DNA repair.³⁹ In addition, members of the Histone H1, H2A, H2b and H4 family, all of which located in the region of the CNA-signature that covers chromosomal band 6p22.2, were also known to be involved in these processes.⁴⁰ These findings point to chromosomal instability as a major consequence of deregulated DNA repair processes, which is a well-known feature of cells exposed to ionizing radiation.⁴¹

Interestingly, copy number loss of the signature region on 2q35 contains miRNA hsa-miRNA-26b-5p, which recently was published as a breast cancer radiation marker.¹⁹ Hsa-miRNA-26b-5p expression was significantly reduced in cases showing the loss, indicating, that its expression is mainly determined by the copy number of the underlying miRNA gene (Supporting Information, Figure S3).

In summary, our study presents a novel approach to predict the radiation exposure status of breast cancer patients using a genomic copy number signature composed of nine genomic copy number regions. The identified CNA-signature may allow the detection of radiation-induced breast cancers and could serve as a diagnostic marker for radiation exposure in breast cancer. In further studies, an integration of copy number data with transcriptome data would be desirable to in-depth investigate if radiation-induced breast cancers represent a potential new molecular subtype.

Acknowledgement


The authors thank C. Innerlohinger, E. Konhäuser, L. Dajka, S. Heuer, A. Selmeier, L. Rybchenko and B. Klymuk for excellent technical support.

References

- Ronckers CM, Erdmann CA, Land CE. Radiation and breast cancer: a review of current evidence. *Breast Cancer Res* 2005;7:21–32.
- Ibrahim EM, Abouelkhair KM, Kazkaz GA, et al. Risk of second breast cancer in female Hodgkin's lymphoma survivors: a meta-analysis. *BMC Cancer* 2012;12:197.
- McGregor H, Land CE, Choi K, et al. Breast cancer incidence among atomic bomb survivors, Hiroshima and Nagasaki, 1950–69. *J Natl Cancer Inst* 1977;59:799–811.
- Prsyazhnyuk A, Gristchenko V, Fedorenko Z, et al. Twenty years after the Chernobyl accident: solid cancer incidence in various groups of the Ukrainian population. *Radiat Environ Biophys* 2007;46:43–51.
- Prsyazhnyuk AY, Bazyka DA, Romanenko AY, et al. Quarter of century since the Chernobyl accident: small es, Cyrilli cancer risks in affected groups of population. *Probl Radiac Med Radiobiol*. 2014;19:147–69.
- Pukkala E, Kesminiene A, Poliakov S, et al. Breast cancer in Belarus and Ukraine after the Chernobyl accident. *Int J Cancer* 2006;119:651–8.
- Miura S, Nakashima M, Ito M, et al. Significance of HER2 and C-MYC oncogene amplifications in breast cancer in atomic bomb survivors: associations with radiation exposure and histologic grade. *Cancer* 2008;112:2143–51.
- Oikawa M, Yoshiura K, Kondo H, et al. Significance of genomic instability in breast cancer in atomic bomb survivors: analysis of microarray-comparative genomic hybridization. *Radiat Oncol* 2011;6:168.
- Horst KC, Hancock SL, Ognibene G, et al. Histologic subtypes of breast cancer following radiotherapy for Hodgkin lymphoma. *Ann Oncol* 2014;25:848–51.
- Yang XR, Killian JK, Hammond S, et al. Characterization of genomic alterations in radiation-associated breast cancer among childhood cancer survivors, using comparative genomic hybridization (CGH) arrays. *PLoS One* 2015;10:e0116078.
- Srihari S, Kalimutho M, Lal S, et al. Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach. *Mol Biosyst* 2016;12:963–72.
- Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009;1:62.
- Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
- Natrajan R, Weigelt B, Mackay A, et al. An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, HER2 and luminal cancers. *Breast Cancer Res Treat* 2010;121:575–89.
- Bergamaschi A, Kim YH, Wang P, et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* 2006;45:1033–40.
- Hess J, Thomas G, Braselmann H, et al. Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proc Natl Acad Sci USA* 2011;108:9595–600.
- Kryuchkov V, Chumak V, Maceika E, et al. Radrue method for reconstruction of external photon doses for Chernobyl liquidators in epidemiological studies. *Health Phys* 2009;97:275–98.
- Chumak VV, Klymenko SV, Zitzelsberger H, et al. Doses of Ukrainian female clean-up workers with diagnosed breast cancer. *Radiat Environ Biophys*. 2018;57:163–68.
- Wilke CM, Hess J, Klymenko SV, et al. Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. *Int J Cancer*. 2018;142:573–83.
- James GWD, Hastie T, Tibshirani R. An introduction to statistical learning. Springer, 2013.
- Team RDC. R: A language and environment for statistical computing 2013.
- Mullenders L, Atkinson M, Paretzke H, et al. Assessing cancer risks of low-dose radiation. *Nat Rev Cancer* 2009;9:596–604.
- Li J, Wang K, Li S, et al. DNA copy number aberrations in breast cancer by array comparative genomic hybridization. *Genomics Proteomics Bioinformatics* 2009;7:13–24.
- Chin SF, Wang Y, Thorne NP, et al. Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene* 2007;26:1959–70.
- Albertson DG. Profiling breast cancer by array CGH. *Breast Cancer Res Treat* 2003;78:289–98.
- Loo LW, Grove DI, Williams EM, et al. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res* 2004;64:8541–9.
- van de Vijver MJ, He YD, van 't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;347:1999–2009.
- Lowery AJ, Miller N, Devaney A, et al. MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res* 2009;11:R27.
- Pronold M, Vali M, Pique-Regi R, et al. Copy number variation signature to predict human ancestry. *Bmc Bioinformatics* 2012;13:336.
- Sung CO, Sohn I. The expression pattern of 19 genes predicts the histology of endometrial carcinoma. *Sci Rep*. 2014;4:5174.
- van Wieringen WN, van de Wiel MA, Ylstra B. Normalized, segmented or called aCGH data? *Cancer Inform* 2007;3:321–7.

32. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics* 2004;5:427–43.
33. Dumitrescu RG, Cotarla I. Understanding breast cancer risk - where do we stand in 2005?. *J Cell Mol Med* 2005;9:208–221.
34. Macacu A, Autier P, Boniol M, et al. Active and passive smoking and risk of breast cancer: a meta-analysis. *Breast Cancer Res Treat* 2015;154: 213–224.
35. Drozdovitch V, Chumak V, Kesminiene A, et al. Doses for post-Chernobyl epidemiological studies: are they reliable? *J Radiol Prot* 2016;36: R36–73.
36. Bauer VL, Braselmann H, Henke M, et al. Chromosomal changes characterize head and neck cancer with poor prognosis. *J Mol Med.* 2008;86: 1353–65.
37. Hess J, Unger K, Orth M, et al. Genomic amplification of Fanconi anemia complementation group A (FancA) in head and neck squamous cell carcinoma (HNSCC): cellular mechanisms of radiosensitivity and clinical relevance. *Cancer Lett* 2017; 386:87–99.
38. D'Andrea AD, Grompe M. The Fanconi anaemia/BRCA pathway. *Nat Rev Cancer* 2003;3:23–34.
39. Hefferin ML, Tomkinson AE. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair (Amst)*. 2005;4:639–48.
40. Scaffidi P. Histone H1 alterations in cancer. *Biochim Biophys Acta* 2016;1859:533–9.
41. Huang L, Snyder AR, Morgan WF. Radiation-induced genomic instability and its implications for radiation carcinogenesis. *Oncogene* 2003;22: 5848–54.

Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer

Christina M. Wilke ¹, Julia Hess^{1,2}, Sergiy V. Klymenko⁴, Vadim V. Chumak⁴, Liubov M. Zakhartseva⁵, Elena V. Bakhanova⁴, Annette Feuchtinger⁶, Axel K. Walch⁶, Martin Selmansberger¹, Herbert Braselmann^{1,2}, Ludmila Schneider^{1,2}, Adriana Pitea^{1,7}, Julia Steinhilber⁸, Falko Fend⁸, Hans C. Bösmüller⁸, Horst Zitzelsberger^{1,2,3} and Kristian Unger^{1,2}

¹Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany

²Clinical Cooperation Group 'Personalized Radiotherapy in Head and Neck Cancer', Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, 85764, Germany

³Department of Radiation Oncology, University Hospital, LMU Munich, München, Germany

⁴National Research Center for Radiation Medicine of National Academy of Medical Sciences of Ukraine, Kyiv, Ukraine

⁵Bogomolets National Medical University, Kyiv, Ukraine

⁶Research Unit Analytical Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany

⁷Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany

⁸Institute of Pathology and Neuropathology, Tübingen, Germany

Ionizing radiation is a well-recognized risk factor for the development of breast cancer. However, it is unknown whether radiation-specific molecular oncogenic mechanisms exist. We investigated post-Chernobyl breast cancers from radiation-exposed female clean-up workers and nonexposed controls for molecular changes. Radiation-associated alterations identified in the discovery cohort ($n = 38$) were subsequently validated in a second cohort ($n = 39$). Increased expression of hsa-miR-26b-5p was associated with radiation exposure in both of the cohorts. Moreover, downregulation of the TRPS1 protein, which is a transcriptional target of hsa-miR-26b-5p, was associated with radiation exposure. As TRPS1 overexpression is common in sporadic breast cancer, its observed downregulation in radiation-associated breast cancer warrants clarification of the specific functional role of TRPS1 in the radiation context. For this purpose, the impact of TRPS1 on the transcriptome was characterized in two radiation-transformed breast cell culture models after siRNA-knockdown. Deregulated genes upon TRPS1 knockdown were associated with DNA-repair, cell cycle, mitosis, cell migration, angiogenesis and EMT pathways. Furthermore, we identified the interaction partners of TRPS1 from the transcriptomic correlation networks derived from gene expression data on radiation-transformed breast cell culture models and sporadic breast cancer tissues provided by the TCGA database. The genes correlating with TRPS1 in the radiation-transformed breast cell lines were primarily linked to DNA damage response and chromosome segregation, while the transcriptional interaction partners in the sporadic breast cancers were mostly associated with apoptosis. Thus, upregulation of hsa-miR-26b-5p and downregulation of TRPS1 in radiation-associated breast cancer tissue samples suggests these molecules representing radiation markers in breast cancer.

Breast cancer is one of the most common cancers in women worldwide. Besides risk factors such as age and lifestyle, it is well-recognized that breast cancer risk increases with exposure to ionizing radiation. Patients with preceding radiotherapy for the treatment of Hodgkin lymphoma exhibit an increased risk to develop breast cancer as a secondary tumor. In the Japanese atomic bomb survivors cohort, a similar finding has been reported for women who were exposed

Key words: TRPS1, hsa-miR-26b-5p, Chernobyl, breast cancer, radiation-associated

Abbreviations: FDR: false discovery rate; FFPE: formalin-fixed paraffin-embedded; GO: gene ontology; IHC: immunohistochemistry; ILC: invasive lobular carcinoma; NST: invasive carcinoma of no special type; PTC: papillary thyroid carcinoma; PVDF: polyvinylidene fluoride; qRT-PCR: quantitative real-time reverse transcription polymerase chain reaction; SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis; SKY: spectral imaging; TBST: Tris-buffered saline Tween20; TNM: primary tumor, lymph node metastases, distant metastases; TRPS1: trichorhinophalangeal syndrome type 1

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Bundesamt für Strahlenschutz; **Grant numbers:** 3615S32454, 3611S30019

DOI: 10.1002/ijc.31072

History: Received 21 Mar 2017; Accepted 31 Aug 2017; Online 25 Sep 2017

Correspondence to: Kristian Unger, Helmholtz Zentrum München German Research Center for Environmental Health GmbH, Research Unit Radiation Cytogenetics, Neuherberg, Bayern, Germany, E-mail: unger@helmholtz-muenchen.de

What's new?

While ionizing radiation is an established risk factor for breast cancer, little is known about mechanisms of radiation-specific breast carcinogenesis related to low-dose exposure. Here, investigation of molecular changes in breast cancers from female post-Chernobyl clean-up workers exposed to radiation revealed two radiation-specific molecular markers: increased expression of hsa-miR-26b-5p and downregulation of its target TRPS1. In human radiation-transformed breast cells, TRPS1 knockdown was found to be associated with enrichment of DNA repair, cell cycle, mitosis, angiogenesis, migration and EMT pathways. Further investigation of specific markers could facilitate the identification of radiation-induced breast cancer and potentially provide a basis for individualized therapy.

to ionizing radiation during adolescence.¹ Also in the aftermath of the Chernobyl accident in 1986, a significant increase of the breast carcinoma rate (standardized incidence ratio 190.6%) in female clean-up workers was noticed in comparison to sporadic breast cancer rates in Ukraine.^{2,3} To date, despite this epidemiologic evidence, the knowledge about radiation-specific mechanisms of breast carcinogenesis after low-dose exposure is sparse.

In contrast to environmental exposures of patients from this study, aberrant expressions of miRNAs after exposure to therapeutic doses of ionizing radiation have already been reported.⁴ miRNAs are 19–25 nucleotides long, noncoding, highly conserved RNA molecules, that are known to play an important role in the regulation of gene expression at the post-transcriptional level. Numerous studies have shown a deregulation of miRNAs in tumors, thereby demonstrating that miRNAs are involved in the process of carcinogenesis and act as oncogenes or as tumor suppressors.⁵ Breast cancer-specific miRNA profiles have been observed for different breast cancer subtypes, enabling a classification into different molecular subtypes.⁶ However, the role of miRNAs in radiation-associated breast cancer has not been investigated so far. Therefore, our study intended to investigate the miRNA profiles of breast cancers from a cohort of female clean-up workers who were exposed to ionizing radiation from the Chernobyl reactor accident and nonexposed controls matched for residence, tumor type, age at diagnosis, TNM classification and grading. We explored if among miRNAs that are known to play a role in sporadic breast cancer there are specifically radiation-associated ones. We discovered that expression of hsa-miR-26b-5p was increased in radiation-associated breast cancers compared to nonexposed controls. Further, we showed that expression of one of the hsa-miR-26b-5p target proteins TRPS1 was significantly decreased in radiation-exposed cases. TRPS1 is a GATA-type transcription factor and consists of nine zinc-finger domains, including a single GATA-type DNA-binding domain. Either mutation or deletion of this gene causes a disease called trichorhinophalangeal syndrome. Previous studies have shown that TRPS1 is expressed in several human malignant tumors and implied an important function in tumor growth, cell cycle, angiogenesis, apoptosis, cell proliferation, migration and metastasis.^{7–13}

In this study, we were able to identify for the first time one miRNA and one of its target proteins to be significantly associated with radiation-associated breast cancer.

Material and Methods**Patients tumor tissues and clinical data**

Formalin-fixed paraffin-embedded (FFPE) breast cancer tissue samples ($n = 76$) from 38 female Ukrainian patients that were exposed to radiation after the Chernobyl reactor accident and a matched set of 38 breast cancer samples from nonexposed patients from Ukraine were collected (discovery set). The vast majority (34 out of 38) of patients have been exposed as clean-up workers after the accident for which an elevated breast cancer incidence has been reported.^{2,3} Four patients were exposed as evacuees after the accident. The mean age at time of exposure was 33 years (range 18–45 years), the mean age at time of diagnosis was 49 years (range 33–59 years), and the mean latency of tumors was 17 years. None of the patients from the discovery set received neoadjuvant radio(chemo)therapy (Table 1).

A validation cohort consisting of FFPE breast cancer tissue samples, 39 from post-Chernobyl clean-up workers and 39 matched nonexposed Ukrainian control cases, was established. The mean age at time of exposure was 33 years (range 23–48 years) and the mean age at diagnosis 51 years (range 35–59 years) and the mean latency was 18 years. Out of 78 patients of the validation cohort, 18 received a neoadjuvant radio(chemo)therapy (Table 1).

The so-called RADRUE method, which was adapted specifically for estimation of breast doses, was used for reconstruction of the absorbed doses of the exposed breast cancer patients.¹⁴ Information about the absorbed doses were only available for a subset of the exposed breast cancer patients ($n = 54$). The absorbed doses showed a large interindividual variability between patients ranging from 0.06 to 929 mGy (median 8.53 mGy) in the clean-up workers and from 5.72 to 36.85 mGy (median 18.68 mGy) in the evacuees (unpublished data).

In both cohorts, all patients were younger than 60 years at the time of diagnosis. Exposed patients and nonexposed controls for this case–case study were frequency matched for residence, tumor subtype, age at diagnosis, TNM-classification and histological grading. The majority of tumors was diagnosed as invasive carcinoma of no special type (NST; discovery cohort:

Table 1. Patient characteristics of the Chernobyl discovery and validation cohort

Characteristics	Chernobyl Discovery Cohort			Chernobyl Validation Cohort			
	Exposed	Not exposed	<i>p</i> value ¹	Exposed	Not exposed	<i>p</i> value ¹	
Number of patients	38	38		39	39		
Tumor type, no. (%)							
	invasive carcinoma of no special type	36 (95)	36 (95)	1	35 (90)	35 (90)	1
	lobular	1 (3)	1 (3)		3 (8)	3 (8)	
	intracystic	0 (0)	0 (0)		1 (3)	1 (3)	
	medullar	1 (3)	1 (3)		0 (0)	0 (0)	
Estrogen-receptor status, no. (%)							
	positive	24 (63)	24 (63)	1	28 (72)	30 (77)	0.8
	negative	14 (37)	14 (37)		11 (28)	9 (23)	
Progesteron-receptor status, no. (%)							
	positive	22 (58)	26 (68)	0.48	26 (67)	29 (74)	0.62
	negative	16 (42)	12 (32)		13 (33)	10 (26)	
C-kit status, no. (%)							
	positive	7 (18)	6 (16)	1	2 (5)	3 (8)	1
	negative	31 (82)	32 (84)		37 (95)	36 (92)	
Cytokeratin 5/6 status, no. (%)							
	positive	7 (18)	7 (18)	1	6 (15)	1 (3)	0.11
	negative	31 (82)	31 (82)		33 (85)	38 (97)	
P53 status, no. (%)							
	positive	13 (34)	15 (39)	0.81	24 (62)	22 (56)	0.82
	negative	25 (66)	23 (61)		15 (38)	17 (44)	
Ki-67 status, no. (%)							
	positive	31 (82)	32 (84)	1	39 (100)	39 (100)	1
	negative	7 (18)	6 (16)		0 (0)	0 (0)	
BRCA1/2 status, no. (%)							
	positive	2 (5)	3 (8)	0.67	4 (10)	3 (8)	1
	negative	36 (95)	34 (89)		35 (90)	36 (92)	
	not evaluable	0 (0)	1 (3)		0 (0)	0 (0)	
Her2 status, no. (%)							
	positive	2 (5)	4 (11)	0.67	6 (15)	8 (21)	0.77
	negative	36 (95)	34 (89)		29 (74)	31 (79)	
	not evaluable	0 (0)	0 (0)		4 (10)	0 (0)	
pT stage, no. (%)							
	pT1	21 (55)	20 (53)	0.88	11 (28)	12 (31)	0.85
	pT2	14 (37)	16 (42)		27 (69)	25 (64)	
	pT3	3 (8)	2 (5)		1 (3)	2 (5)	
pN stage, no. (%)							
	pN0	24 (63)	24 (63)	1	16 (41)	17 (44)	1
	PN1	13 (34)	14 (37)		18 (46)	19 (49)	
	pN2	1 (3)	0 (0)		3 (8)	3 (8)	
	pN3	0 (0)	0 (0)		1 (3)	0 (0)	
	pNx	0 (0)	0 (0)		1 (3)	0 (0)	
pM stage, no. (%)							
	M0	38 (100)	38 (100)	1	39 (100)	39 (100)	1
Grade, no. (%)							
	G1	1 (3)	1 (3)	1	2 (5)	2 (5)	1
	G2	24 (63)	24 (63)		26 (67)	26 (67)	
	G3	13 (34)	13 (34)		11 (28)	11 (28)	

¹The *p* values were calculated by Fisher's-exact test.

95%, validation cohort: 90%) and invasive lobular carcinoma (ILC; discovery cohort: 2.5%, validation cohort: 8%). Two cases were diagnosed as intracystic papillary breast carcinoma and another two as breast carcinomas with medullary features. Immunohistochemical staining for estrogen and progesterone receptors, C-kit, Cytokeratin 5/6, TP53 and Ki67 antigen expression and HER2 gene status determination by fluorescence

in situ hybridization (FISH) is described in the Supporting Information, Material and Methods part.

Information of all clinicopathologic characteristics of the discovery and validation cohort is presented in Supporting Information, Tables S1 and S2.

Total RNA including the small RNA fraction was isolated using the Qiagen RNeasy FFPE Kit (Qiagen, Hilden, Germany).

Small RNA (miRNA) integrity was analyzed by qRT-PCR of the small noncoding RNA RNU24 using TaqMan chemistry (Life technologies, Carlsbad, CA). Samples with Ct values <35 were considered suitable for analysis.

Fisher's exact test was used to test associations of the exposure status with any clinical characteristics of the patients such as estrogen-receptor status, progesterone-receptor status, cytokeratin-expression status (positive/negative), C-kit-expression status (positive/negative), Ki67-expression status (positive/negative), Her2/neu-status, p53-mutation status, BRCA1/2-mutation status, pT-status, pN-status and grading. Significance was accepted for $p < 0.05$.

Quantitative real-time RT-PCR (qRT-PCR)

Reverse transcription of miRNAs was performed using the MicroRNA Reverse Transcription Kit and microRNA-specific stem-loop primers according to the manufacturer's protocol (Life Technologies). TaqMan MicroRNA assays (Life Technologies) for the following miRNAs were used: *hsa-miR-222-3p* (477982_mir), *hsa-miR-221-3p* (477981_mir), *hsa-miR-372-3p* (478071_mir), *hsa-miR-26b-5p* (478418_mir), *hsa-miR-302d-3p* (478237_mir), *hsa-miR-124-3p* (477879_mir), *hsa-miR-1-3p* (477820_mir) and *hsa-miR-99b-5p* (478343_mir). For endogenous normalization, the assays for *RNU44* (001094) and *RNU48* (001006) were used. qRT-PCR reactions (20 μ l) were carried out in triplicates using the ViiA 7 Real Time PCR System in combination with the ViiA 7 Software v.1.2.2 following the manufacturer's protocol (Life Technologies). Relative expressions were calculated using the $\Delta\Delta$ Ct method. The partial differential test considering intertumor heterogeneity was used to test for statistical significant differences of miRNA expressions between exposed and nonexposed samples and possible associations of miRNA expression with clinicopathological data.¹⁵

TRPS1 (Hs00232645_m1) TaqMan gene expression assay (LifeTechnologies) was used to validate the TRPS1-knockdown and to determine the TRPS1-knockdown efficacy in B42-11 and B42-16 cells at gene expression level.

For technical validation of the gene expression microarray data, qRT-PCR was performed for randomly selected genes ($n = 12$) detected by gene expression microarray in B42-11 and B42-16 cells: *ANXA1* (Hs00167549_m1), *APRT* (Hs00975725_m1), *BBC3* (Hs00248075_m1), *BMP2* (Hs01055564_m1), *CLNS1A* (Hs00818054_m1), *DTL* (Hs00978565_m1), *DUSP6* (Hs00169257_m1), *F2R* (Hs00169258_m1), *PLK2* (Hs01573405_g1), *RFC5* (Hs00738859_m1), *TRPS1* (Hs00232645_m1) and *TUBB3* (Hs00801390_s1). For endogenous normalization, the assays for *ACTB* (Hs99999903_m1) and *B2M* (Hs99999907_m1) were used. RNA was reverse transcribed using the QuantiTect Reverse Transcription Kit (Qiagen). qRT-PCR reactions (10 μ l) and calculations of relative expressions were carried out as described above. For technical validation of the gene expression microarray data, Pearson correlation analyses of expression determined by qRT-PCR with that determined by microarray were

performed. Validation was considered successful for correlation coefficients >0.5 and p values <0.05.

Immunohistochemistry

The expression of the TRPS1 protein in both tumor cohorts was measured by immunohistochemical staining (IHC) of FFPE tumor sections with a primary antibody against TRPS1 (Abcam: ab111439, Cambridge, UK). The antibody was selected from Abcam with information about antibody specificity and staining patterns.⁹ The primary antibody was used in a dilution of 1:100 and Discovery-Universal (Roche, Ventana, Tucson, AZ) as a secondary antibody. IHC staining was performed with the automated staining instrument Discovery XT (Roche, Ventana) system using peroxidase-DAB-(diaminobenzidine)-MAP chemistry (Roche, Ventana) for signal detection. The stained tissue sections were fixed in an ethanol series and coated by a coverslip. All stained slides were scanned at 20 \times objective magnification using the Leica SCN400 digital slide scanning system (Leica, Houston, TX).

Digital image analysis

The evaluation of the immunohistochemical staining was performed using the digital image analysis platform DefiniensTissueStudio 3.5 (Definiens AG, Munich, Germany). For this purpose, the digital slide images were imported into the image analysis software. In the first step regions of interest, that is tumor area, were manually defined. A specific rule set was then created to detect and quantify the TRPS1-stained nuclei within the annotated tissue areas. The quantified parameters were the amount and the mean brown intensity of TRPS1-positive nuclei per annotated tissue area. The averaged TRPS1 staining intensities were tested for significant differences between exposed and nonexposed samples and possible associations of TRPS1 staining intensities with clinicopathological data using partial differential testing, which considers intertumor heterogeneity.¹⁵ p values <0.05 were considered statistically significant.

B42-11 and B42-16 cell lines and spectral karyotyping (SKY)

Human B42-11 and B42-16 radiation transformed breast cells were grown in mammary epithelial growth medium (MEGM) as published previously.¹⁶ The B42-11 and B42-16 cell lines were authenticated by STR-typing and spectral karyotyping (SKY). Metaphase chromosome spreads were prepared and hybridized as described earlier.¹⁷ SKY image analysis was performed with a SpectraCube system and SkyView imaging software (Applied Spectral Imaging).

RNA interference

The B42-11 and B42-16 cells were seeded into six-well plates and were transfected at 70–90% confluency in triplicates with a nonsense scrambled control (Ambion, Carlsbad, CA; Negative control #1) or two specific siRNAs against TRPS1

(Ambion, silencer select siRNA 1: ID: s14428 and siRNA 2: ID: s14427). SiRNA transfections were performed using Lipofectamine RNAiMAX (Life Technologies) according to the manufacturer's instructions. 7.5 μ l lipofectamine and 3.75 μ l of TRPS1 siRNA were used per sample resulting in a siRNA concentration of 75 pmol per well. After 24, 48, 72 and 96 hrs, cells were harvested for total RNA isolation using the Qiagen RNeasy Mini Kit (Qiagen). In addition, protein lysates were generated 48 and 72 hrs after transfection to verify TRPS1-knockdown efficacy by Western blot analysis.

Western blot analysis

Western blot analysis with an antibody against TRPS1 (Abcam: ab111439) was performed to monitor the TRPS1 knockdown at protein level. RIPA-buffer (150 mM NaCl, 1% NP-40, 10 mM MDOC, 0.1% SDS, 50 mM Tris pH 8.0 supplemented with protease, phosphatase and HDAC inhibitors) was used for protein extraction which was performed on ice. Twenty-five micrograms of total protein was used for each Western blot analysis. The proteins were separated on a 10% SDS-PAGE. PVDF-membranes were cut and blocked with 8% skim milk buffer after immunoblotting followed by incubation over night at 4°C with primary antibodies (rabbit polyclonal anti-TRPS1, Abcam: ab111439; 1:2000; mouse monoclonal anti- β -Actin, Sigma: A5441; 1:10000) diluted in Roti-Block (Roth). After four washing steps with TBST-buffer (5 min each), the PVDF-membranes were incubated for 2 hrs with a secondary antibody (anti-rabbit IgG, Jackson ImmunoResearch; 1:50000, anti-mouse IgG Jackson ImmunoResearch; 1:50000), diluted in 8% skim milk buffer. Blots were developed with Amersham ECL Select Western Blotting Detection Reagent (GE Healthcare, Little Chalfont, United Kingdom). Chemiluminescence was detected and images were acquired with a FluorChem HD2 documentation system from Alpha Innotech in combination with the AlphaView software (Biozym, Oldendorf, Germany).

Microarray-based gene expression analysis

To investigate the effect of TRPS1-knockdown on the transcriptome, mRNA microarray expression profiling of biological triplicates of cells after TRPS1-knockdown, a nonsense scrambled control and the B42-11 and B42-16 untreated cell lines 48 hrs after transfection was performed using G3 Human Gene Expression 8x60k v2 microarrays (AMADID 72363, Agilent Technologies, Santa Clara, CA). RNA quality was assessed prior to expression analysis using an Agilent 2100 Bioanalyzer (Agilent Technologies). The obtained RNA integrity numbers (RINs) ranged from 6.7 to 9.7. The analysis was performed according to the manufacturer's instructions using 50 ng of total RNA. Microarrays were scanned using a G2505C Sure Scan Microarray Scanner (Agilent Technologies) followed by raw data extraction with the Feature Extraction 10.7 software (Agilent Technologies). Data quality assessment, preprocessing and normalization were conducted in R using the Bioconductor AgiMicroRNA package.¹⁸ Statistical analyses were

performed using functions from the Bioconductor limma package for the identification of significantly differentially expressed genes after TRPS1-knockdown (siRNA 1 and siRNA 2 taken together) compared to the nonsense scrambled control.¹⁹ A cutoff for FDR-adjusted *p* values of 0.05 and minimum absolute log₂-fold change of 0.5 was applied. Significantly deregulated genes after TRPS1 knockdown were subjected to pathway enrichment analysis using the Cytoscape Reactome Functional Interaction (FI) plugin (version 2016) within the Cytoscape network visualization software (version 3.5.1).^{20,21} For pathway enrichment analysis, only network modules containing more than three genes were considered. The top 50 pathways with an FDR-adjusted *p* values <0.05 were considered for further interpretation.

TRPS1-centered correlation network

To explore potential direct and indirect interaction partners of TRPS1 at the transcriptome level, we generated gene correlation networks from the microarray gene expression data on B42-11 and B42-16 untransfected, scrambled-siRNA transfected and TRPS1-downregulated cells and from global mRNA expression data on sporadic breast cancers of the publicly available The Cancer Genome Atlas (TCGA) breast cancer dataset.^{22,23} The latter of which were matched to the breast cancer post-Chernobyl cohort for the parameters tumor type, hormone receptor status, age, TNM-classification, grading, BRCA1/2- and Her2-status. For both data sets, correlation (Pearson) of the TRPS1 expression vector and all other genes was determined and a correlation test was applied. The resulting *p* values were corrected for multiple-testing error determining the Benjamini-Hochberg FDR.²⁴ A cutoff for FDR-adjusted *p* values of 0.05 was applied. The top 100 correlating genes were selected and subjected to GO-term and pathway enrichment analysis using the ClueGo plugin (version 2.3.2, 2016) of the Cytoscape network analysis software (version 3.0.2).^{21,25} The top 50 pathways with an FDR-adjusted *p* value <0.05 were considered for further interpretation.

Results

Selection of candidate miRNAs

We explored the literature by PubMed research and identified the following miRNAs to be most frequently published as being associated with breast cancer and radiation exposure: hsa-miR-26b-5p, hsa-miR-99b-5p, hsa-miR-221-3p and hsa-miR-222-3p.^{13,26–29} Commonly regulated target genes of these miRNAs were identified using MiRTarBase (version 4.0, 2014) and revealed the gene TRPS1 (The trichorhinophalangeal syndrome 1).³⁰ According to MiRTarBase (version 4.0, 2014), TRPS1 is regulated by additional four miRNAs: hsa-miR-124-3p, hsa-miR-302d-3p, hsa-miR-1-3p and hsa-miR-372-3p. We selected these eight TRPS1-regulating miRNAs and the target protein TRPS1 for further analysis in the discovery and validation cohorts.

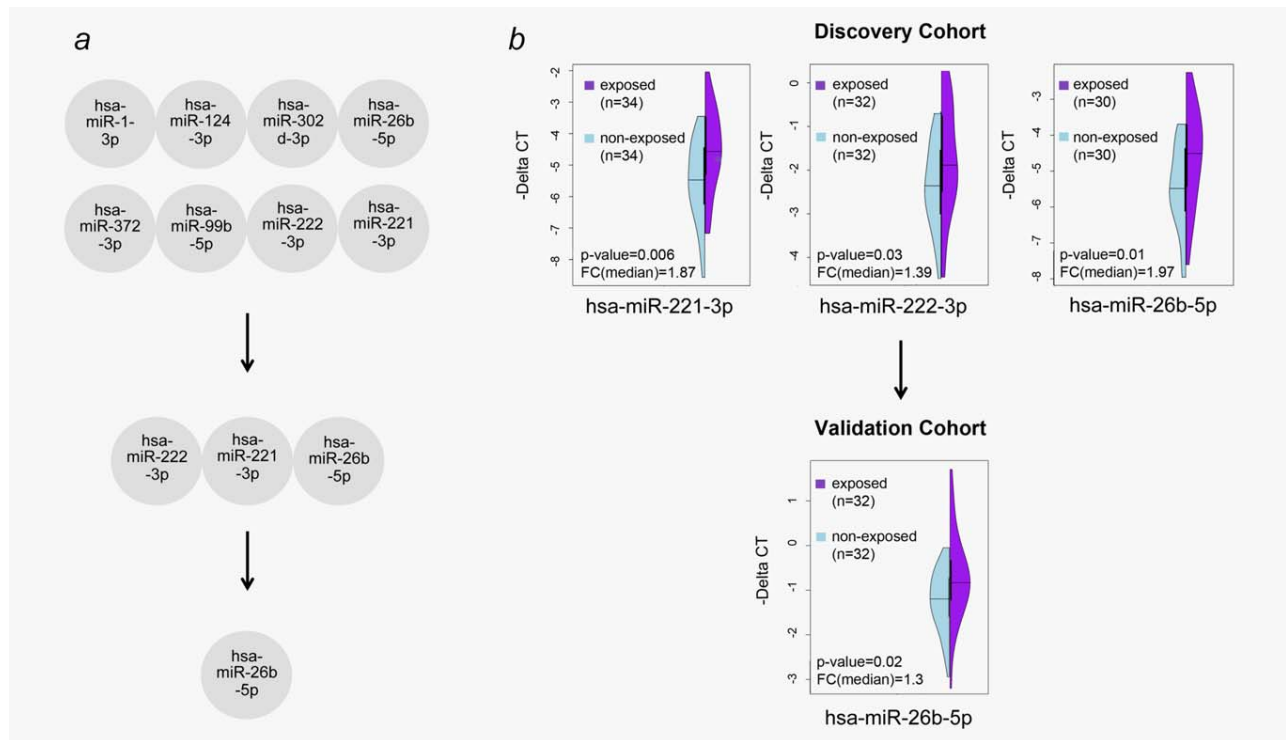


Figure 1. (a) The expression levels of all eight TRPS1-regulating miRNAs were analyzed in the Chernobyl discovery cohort by qRT-PCR. Hsa-miR-222-3p, hsa-miR-221-3p and hsa-miR-26b-5p showed a significant differential expression between exposed and nonexposed samples. The expression levels of these three microRNAs were also tested in the Chernobyl validation cohort. The expression of hsa-miR-26b-5p was associated with exposure to ionizing radiation in the validation cohort. (b) Violin plots displaying the expressions of hsa-miR-26b-5p, hsa-miR-221-3p and hsa-miR-222-3p in the Chernobyl discovery cohort and hsa-miR-26b-5p in the Chernobyl validation cohort measured by qRT-PCR ($-\Delta\text{CT}$ values) are shown (right panel). The nonexposed control group is labeled in light blue and the exposed group in purple. The middle dark line represents the median of expression values. The vertical black line represents the interquartile. [Color figure can be viewed at wileyonlinelibrary.com]

Increased hsa-miR-26b-5p expression is associated with radiation exposure

The analysis of the selected miRNAs was performed by qRT-PCR and subsequent partial differential testing between the exposed and nonexposed tumor sets. Hsa-miR-221-3p (FC = 1.87, partial differential test $p = 0.006$), hsa-miR-222-3p (FC = 1.39, partial differential test $p = 0.03$) and hsa-miR-26b-5p (FC = 1.97, partial differential test p value = 0.01) were significantly upregulated in the exposed compared to the nonexposed tumor set of the discovery cohort. The other miRNAs did not show statistically significant deregulation between exposed cases and controls. From the three miRNAs that were found to be significantly associated with radiation exposure in the discovery cohort, upregulation of hsa-miR-26b-5p could be confirmed in the exposed cases of the validation cohort (FC = 1.3, partial differential test $p = 0.02$, Figs. 1a and 1b). Hsa-miR-26b-5p expression was not associated with estrogen-receptor status, progesterone-receptor status, cytokeratin-expression (positive/negative), C-kit-expression (positive/negative), Ki67-expression (positive/negative), Her2/neu-status, TP53-status and BRCA1/2-mutation status in the discovery or the validation cohort. Moreover, no dose-response effect was observed for hsa-miR-26b-5p (data not shown). We also tested if the exposure status was associated with any clinical

characteristics of the patients, whereby no significant association between exposure status and any of the clinical characteristics could be detected (Table 1).

Decreased TRPS1 protein expression is associated with radiation exposure

The expression of the TRPS1 protein, which was identified as a target of the literature-derived candidate miRNAs, was determined by immunohistochemical staining of serial FFPE tissue sections and subsequently tested for association with radiation exposure. After software-based quantification of staining intensities a significant downregulation of TRPS1 protein expression in breast cancer tissues from exposed patients was detected (partial differential test $p = 0.028$). This finding was confirmed in the validation cohort (partial differential test $p = 0.027$). Visualization of these results can be found in Figures 2 and 3a and 3b. Furthermore, no dose-response effect was observed for TRPS1 (data not shown).

Association of TRPS1 expression with clinical and histological data

For all tumor samples of the discovery and validation cohorts, an association of the TRPS1 protein expression with other clinical parameters was tested (partial differential test). TRPS1

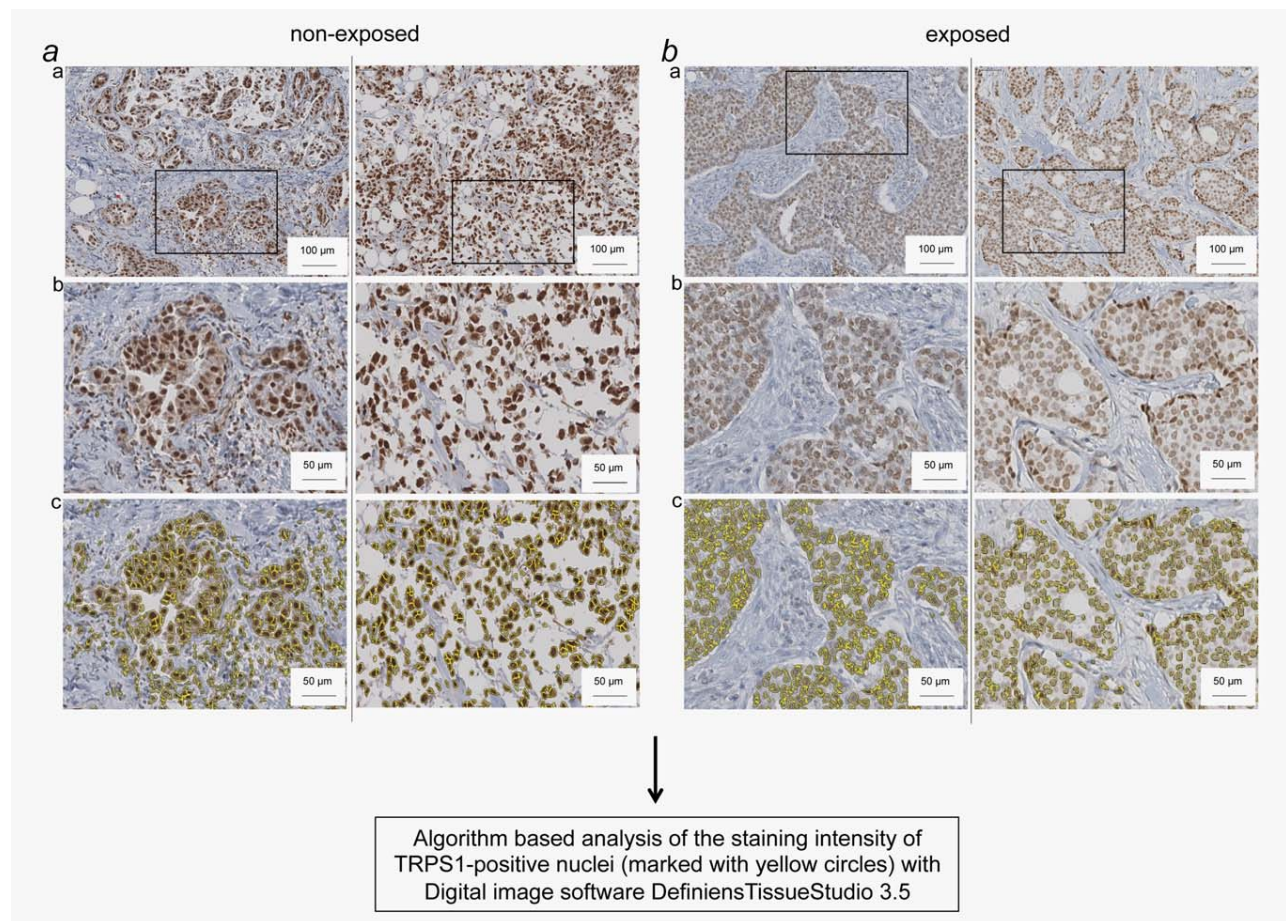


Figure 2. Digital image analysis of immunohistochemically stained FFPE tumor sections from non-exposed and exposed breast cancer samples using an antibody against TRPS1. (a/b) Two representative immunohistochemically stained breast carcinoma cases are shown for non-exposed (a) and exposed (b) cases. Image details of Aa and Ba (black frames) are shown in Ab and Bb. Detection and quantification of TRPS1-stained nuclei was performed using the digital image software Definiens. Nuclei of tumor cells, for which the staining intensities were calculated based on the algorithm, are labeled in yellow (Ac, Bc). [Color figure can be viewed at wileyonlinelibrary.com]

protein expression was not associated with estrogen-receptor status, progesterone-receptor status, cytokeratin-expression (positive/negative), C-kit-expression (positive/negative), Ki67-expression (positive/negative), Her2/neu-status, TP53-status and BRCA1/2-mutation status in the discovery and the validation cohort, suggesting an independent association of TRPS1 downregulation with radiation exposure of patients.

Characterization of the B42-11 and B42-16 cell lines

SKY analysis revealed the following karyotype for B42-16 resulting from evaluation of 15 metaphases: 47,XX,der(4)t(4;12)(p31;?),i(8)(q10),+der(8)t(8;10)(q21;?),der(10)t(8;10;12)(?:p12;q23;?),der(12)t(8;10;12)(?:q22) and for B42-11:47,XX,+i(8)(q10),der(7)t(7;10)(q11.1;11.2). A representative metaphase for each is shown in Supporting Information, Figure S1.

TRPS1 knockdown in B42-11 and B42-16 cells

To characterize the impact of TRPS1 on the transcriptome in radiation-transformed breast cells, siRNA-knockdown of TRPS1 was performed in the radiation-transformed breast

cell lines B42-11 and B42-16. The knockdown reached a maximum after 48 hrs (Fig. 4 and Supporting Information, Fig. S2); therefore, this timepoint was chosen for differential expression analysis between TRPS1-knockdown and scrambled control of B42-11 and B42-16 cells. The analysis revealed 281 significantly differentially expressed microarray probes (144 downregulated and 137 upregulated) relating to 267 different genes (Supporting Information, Table S3). Randomly selected genes ($n = 12$) detected by gene expression microarray in B42-11 and B42-16 cells were chosen for technical validation of the microarray data. Correlation analysis of expression of the genes selected for validation determined by qRT-PCR and mRNA microarray showed strong correlation for ten out of 12 analyzed genes (Supporting Information, Table S4). Furthermore, pathway enrichment analysis was conducted based on the Reactome network, resulting in nine modules containing the 267 significantly deregulated genes after TRPS1 knockdown. Significantly enriched pathways involving DNA-repair, cell cycle, mitosis, cell migration, angiogenesis and EMT were detected (Supporting

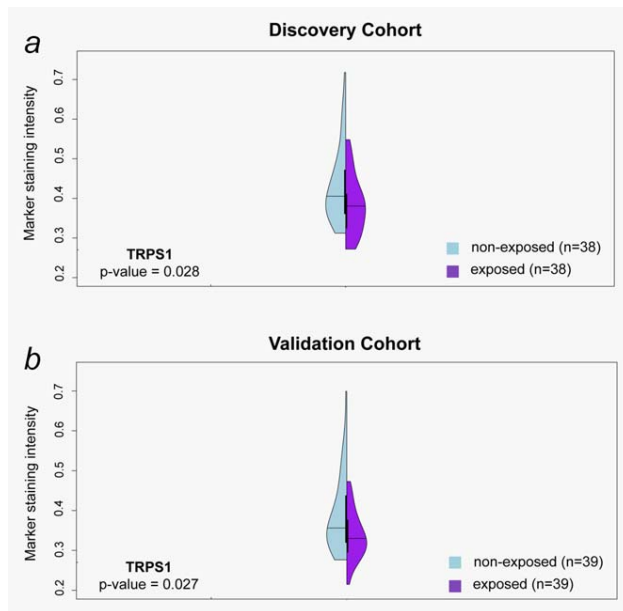


Figure 3. Significantly increased TRPS1 protein expression represented by the marker staining intensity was observed in breast cancer tissues from the nonexposed groups (light blue) compared to the exposed groups (purple) in the discovery (*a*, $p = 0.028$) and validation cohorts (*b*, $p = 0.027$). *p* values were calculated using the partial differential test considering intertumor heterogeneity. [Color figure can be viewed at wileyonlinelibrary.com]

Information, Table S5). Downregulated genes were mainly involved in DNA-repair, cell cycle and mitosis while upregulated genes mostly showed up in cell migration, angiogenesis and EMT pathways (Supporting Information, Table S5).

TRPS1-centered correlation network

To explore putative direct and indirect interaction partners of TRPS1 in the sporadic and radiation-associated context at the transcriptome level, two TRPS1-centered correlation networks were generated and subsequently analyzed for involved pathways. To examine the role of TRPS1 in sporadic breast cancer, we deployed the RNAseq-derived global gene expression data set on breast cancer from the The Cancer Genome Atlas (TCGA) dataset.^{22,23} From the 1106 available cases, a subset that matched our radiation-associated breast cancer cohort ($n = 382$) was used. In total, 12,106 genes showed a statistical significant correlation with TRPS1 expression in sporadic breast cancers of the publicly available TCGA dataset and 1,270 genes in the B42-11 and B42-16 cells ($FDR < 0.05$) (Supporting Information, Table S6).

From both correlation networks, we selected the top 100 correlating genes with regard to FDR (Figs. 5*a* and 5*b* and Supporting Information, Table S6). GO and pathway enrichment analysis including the top 100 correlating genes of the sporadic breast cancer correlation network revealed mainly significant enrichment of apoptosis related pathways such as *TRADD-TRAF2:RIP1 complex binds FADD* and *RIPK1 is*

deubiquitinated. The radiation-associated cell lines B42-11 and B42-16 showed mainly significant enrichment of GO terms related to the process of chromosome segregation and DNA repair.

Discussion

Radiation-specific markers have already been reported in young patients suffering from papillary thyroid carcinomas in the aftermath of the Chernobyl accident.³¹ Although ionizing radiation is also known to be a risk factor for the development of breast cancer, radiation-specific markers in these tumors are still unknown.^{2,3,32} This study aimed for the discovery of radiation-specific changes of miRNA and protein expressions in breast cancer samples from Ukrainian cleanup workers, who were exposed to ionizing radiation from the Chernobyl accident by comparison with nonexposed Ukrainian control cases matched for age and clinical parameters.

From the published literature, we identified four miRNAs (hsa-miR-26b-5p, hsa-miR-99b-5p, hsa-miR-221-3p and hsa-miR-222-3p) that were associated with breast cancer and radiation exposure.^{13,26–29} The TRPS1 gene was recognized as a common target gene that is regulated by additional four miRNAs (hsa-miR-124-3p, hsa-miR-302d-3p, hsa-miR-1-3p and hsa-miR-372-3p).³⁰ The eight TRPS1-regulating miRNAs in total along with the TRPS1 protein were investigated on two independent post-Chernobyl breast cancer cohorts from cleanup workers. Consistently, a significant coregulation of hsa-miR-26b-5p in exposed compared to matched nonexposed patients became apparent in both cohorts and thus, an association of hsa-miR-26b-5p with radiation exposure could be validated independently (Fig. 1). Hsa-miR-26b-5p plays a pivotal role in sporadic breast cancer.²⁹ In sporadic breast cancer, decreased hsa-miR-26b-5p expression was reported, and could be confirmed in our sporadic breast cancer control cases. Hsa-miR-26b-5p obviously plays a tumor-suppressive role by the promotion of apoptosis and the suppression of cell growth.^{29,33} An opposed observation in post-Chernobyl cases points to a radiation-specific deregulation of hsa-miR-26b-5p and renders the question whether TRPS1 is also affected. Surprisingly, also the TRPS1 expression was significantly downregulated in both exposed breast cancer cohorts compared to the nonexposed cohorts. As this finding was confirmed in two independent cohorts, it suggests an important role of TRPS1 in radiation-associated breast cancer (Figs. 2 and 3). To our knowledge TRPS1 and hsa-miR-26b-5p alterations have not been investigated in radiation-associated breast cancers so far. In sporadic breast cancer, an upregulated TRPS1 expression was previously reported which is in line with our findings in the sporadic subset of control cases.³⁴ In sporadic breast cancer TRPS1 is linked to the stimulation of cell proliferation and angiogenesis and the promotion of cell cycle progression.^{7,10,12} Furthermore, TRPS1 overexpression was proposed as a prognostic marker in early stage breast cancer due to an association with improved overall survival and disease-free survival in these tumors.³⁵ Moreover, TRPS1 expression was found to correlate with ER,

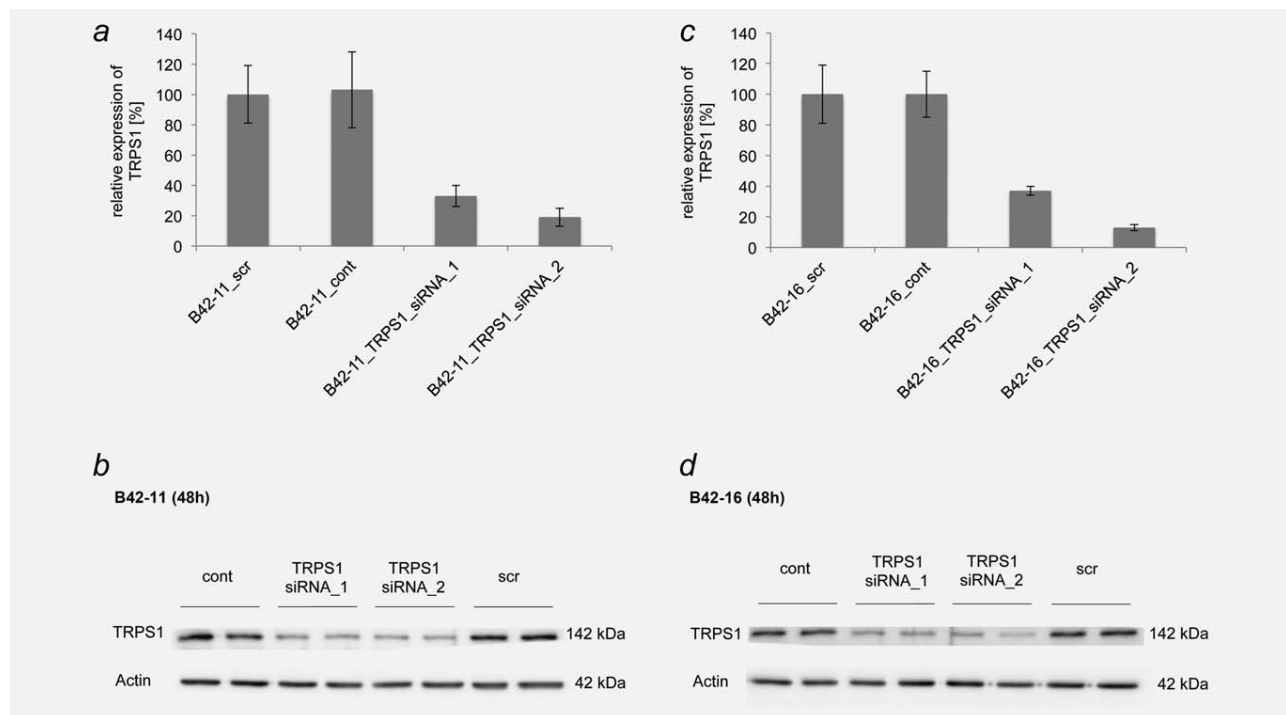


Figure 4. (a/c) Levels of TRPS1-mRNA-expression in untransfected (cont), scrambled-siRNA transfected (scr) and TRPS1-siRNA transfected B42-11 and B42-16 cells 48 hrs after transfection. (b/d): Western Blot images show levels of TRPS1-protein expression in untransfected (cont), scrambled-siRNA transfected (scr) and TRPS1-siRNA transfected B42-11 and B42-16 cells 48 hrs after transfection.

PgR, Ki67, GATA-3 and Her2 expression, which we could not confirm in our data.^{36,37} At the same time, TRPS1 acts as a negative regulator of EMT and thus could reduce the metastatic potential of breast cancers by suppressing transcriptionally the processes of migration and invasion.^{11,13} Taken together the published data on TRPS1 overexpression in sporadic breast cancer and its impact on tumor progression suggests in turn a more aggressive tumor behavior in radiation-associated breast cancers with downregulated TRPS1.

To clarify the functional consequences of TRPS1 downregulation in the radiation-associated context, we performed siRNA-knockdown experiments in radiation transformed breast cells B42-11 and B42-16. A time-course analysis of TRPS1 expression after siRNA-transfection (Supporting Information, Fig. S2) showed a downregulation of TRPS1 compared to the scrambled control at the mRNA and protein levels (Fig. 4). The major goal of this knockdown experiment was to establish a gene-correlation network in radiation-associated B42-11 and B42-16 cells based on global transcriptomic data for functional insights. A pathway enrichment analysis of differentially expressed genes revealed a significant enrichment of pathways related to DNA-repair, cell cycle, mitosis, cell migration, angiogenesis and EMT (Supporting Information, Table S5). This is in good agreement with the expectations from the published data as discussed above. However, a novel finding of this study is the effect of TRPS1 downregulation on DNA-repair pathways in radiation-

associated B42-11 and B42-16 cells pointing to radiation-induced effects in these cells. Furthermore, gene-expression-microarray data could be technically validated by qRT-PCR (Supporting Information, Table S4).

The gene interaction network of TRPS1 from global transcriptomic data of the TRPS1-knockdown in B42-11 and B42-16 cells was compared to a TRPS1-centered correlation network based on global mRNA expression data from matched sporadic breast cancers of the publicly available TCGA dataset (Figs. 5a and 5b and Supporting Information, Table S6). The main difference between both networks was a significant enrichment of apoptosis-related processes in sporadic tumors, while a link to DNA repair, chromosome segregation and genomic instability became apparent in the radiation transformed cell lines B42-11 and B42-16 (Supporting Information, Table S7). The involvement of TRPS1 in chromosome segregation has already been described in chondrocytes.³⁸ Many of the top ten genes interacting with TRPS1 are known to be involved in fundamental carcinogenic processes such as DNA repair and cell migration. For example, GPR64 and LYAR (TRPS1-interaction partners in B42-11 and B42-16 cells showing a positive correlation with TRPS1) are known to be involved in the process of migration. GPR64 is known to be involved in the adhesion and migration of breast cancer cells through mechanisms including a noncanonical NFkB pathway.³⁹ Furthermore, it was reported that transcription factor LYAR promote tumor cell migration and invasion by

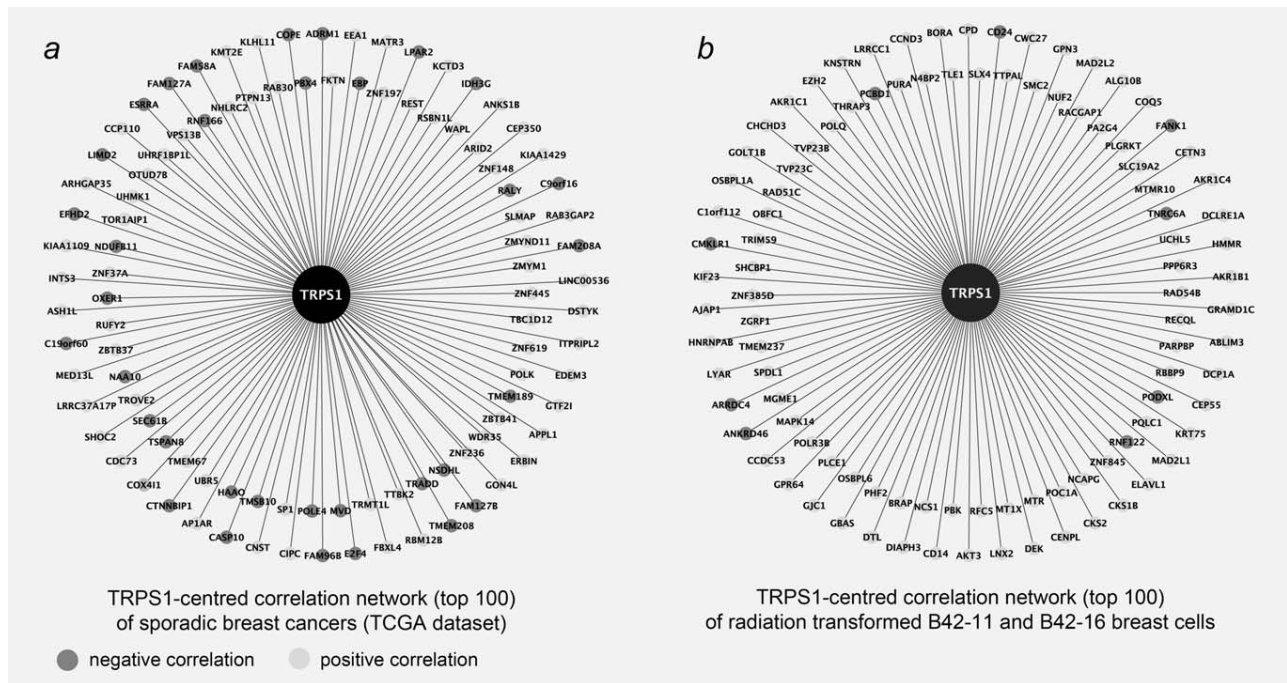


Figure 5. TRPS1-centered correlation networks consisting of the top 100 correlating genes with an FDR <0.05. The expression of genes labeled with dark grey circles showed negative correlation with TRPS1 expression and that of genes labeled with light grey circles showed positive correlation with TRPS1 expression. (a) TRPS1-centered correlation network based on global mRNA expression data from matched sporadic breast cancers of the publicly available TCGA dataset. (b) TRPS1-centered correlation network based on microarray gene expression data from B42-11 and B42-16 untransfected, scrambled-siRNA transfected and TRPS1-downregulated cells.

upregulating galectin-1 gene expression in colorectal cancer.⁴⁰ Another interesting network link was RFC5 (activated by TRPS1 in B42-11 and B42-16) as it appeared in many (10 out of 14) pathways related to DNA repair and cell cycle/mitosis from the differential expression analysis and is among the top five genes correlating with TRPS1. The RFC5 gene belongs to the replication factor C family and was described to reflect the hallmark of cancer “genomic instability.”⁴¹ It was already reported that RFC5 recognize DNA damage and is involved in pathways related to the process of mismatch repair.^{42,43} Furthermore, an aberrant expression of this gene was already observed in several tumor entities.^{42,44,45}

This suggests deregulation of cellular processes involved in radiation-induced damage response. In all, there are several hints that TRPS1 plays a specific role in DNA repair, chromosome segregation and genomic instability which is a well-established phenotype after irradiation and in radiation-associated carcinogenesis.⁴⁶ A link of TRPS1-interaction partners to DNA repair and chromosome segregation is not obvious from the TRPS1-centered correlation network derived from the sporadic breast cancer TCGA dataset suggesting this being a specific effect of TRPS1 deregulation in radiation-associated breast cancer. Moreover, most of the top ten TRPS1-interaction partners derived from the sporadic dataset are known to be involved in apoptosis, cell migration and cell cycle which is in agreement with the published literature on TRPS1 in sporadic breast cancer and prostate cancer.^{47–49}

It was already shown in MCF7 breast cancer cells that TRPS1 functions as a transcription activator of FOXA1 and negatively regulates the expression of ZEB2.^{11,13} An interaction of FOXA1 with TRPS1 was also detected in the correlation network of the sporadic TCGA breast cancer dataset (FOXA1, Pearson correlation = 0.17, FDR = 0.02). The weak but significant correlation could be due to the fact that the TRPS1-interaction network for sporadic breast cancer in this study was developed from mRNA expressions of tumor tissues in contrast to proteomics data from *in vitro* models as published by Huang *et al.*¹¹ The negative association of TRPS1 with ZEB2, however, was not detected in our data. It is interesting to note that there is no common gene between the correlation networks of B42-11 and B42-16 cells and of the sporadic TCGA dataset which again points to specific radiation-associated functions of TRPS1.

In conclusion, this study reveals radiation markers in breast carcinogenesis consisting of an upregulated hsa-miR-26b-5p and a downregulation of the validated target protein TRPS1. Both markers could be validated in independent tumor cohorts of radiation-associated post-Chernobyl breast cancers, suggesting an important role in radiation-induced carcinogenesis. Moreover, we could identify interaction partners of TRPS1 in TRPS1-knockdown models that point to a functional role of TRPS1 in radiation-associated breast carcinogenesis in DNA damage response and tumor progression.

Acknowledgements

The authors thank U. Buchholz, C. Innerlohinger, E. Konhäuser, L. Dajka, C.M. Pflüger, I. Zagorski and A. Selmeier for excellent technical support.

The results shown here are in whole or part based on data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>.

References

- Ronckers CM, Erdmann CA, Land CE. Radiation and breast cancer: a review of current evidence. *Breast Cancer Res* 2005;7:21–32.
- Prisyazhnyuk A, Gristchenko V, Fedorenko Z, et al. Twenty years after the Chernobyl accident: solid cancer incidence in various groups of the Ukrainian population. *Radiat Environ Biophys* 2007;46:43–51.
- Prisyazhnyuk AY, Bazyka DA, Romanenko AY, et al. Quarter of century since the Chernobyl accident: small es, Cyrillicancer risks in affected groups of population. *Probl Radiac Med Radiobiol* 2014;19:147–69.
- Niemoeller OM, Niyazi M, Corradini S, et al. MicroRNA expression profiles in human cancer cells after ionizing radiation. *Radiat Oncol* 2011;6:29.
- Acunzo M, Romano G, Wernicke D, et al. MicroRNA and cancer—a brief overview. *Adv Biol Regul* 2015;57:1–9.
- Blenkiron C, Goldstein LD, Thorne NP, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol* 2007;8:R214.
- Bach AS, Derocq D, Laurent-Matha V, et al. Nuclear cathepsin D enhances TRPS1 transcriptional repressor function to regulate cell cycle progression and transformation in human breast cancer cells. *Oncotarget* 2015;6:28084–103.
- Chang GT, van den Bemd GJ, Jhamai M, et al. Structure and function of GC79/TRPS1, a novel androgen-repressible apoptosis gene. *Apoptosis* 2002;7:13–21.
- Hong J, Sun J, Huang T. Increased expression of TRPS1 affects tumor progression and correlates with patients' prognosis of colon cancer. *Biomed Res Int*. 2013;2013:454085.
- Hu J, Su P, Jia M, et al. TRPS1 expression promotes angiogenesis and affects VEGFA expression in breast cancer. *Exp Biol Med (Maywood)* 2014;239:423–9.
- Huang JZ, Chen M, Zeng M, et al. Down-regulation of TRPS1 stimulates epithelial-mesenchymal transition and metastasis through repression of FOXA1. *J Pathol* 2016;239:186–96.
- Wu L, Wang Y, Liu Y, et al. A central role for TRPS1 in the control of cell cycle and cancer development. *Oncotarget* 2014;5:7677–90.
- Stinson S, Lackner MR, Adai AT, et al. TRPS1 targeting by miR-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. *Sci Signal* 2011;4:ra41.
- Kryuchkov V, Chumak V, Maceika E, et al. Radrue method for reconstruction of external photon doses for Chernobyl liquidators in epidemiological studies. *Health Phys* 2009;97:275–98.
- Van Wieringen WN, VdWM, Van der Vaart AW. A test for partial differential expression. *J Am Stat Assoc* 2008;103:1039–49.
- Unger K, Wienberg J, Riches A, et al. Novel gene rearrangements in transformed breast cells identified by high-resolution breakpoint analysis of chromosomal aberrations. *Endocr Relat Cancer* 2010;17:87–98.
- Zitzelsberger H, Lehmann L, Hieber L, et al. Cytogenetic changes in radiation-induced tumors of the thyroid. *Cancer Res* 1999;59:135–40.
- Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5:R80.
- Gentleman RCV, Dudoit S, Izirary RA, et al. Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer, 2005.
- Jupe S, Akkerman JW, Soranzo N, et al. Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets. *J Thromb Haemost* 2012;10:2399–402.
- Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–504.
- Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2:401–4.
- Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6:pl1.
- BYaH Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JR Stat* 1995;57:289–300.
- Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009;25:1091–3.
- Chaudhry MA. Radiation-induced microRNA: discovery, functional analysis, and cancer radiotherapy. *J Cell Biochem* 2014;115:436–49.
- van Schooneveld E, Wildiers H, Vergote I, et al. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. *Breast Cancer Res* 2015;17:21.
- Gandellini P, Rancati T, Valdagni R, et al. miRNAs in tumor radiation response: bystanders or participants? *Trends Mol Med* 2014;20:529–39.
- Liu XX, Li XJ, Zhang B, et al. MicroRNA-26b is underexpressed in human breast cancer and induces cell apoptosis by targeting SLC7A11. *FEBS Lett* 2011;585:1363–7.
- Hsu SD, Tseng YT, Shrestha S, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 2014;42:D78–85.
- Selmansberger M, Feuchtinger A, Zurnadzy L, et al. CLIP2 as radiation biomarker in papillary thyroid carcinoma. *Oncogene* 2015;34:3917–25.
- Barcellos-Hoff MH, Park C, Wright EG. Radiation and the microenvironment - tumorigenesis and therapy. *Nat Rev Cancer* 2005;5:867–75.
- Li J, Kong X, Zhang J, et al. MiRNA-26b inhibits proliferation by targeting PTGS2 in breast cancer. *Cancer Cell Int* 2013;13:7.
- Radvanyi L, Singh-Sandhu D, Gallichan S, et al. The gene associated with trichorhinophalangeal syndrome in humans is overexpressed in breast cancer. *Proc Natl Acad Sci USA* 2005;102:11005–10.
- Chen JQ, Bao Y, Lee J, et al. Prognostic value of the trichorhinophalangeal syndrome-1 (TRPS-1), a GATA family transcription factor, in early-stage breast cancer. *Ann Oncol* 2013;24:2534–42.
- Su P, Hu J, Zhang H, et al. Association of TRPS1 gene with different EMT markers in ERalpha-positive and ERalpha-negative breast cancer. *Diagn Pathol* 2014;9:119.
- Chen JQ, Bao Y, Litton J, et al. Expression and relevance of TRPS-1: a new GATA transcription factor in breast cancer. *Horm Cancer* 2011;2:132–43.
- Wuelling M, Pasdziernik M, Moll CN, et al. The multi zinc-finger protein Trps1 acts as a regulator of histone deacetylation during mitosis. *Cell Cycle* 2013;12:2219–32.
- Peeters MC, Fokkelman M, Boogaard B, et al. The adhesion G protein-coupled receptor G2 (ADGRG2/GPR64) constitutively activates SRE and NFkappaB and is involved in cell adhesion and migration. *Cell Signal* 2015;27:2579–88.
- Wu Y, Liu M, Li Z, et al. LYAR promotes colorectal cancer cell motility by activating galectin-1 expression. *Oncotarget* 2015;6:32890–901.
- Cook AC, Tuck AB, McCarthy S, et al. Osteopontin induces multiple changes in gene expression that reflect the six “hallmarks of cancer” in a model of breast cancer progression. *Mol Carcinog* 2005;43:225–36.
- Qian L, Luo Q, Zhao X, et al. Pathways enrichment analysis for differentially expressed genes in squamous lung cancer. *Pathol Oncol Res* 2014;20:197–202.
- Niida H, Nakanishi M. DNA damage checkpoints in mammals. *Mutagenesis* 2006;21:3–9.
- Badura M, Braunstein S, Zavadil J, et al. DNA damage and eIF4G1 in breast cancer cells reprogram translation for survival and DNA repair mRNAs. *Proc Natl Acad Sci USA* 2012;109:18767–72.
- Mazumdar T, DeVecchio J, Agyeman A, et al. The GLI genes as the molecular switch in disrupting Hedgehog signaling in colon cancer. *Oncotarget* 2011;2:638–45.
- Huang L, Snyder AR, Morgan WF. Radiation-induced genomic instability and its implications for radiation carcinogenesis. *Oncogene* 2003;22:5848–54.
- Wang H, Rana S, Giese N, et al. Tspan8, CD44v6 and alpha6beta4 are biomarkers of migrating pancreatic cancer-initiating cells. *Int J Cancer* 2013;133:416–26.
- Sarveswaran S, Ghosh J. OXER1, a G protein-coupled oxoecosatetraenoid receptor, mediates the survival-promoting effects of arachidonate 5-lipoxygenase in prostate cancer cells. *Cancer Lett* 2013;336:185–95.
- Angus SP, Nevins JR. A role for Mediator complex subunit MED13L in Rb/E2F-induced growth arrest. *Oncogene* 2012;31:4709–17.

RESEARCH ARTICLE

Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data

Agata Michna¹, Herbert Braselmann^{1,2}, Martin Selmsberger¹, Anne Dietz³, Julia Hess^{1,2}, Maria Gomolka³, Sabine Hornhardt³, Nils Blüthgen⁴, Horst Zitzelsberger^{1,2}, Kristian Unger^{1,2*}

1 Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany, **2** Clinical Cooperation Group "Personalized Radiotherapy in Head and Neck Cancer", Helmholtz Zentrum München, Neuherberg, Germany, **3** Department of Radiation Protection and Health, Federal Office for Radiation Protection, Neuherberg, Germany, **4** Institute of Pathology, Charité—Universitätsmedizin Berlin, Berlin, Germany

* unger@helmholtz-muenchen.de



CrossMark
click for updates

OPEN ACCESS

Citation: Michna A, Braselmann H, Selmsberger M, Dietz A, Hess J, Gomolka M, et al. (2016) Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data. PLoS ONE 11(8): e0160791. doi:10.1371/journal.pone.0160791

Editor: Geraldo A Passos, University of São Paulo, BRAZIL

Received: April 26, 2016

Accepted: June 14, 2016

Published: August 9, 2016

Copyright: © 2016 Michna et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Expression microarray data files are available from the ArrayExpress database (accession number E-MTAB-4829).

Funding: ZISS project, 02NUK024B, <https://www.bmbf.de/en/index.html>, Federal Ministry of Education and Research, KU JH HZ. 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.

Abstract

Gene expression time-course experiments allow to study the dynamics of transcriptomic changes in cells exposed to different stimuli. However, most approaches for the reconstruction of gene association networks (GANs) do not propose prior-selection approaches tailored to time-course transcriptome data. Here, we present a workflow for the identification of GANs from time-course data using prior selection of genes differentially expressed over time identified by natural cubic spline regression modeling (NCSRМ). The workflow comprises three major steps: 1) the identification of differentially expressed genes from time-course expression data by employing NCSRМ, 2) the use of regularized dynamic partial correlation as implemented in GeneNet to infer GANs from differentially expressed genes and 3) the identification and functional characterization of the key nodes in the reconstructed networks. The approach was applied on a time-resolved transcriptome data set of radiation-perturbed cell culture models of non-tumor cells with normal and increased radiation sensitivity. NCSRМ detected significantly more genes than another commonly used method for time-course transcriptome analysis (BETR). While most genes detected with BETR were also detected with NCSRМ the false-detection rate of NCSRМ was low (3%). The GANs reconstructed from genes detected with NCSRМ showed a better overlap with the interactome network Reactome compared to GANs derived from BETR detected genes. After exposure to 1 Gy the normal sensitive cells showed only sparse response compared to cells with increased sensitivity, which exhibited a strong response mainly of genes related to the senescence pathway. After exposure to 10 Gy the response of the normal sensitive

cells was mainly associated with senescence and that of cells with increased sensitivity with apoptosis. We discuss these results in a clinical context and underline the impact of senescence-associated pathways in acute radiation response of normal cells. The workflow of this novel approach is implemented in the open-source Bioconductor R-package `splineTimeR`.

Introduction

In general terms, the expression of genes can be studied from a static or temporal point of view. Static microarray experiments allow measuring gene expression responses only at one single time point. Therefore, data obtained from those experiments can be considered as more or less randomly taken snapshots of the molecular phenotype of a cell. However, biological processes are dynamic and thus, the expression of a gene is a function of time [1]. To be able to understand and model the dynamic behavior and association of genes, it is important to study gene expression patterns over time.

However, compared to static microarray data, the analysis of time-course data introduces a number of new challenges. First, the experimental costs for the generation of data as well as the computational cost increases with the increase in the number of introduced time points. Second, hidden correlation caused by co-expression of genes makes the data linearly dependent [2]. Finally, one has to be aware of additional correlations existing between neighboring time points clearly revealed in published gene expression profiles [3].

Several different algorithms have been suggested to analyze gene time-course microarray data with regard to differential expression in two or more biological groups (e.g. exposed to radiation vs. non-exposed) [4–7]. Nevertheless solitary identification of differentially expressed genes does not help to determine the molecular mechanisms in the investigated biological groups. Therefore, it is not only important to know differentially expressed genes per se, but also how those genes interact and regulate each other in order to determine specifically deregulated molecular networks.

Currently, many different algorithms including cluster analysis [8–13] and supervised classification [14–16] are used to identify relationships between genes. However, both of these methods suffer from serious limitations. First, the timing information of the measurements is not incorporated and, therefore, the intrinsic temporal structure of the time-course data is neglected. Second, the available standard clustering and classification methods are not designed to measure statistical significance of the results based on a statistical hypothesis test. By nature of these methods, clusters or classes of genes with similar expression patterns will always be identified but they do not provide a measure of how reliable this information is. For this reason, we preferred usage of a dynamic network modeling approach that allows delineation of relationships between genes along with providing statistical significance for these relationships.

The aim of the present study was to identify and compare signaling pathways involved in the radiation responses of normal cells differing in their radiation sensitivity that could be used to modulate cell sensitivity to ionizing radiation. For this, we propose an approach that combines the detection of genes differentially expressed over time based on statistics determined by natural cubic spline regression (NCSRM) [17] followed by dynamic gene association network (GAN) reconstruction based on a regularized dynamic partial correlation as implemented in the GeneNet R-package [18].

Most exploratory gene expression studies focus only on the identification of differentially expressed genes by treating them as independent events and do not seek to study the interplay of identified genes. This makes it difficult to tell which genes are part of the interaction network

causal of the studied phenotype and which are the most “important” with regard to the context of the investigation. The herein present approach combines the identification of differentially expressed genes and reconstruction of possible associations between them. Further analysis of identified GANs then allows hypothesizing which genes may play a crucial role in the investigated processes. This should markedly increase the likelihood to find meaningful results from an initial observation and help to understand the underlying molecular mechanisms. We applied our workflow on time-course transcriptome data of two normal and well-characterized lymphoblastoid cell lines with normal (20037–200) and increased radiation sensitivity (4060–200), in order to identify molecular mechanisms and potential key players responsible for different radiation responses [19, 20]. Our exploratory approach provides novel and informative insights in the biology of radiation sensitivity of non-tumor cells after exposure to ionizing radiation with regard to the identified signaling pathways and their key drivers. Moreover, we could demonstrate that spline regression in differential gene expression analysis for the purpose of prior selection in gene-association network reconstruction outperforms another commonly used existing approach for time-course gene expression analysis.

Results

The schematic workflow of the presented novel approach for time-course gene expression data analysis is presented in [Fig 1](#).

Identification of ionizing radiation-responsive genes using NCSRМ method

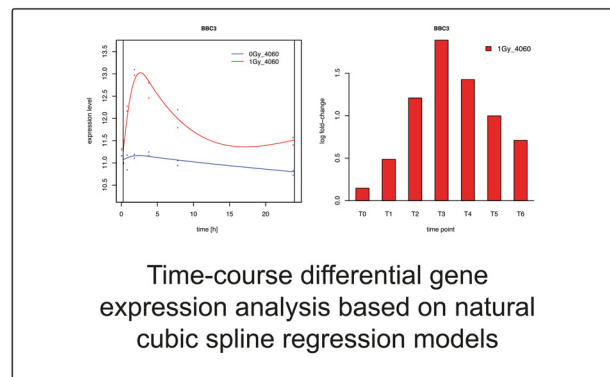
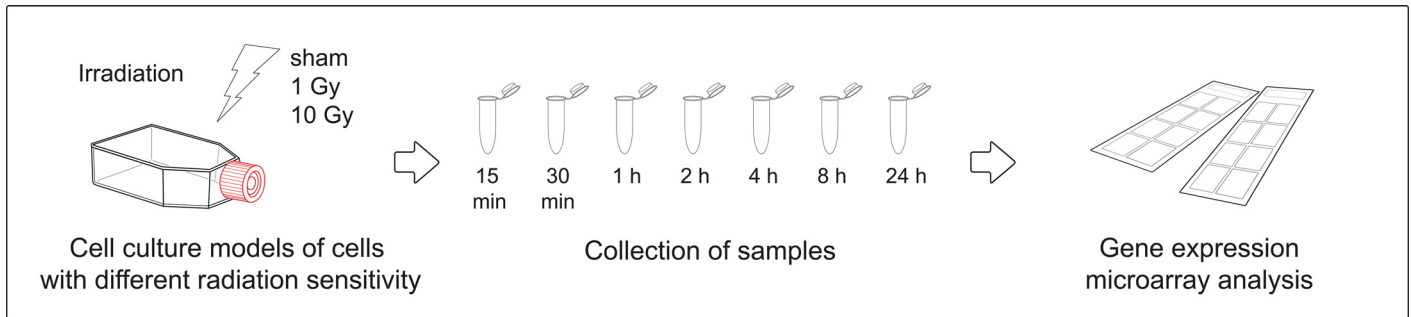
A fraction of the probes was removed due to low expression levels, with not detectable signal intensities as described in [21]. [Table 1](#) shows the number of probes remained after quality filtering from the total number of 25220 unique probes representing HGNC annotated genes. Differential analysis was performed relative to the corresponding sham irradiated cells as a reference. In general, more genes were detected as differentially expressed in the cells with increased radiation sensitivity compared to cells with normal radiation sensitivity after each dose of gamma irradiation ([Table 1](#)). The most prominent difference was observed when comparing the responses after 1 Gy irradiation. In the cells with increased radiation sensitivity 2335 genes showed differential expression compared to only seven genes in cells with normal radiation sensitivity. We observed the same trend after irradiation with 10 Gy where the cells with increased sensitivity showed 6019 and the normal sensitive cells 3892 differentially expressed genes.

Pathway enrichment analysis of NCSRМ identified genes

Pathway enrichment analysis was performed on differentially expressed genes to identify over-represented biological pathways. The analysis on genes identified with NCSRМ revealed 634 and 964 significantly enriched pathways for the cells with increased radiation sensitivity after 1 Gy and 10 Gy irradiation dose, respectively and 758 pathways for the normal sensitive cell line after 10 Gy irradiation. For the seven differentially expressed genes (i.e. FDXR, BBC3, VWCE, PHLDA3, SCARF2, HIST1H4C, PCNA) of the cell line with normal radiation sensitivity after 1 Gy dose of irradiation we did not find any significantly enriched pathways. A summary of the pathway enrichment results can be found in [S2 Table](#).

Gene association network reconstruction

None of the edge probabilities calculated for the seven differentially expressed genes in the cell line with normal radiation sensitivity after 1 Gy irradiation exceeded the considered



Pathway enrichment analysis of differentially expressed genes over time (Reactome)

Plausibility of differential gene expression results

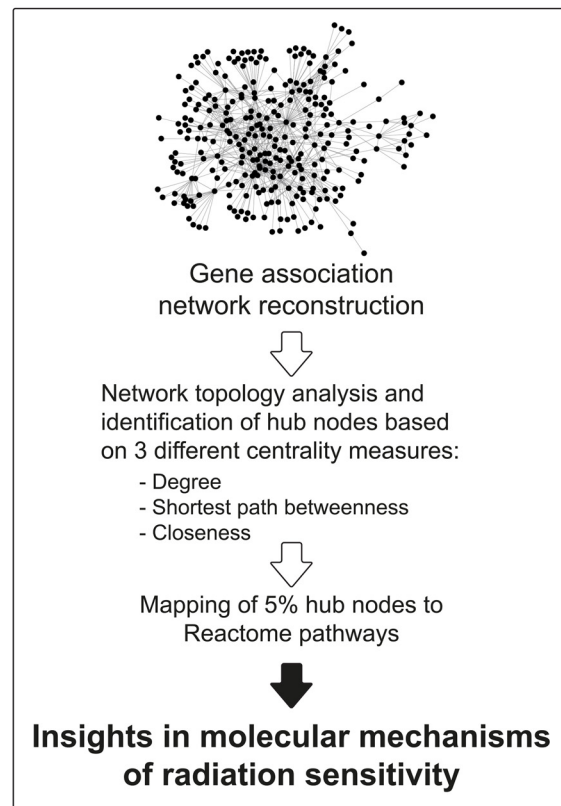


Fig 1. Schematic workflow of the analysis of gene expression time-course data. Samples were collected 0.25, 0.5, 1, 2, 4, 8 and 24 hours after sham or actual irradiation. Transcriptional profiling was performed using Agilent gene expression microarrays and comprises three major steps: the identification of differentially expressed genes from time-course expression data by employing a natural cubic spline regression model; the use of regularized dynamic partial correlation method to infer gene associations networks from differentially expressed genes and the topological identification and functional characterization of the key nodes in the reconstructed networks.

doi:10.1371/journal.pone.0160791.g001

significance threshold and hence no network was obtained. For the remaining conditions we were able to obtain association networks as presented in [Table 2](#). Obtained networks are provided as igraph R-objects in the supplementary data ([S1 File](#)). The graph densities for all resulting networks were in the same range as the density of the Reactome interaction network ([Table 2](#)).

Identification and functional characterization of the most important genes in the reconstructed association networks

The combined topological centrality measure was used to characterize the biological importance of nodes (genes) in the reconstructed association networks. The 5% of the highest ranked genes listed in supplementary [S3 Table](#) were mapped to Reactome pathways in order to further evaluate their biological roles. The top 10 most relevant pathways according to the FDR values are shown in [Table 3](#). For the cell line with increased radiation sensitivity after irradiation with 1 Gy and for the normal sensitive cell line after 10 Gy the induction of pathways associated with senescence response was detected. For the cell line with increased radiation sensitivity after 10 Gy of irradiation we mostly observed pathways associated with apoptosis. All pathways are listed in supplementary [S4 Table](#).

False detected differentially expressed genes between technical replicates

In order to assess the false positive rate, the spline regression based differential analyses between technical replicates of each treatment conditions and cell lines were performed. Here, we can state that the null-hypothesis of no differential expression is true for all genes. Then the q^* -level of 0.05 for Benjamini-Hochberg method controls also the FWER at alpha-level equal to 0.05 (type I error) [22]. For all compared technical replicates not more than 3% rejections of null hypothesis were detected, which is in good accordance to the expected or nominal type I error.

Evaluation of spline regression model in comparison to BETR method

[Table 1](#) compares the numbers of differentially expressed genes obtained from both methods applied on the same gene expression data set and FDR thresholds. For almost all treatment conditions the BETR method detected less differentially expressed genes in comparison to

Table 1. Number of detected and differentially expressed genes for each dose and cell lines for NCSRm and BETR methods.

cell line and applied radiation dose	increased sensitivity (1 Gy vs 0 Gy)	Normal sensitivity (1 Gy vs 0 Gy)	increased sensitivity (10 Gy vs 0 Gy)	Normal sensitivity (10 Gy vs 0 Gy)
total number of detected probes after preprocessing	10388	11311	10330	11446
differentially expressed genes detected with NCSRm	2335	7	6019	3892
differentially expressed genes detected with BETR	923	12	3889	1256
intersection of differentially expressed genes resulting from both methods	855	4	3875	1233

doi:10.1371/journal.pone.0160791.t001

Table 2. Number of genes subjected to GAN reconstruction and properties of resulted GANs.

method	NCSRМ				BETR			
	Increased sensitivity (1 Gy)	normal sensitivity (1 Gy)	Increased sensitivity (10 Gy)	normal sensitivity (10 Gy)	Increased sensitivity (1 Gy)	normal sensitivity (1 Gy)	Increased sensitivity (10 Gy)	normal sensitivity (10 Gy)
number of genes taken for network reconstruction	2335	7	6019	3892	923	12	3889	1256
number of nodes remained in the network	1140	-	3483	2735	336	-	2299	773
number of edges in the network	12198	-	114629	84695	3268	-	126378	16862
network density	0.00939	-	0.00945	0.01133	0.02903	-	0.02392	0.02826
density of the Reactome interaction network	0.00536							

Gene association network reconstructions were performed using the GeneNet method [18]. Association between two genes was considered as significant if posterior edge probability was equal or greater than 0.95. Densities of the reconstructed networks were compared with the density of the Reactome interaction network in order to assess their complexity.

doi:10.1371/journal.pone.0160791.t002

NCSRМ. Only for the normal cell line after irradiation with 1 Gy BETR identified 12 genes whereas NCSRМ identified only 7 genes. As a consequence of the lower numbers of detected differentially expressed genes with BETR, the obtained networks are smaller than those obtained after spline regression. The detailed comparison results including numbers of detected differentially expressed genes and the sizes of reconstructed association networks are presented in the Table 2. The lists of differentially expressed genes obtained with the two methods are shown in supplementary S1 Table. The top 10 pathways to which the 5% of the most important genes in the reconstructed association networks were mapped to are shown in Table 3. With NCSRМ we were not only able to detect almost all genes that were detected also by BETR (Table 1), but also an additional set of genes resulting in almost twice the number of genes compared to BETR. Nevertheless, the top 5% hub genes of the networks derived from the differentially expressed genes defined by BETR were associated with similar biological processes as those from the spline differential expression analysis derived networks. The numbers and names of overlapping hub genes in the GANs are presented in Table 4 and in supplementary S3 Table, respectively.

Evaluation of reconstructed networks

The evaluation of the two networks derived after 1 Gy irradiation of the cell line with increased sensitivity showed that the network reconstructed with the differentially expressed genes determined using BETR did not contain significantly more common edges than random networks ($p = 0.529$), whereas the network reconstructed with the differentially expressed genes determined by NCSRМ did ($p = 0.048$). The networks derived after 10 Gy irradiation of the cell line with increased sensitivity and 10 Gy irradiation of the normal sensitive cell line contained significantly more edges that were common with the Reactome network compared to random networks for both methods.

Discussion

The success of tumor radiation therapy predominantly depends on the total applied radiation dose, but also on the tolerance of the tumor surrounding normal tissues to radiation. Toxicity

Table 3. Comparison of NCSR and BETR methods with respect to the top 10 pathways after mapping of 5% highest ranked genes from the reconstructed gene association networks.

with NCSR method			with BETR method		
increased sensitivity (1 Gy)	increased sensitivity (10 Gy)	normal sensitivity (10 Gy)	increased sensitivity (1 Gy)	increased sensitivity (10 Gy)	normal sensitivity (10 Gy)
Signal Transduction	Signal Transduction	Generic Transcription Pathway	DNA Damage/Telomere Stress Induced Senescence ^a	Activation of BH3-only proteins ^b	DNA Damage/Telomere Stress Induced Senescence ^a
Cellular Senescence ^a	Activation of BH3-only proteins ^b	DNA Damage/Telomere Stress Induced Senescence ^a	Senescence-Associated Secretory Phenotype (SASP) ^a	Activation of PUMA and translocation to mitochondria ^b	Generic Transcription Pathway
DNA Damage/Telomere Stress Induced Senescence ^a	Activation of PUMA and translocation to mitochondria ^b	Immune System	Signal Transduction	Cytokine Signaling in Immune system	Cellular Senescence ^a
Formation of Senescence-Associated Heterochromatin Foci (SAHF) ^a	Fatty acid, triacylglycerol, and ketone body metabolism	Gene Expression	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	Immune System	Gene Expression
Cellular responses to stress	Metabolism	Inositol phosphate metabolism	Cell Cycle Checkpoints	Intrinsic Pathway for Apoptosis ^b	Meiotic recombination
RAF-independent MAPK1/3 activation	Metabolism of proteins	IRF3-mediated induction of type I IFN	Cellular Senescence ^a	Signal Transduction	Signal Transduction
Signaling by ERBB4	PPARA activates gene expression	Cellular Senescence ^a	DNA methylation	Gene Expression	Cell Cycle
DAP12 interactions	Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	Formation of Senescence-Associated Heterochromatin Foci (SAHF) ^a	Packaging Of Telomere Ends	BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members ^b	Transcriptional activation of cell cycle inhibitor p21
PRC2 methylates histones and DNA	Activation of gene expression by SREBF (SREBP)	STING mediated induction of host immune responses	RNA Polymerase I Promoter Opening	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	Transcriptional activation of p53 responsive genes
Apoptotic execution phase ^b	BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members ^b	Metabolism	SIRT1 negatively regulates rRNA Expression	Endosomal/Vacuolar pathway	Senescence-Associated Secretory Phenotype (SASP) ^a

^aPathways associated with senescence responses.

^bPathways associated with apoptotic processes.

doi:10.1371/journal.pone.0160791.t003

towards radiation, which greatly varies on an individual level due to inherited susceptibility, is one of the most important limiting factors for dose escalation in radiooncology treatment [23, 24]. To account for radiation sensitivity of normal tissue in personalized treatment approaches the underlying molecular mechanisms need to be thoroughly understood in order to identify

Table 4. Comparison of hub genes in networks resulting from different methods.

cell line and applied radiation dose	increased sensitivity (1 Gy)	increased sensitivity (10 Gy)	Normal sensitivity (10 Gy)
5% hub genes in the NCSR resulting network in numbers	57	174	137
5% hub genes in the BETR resulting network in numbers	17	115	39
number of common hub genes resulting from both methods	9	111	31

doi:10.1371/journal.pone.0160791.t004

molecular targets for the modulation of radiation sensitivity and molecular markers for the stratification of patients with different intrinsic radiation sensitivity. In the present study we identified significantly differentially expressed genes over time between the radiation-treated group and the control group to be used as prior genes for GAN reconstruction. Two doses of gamma irradiation were used to characterize the differences in radiation response of the two lymphoblastoid cell lines with known differences in radiation sensitivity. The dose of 10 Gy was selected following the fact that the same dose has been applied in a previous research project examining the radiation sensitivity of the same lymphoblastoid cell lines analyzed in the study at hand [20]. The dose of 1 Gy reflects the dose that is delivered as part of the so called “low-dose bath” to the tumor-surrounding tissue during the radiotherapy of the tumors [25].

Here, we conducted time-resolved transcriptome analysis of radiation-perturbed cell culture models of non-tumor cells with normal and with increased radiation sensitivity in order to work out the molecular phenotype of radiation sensitivity in normal cells. Moreover, we present an innovative approach for the identification of GANs from time-course perturbation transcriptome data. The approach comprises three major steps: 1) the identification of differentially expressed genes from time-course gene expression data by employing a natural cubic spline regression model (NCSRM); 2) the use of a regularized dynamic partial correlation method to infer gene associations network from differentially expressed genes; 3) the identification and functional characterization of the key nodes (hubs) in the reconstructed gene dependency network (Fig 1).

Our proposed method for the detection of differentially expressed genes over time is based on NCSRM with a small number of basis functions. A relatively low number of basis functions generally results in a good fit of data and, at the same time, reduces the complexity of the fitted models. Treating time in the model as a continuous variable, a non-linear behavior of gene expressions was approximated by spline curves fitted to the experimental time-course data. Considering temporal changes in gene expression as continuous curves and not as single time points greatly decreases the dimensionality of the data and thereby decreases computational cost. In addition, the proposed NCSRM does not require identical sampling time points for the compared treatment conditions. Furthermore, no biological replicates are needed. Therefore, the method is applicable to data generated according to a tailored time-course differential expression study design and to data that were not specifically generated for time-course differential expression analysis, e.g. existing/previously generated data from clinical samples. Thus, the adaption of the method to differential expression analysis comprises the potential to reanalyze existing data, address new questions *in silico* and thereby potentially add new or additional value to existing data. Incomplete time-course data, e.g. due to the exclusion of samples for technical reasons, that often create major problems for the estimation of the model, are also suitable for fitting the spline regression model as long as enough data points remain in the data set. This is especially valuable when data on certain time points, derived from a very limited sample source, have been excluded from a time-course data set and cannot be repeatedly generated.

Since gene expression is not only dynamic in the treatment group but also in the control group, the inclusion of the time-course control data greatly improves the ability to detect truly differentially expressed genes, as the gene expression values are not referred to a single time point with static gene expression levels only. Comparing a treatment group to time point zero does not provide a proper control over the entire time-course, although it is widely practiced [26–28]. The proposed workflow is implemented in an open-source R-package `splineTimeR` and is available through Bioconductor (<https://www.bioconductor.org>).

Amongst a panel, the two lymphoblastoid cell lines that were different with regard to radiation sensitivity after irradiation with 10 Gy [20], also responded differently with regard to the

quantity of differentially expressed genes. Interestingly, cells with normal radiation sensitivity barely responded to 1 Gy irradiation at the transcriptome level. Only seven genes (FDXR, BBC3, VWCE, PHLDA3, SCARF2, HIST1H4C, PCNA) were identified as differentially expressed, whereas for the cell line with increased sensitivity 2335 differentially expressed genes were detected after exposure to the same dose. A similar behavior was observed for those two cell lines after irradiation with 10 Gy. We detected 6019 and 3892 genes as differentially expressed in the sensitive and normal cell lines, respectively (Table 2). Those results are in a good agreement with the previous proteomic study where more differentially expressed proteins were detected for the same sensitive cell line compare to the cell line with normal radiation sensitivity 24 hours after irradiation with 10 Gy [29]. Thus, for both applied doses, the radiation sensitive cells exhibited much more pronounced transcriptional response compared to the cells with normal radiation sensitivity and thereby underlines the expected radiation response of those two cell lines.

Concerning qualitative differences in the transcriptomic response of normal sensitive cells and cells with increased sensitivity after treatment with 1 Gy and 10 Gy pathway enrichment analysis was performed. Differentially expressed genes identified for all considered treatment conditions except for the normal sensitive cells after exposure to 1 Gy radiation showed statistically significant enrichment of pathways. Most of which were in agreement with known radiation responses such as DNA repair, cell cycle regulation, oxidative stress response or pathways related to apoptosis (S2 Table) [30–32]. Therefore, the pathway enrichment analysis results suggest plausibility of generated data and, more importantly, underline the meaningfulness of our suggested approach based on cubic spline regression for differential gene expression analysis of time-course data. However, differential expression analysis alone followed by pathway enrichment analysis does not provide any mechanistic insights. For this reason we performed GAN reconstruction using identified differentially expressed genes. Based on the assumption that the expression levels of functionally related genes are highly correlated, partial correlation was used for GAN reconstruction. In simple correlation, the strength of the linear relationship between two genes is measured, without taking into account that those genes may be actually influenced by other genes. Partial correlation eliminates the influence of other genes when one particular relationship between pair of genes is considered. Network reconstruction was performed separately for the cell line with increased radiation sensitivity after 1 Gy and 10 Gy and for the cell line with normal radiation sensitivity after 10 Gy of radiation dose. Due to the sparseness of the set of genes differentially expressed after irradiation of the normal-sensitive cell line with 1 Gy, no GAN was obtained.

Subsequently, we identified the network hubs (i.e. most important genes) of the GANs by combining three network centrality measures: degree, closeness and shortest path betweenness [33]. Combining different centrality measures is a widely used approach to identify nodes that are likely to control the network [34]. Also, this approach allows identification of nodes that are connected to the central nodes at the same time which can be informative for the interpretation of the whole GAN or single modules making up the network [33, 34].

Identification of key pathways associated with radiation sensitivity

In order to get functional insights into the reconstructed GANs the 5% top important nodes were identified after a ranking with the combined centrality measure and mapped to the pathways from the interactome database Reactome [35]. The obtained results revealed different pathways considered as the most important in cells with different radiation sensitivity after different doses of ionizing radiation. For the radiation sensitive cell line 4060–200 and 1 Gy irradiation, we mainly detected pathways associated with senescence (Table 3).

A different outcome was observed after irradiation with 10 Gy. For the radiation sensitive cells three out of the ten top pathways were linked to apoptotic processes with the genes BBC3, BCL2, TP53 as key players, whereas for the normal sensitive cell line we mainly observed the induction of senescence related pathways. This indicates that different doses are necessary to induce a similar response in the two cell lines. The activation of senescence genes is a damage response mechanism, which stably arrests proliferating cells and protects them from apoptotic cell death [36]. Together with the senescence pathway we observed increased levels of chemokine, cytokine and interleukin genes that are known to activate an immune response and signal transduction pathways in response to irradiation.

Although the senescence-associated pathways were not seen as the most important ones for the treatment condition 10 Gy/increased sensitivity, they were significantly enriched in the GANs of the three conditions 1 Gy/increased sensitivity, 10 Gy/ increased sensitivity and 10 Gy/normal sensitivity. All differentially expressed genes that related to senescence-associated pathways are shown in supplementary S5 Table. The observation that cells with increased radiation sensitivity compared to cells with normal sensitivity, become senescent after exposure to doses in the range of 1 Gy, rises the question whether this has a positive or negative influence on the tumor therapy. On the one hand side, senescent cell may secrete the so-called SASP (“senescence-associated secretory phenotype”) factors, including growth factors, chemokines and cytokines, which participate in intercellular signaling leading to the attraction of immune cells to the tumor location that, in turn, eliminate the tumor cells and, thereby, positively contribute to the tumor therapy [37, 38]. On the other hand side, senescent cells and the SASP are reported to promote proliferation, survival, invasion and migration of neighboring cells by the release of pro-inflammatory cytokines leading to sustained inflammation [36]. In this way senescence cells can damage their local environment and stimulate angiogenesis and tumor progression [39, 40]. Besides, there are some evidences that the induction of senescence in surrounding normal tissue may lead to an increased radio-tolerance or even radioresistance of the tumor and is, therefore, not desirable and negatively influences the tumor radiotherapy [41]. Thus, it might be beneficial to block senescence in order to prevent the radio-hyposensibilization of tumor cells. Therefore, we suggest a detailed investigation of the consequences of senescent non-tumor cells with the aim to improve the radiotherapy of tumors in radiosensitive patients.

Identification of senescence associated genes involved in cell radiation responses

CDKN1A gene was identified as one of the most important key players linked to the identified senescence associated pathways for both 1 Gy/sensitive and 10 Gy/normal treatment conditions. For both conditions the expression of the CDKN1A was up-regulated for all considered time points. CDKN1A is a well-known damage response gene for which aberrant transcriptional response has been associated with abnormal sensitivity to ionizing radiation [42, 43]. The study by Badie et al. (2008) has shown that a subgroup of breast cancer patients, who developed severe reactions to radiation therapy, could be identified by aberrant overexpression of CDKN1 in peripheral blood lymphocytes [43].

LMNB1 is another genes we identified as a response hub gene after irradiation of sensitive cell line with 1 Gy radiation dose that is associated with senescence. Although the LMNB1 gene was not identified as hub gene in the GAN of the 10 Gy/normal treatment condition, it was still differentially expressed. For both treatment conditions we observed significant downregulation of this gene 24 hours after irradiation. Shah et al (2013) has suggested that downregulation of LMNB1 in senescence is a key trigger of chromatin changes affecting gene expression [44]. In fact also in our data we observed strong downregulation of a group of histone genes associated

with senescence (S5 Table) for the treatment conditions 1 Gy/increased sensitivity and 10 Gy/normal sensitivity. Furthermore, Lee et al. (2012) has shown that histone protein modification may have an impact on the radiation sensitivity of a tissue [45]. Moreover, evidence has been provided that mutation of LMNA can cause increased sensitivity to ionizing radiation [46], however, to our knowledge there are no data showing the role of LMNB gene in the context of radiation sensitivity.

Another potential therapeutic candidate associated with senescence that was identified for the 10 Gy/normal sensitivity treatment condition was MRE11A for which cell culture data suggest that treatment of cells with Mre11 siRNA increases radiation sensitivity and reduces heat-induced radiosensitization [47, 48]. However, the clinical applicability of MRE11, has not been confirmed [49].

Assessment of the false positive rate and validation of the NCSRМ method

The spline regression based differential analyses between technical replicates were performed in order to estimate the extent of random fluctuations of gene expression values. The detected 3% rejections of the overall null hypothesis of no differential gene expression are in accordance with the alpha-level of 5% of the familywise error rate (FWER) and can be considered as false positives. On the other hand, it shows that type I error, due to technical variation, is covered by the model and test assumptions (moderated F-test, [50]) so that it was not necessary to include an extra parameter for technical replicates into the model.

In order to validate the previously mentioned biological results using NCSRМ, we performed the differential expression analysis with another established method for time-course data analysis called BETR (Bayesian Estimation of Temporal Regulation) [6]. The number of genes detected by BETR was considerably lower compared to NCSRМ (Table 1), however the majority of which were also detected with NCSRМ (S1 Table). This is in line with the calculations on the false positive rates that have been conducted on the simulated data presented in the BETR study. In an analysis of the simulated data set, 65% of truly differentially expressed genes have been identified after accepting a false positive rate of 5% [6]. This means that a substantial proportion of differentially expressed genes remained undetected, which is likely to be also the case for the herein analyzed data with BETR. Although the numbers of differentially expressed genes and genes remained in the reconstructed networks greatly differ (Table 1), the qualitative results are well comparable (Table 3). For all treatment conditions where for which we were able to reconstruct GANs, we observed a great overlap of pathways where the 5% of hub genes were mapped to (Table 3). The detection of a higher number of differentially expressed genes with NCSRМ resulted in larger GANs with additional information compared to the smaller GANs that were reconstructed on the basis of genes detected with BETR. This is underlined by the results of the conducted evaluation of GANs. Except one network based on the differentially expressed genes using BETR, all investigated networks consist significantly more common edges with the Reactome reference network compared to random networks with identical network topology and genes. This shows that the additionally detected genes with NCSRМ add additional information rather than adding false positives or noise to the set of differentially expressed genes. Moreover the spline regression method is much more flexible and allows for more freedom during the data collection process. As already mentioned, NCSRМ does not require the same sampling time for treated and control groups and can easily deal with incomplete data, whereas BETR method is not able to overcome or bypass those limitations. Thus, NCSRМ is very robust against the frequently occurring shortcomings in study design and subsequent data generation occurring in life sciences.

Conclusion

Prospectively, we suggest and plan a detailed *in silico* and *in vitro* analysis of the interactions in the proposed gene association networks in order to add meaningful knowledge to the mechanism of radiosensitivity at the experimental level. This novel knowledge has the potential to improve cancer radiation therapy by preventing or lowering the acute responses of normal cells resulting from radiation therapy. The results add novel information to the understanding of mechanisms that are involved in the radiation response of human cells, with the potential to improve tumor radiotherapy. Besides, the presented workflow is not limited to presented study only, but may be applied in other special fields with different biological questions to be addressed.

The software is provided as R-package “splineTimeR” and freely available via the Bioconductor project at <http://www.bioconductor.org>.

Material and Methods

Cell culture

Experiments were conducted with two monoclonal lymphoblastoid Epstein-Barr virus-immortalized cell lines (LCL) obtained from young lung cancer patients of the LUCY study (LUNG Cancer in Young) that differ in radiosensitivity, as tested with Trypan Blue and WST-1 assays [19, 20]. The non-cancer cell lines LCL 4060–200 with increased radiation sensitivity and LCL 20037–200 with normal radiation sensitivity were cultured at 37°C/5% CO₂ in RPMI 1640 medium (Biochrom) supplemented with 10% fetal calf serum (FCS; PAA). Mycoplasma contamination was routinely tested using luminescence-based assays (MycoAlert, Lonza).

Irradiation and sample preparation

The cells were seeded in 75 cm² flasks at a concentration of 0.5 x 10⁶ cells/ml in a total volume of 60 ml. Exponentially growing cells were irradiated with sham, 1 Gy and 10 Gy of gamma-irradiation (¹³⁷Cs-source HWM-D 2000, Markdorf, Germany) at a dose rate of 0.49 Gy/min. Samples were collected 0.25, 0.5, 1, 2, 4, 8 and 24 hours after sham or actual irradiation. Between the time of collection cells were kept in the incubator. Collected cells were washed with PBS and frozen at -80°C. Total RNA was isolated from frozen cell pellets obtained from two independent experiments using the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen) including an DNase digestion step, according to the manufacturer's protocol. The concentration of RNA was quantified with a Qubit 2.0 Fluorometer (Life Technologies), and integrity was determined using a Bioanalyzer 2100 (Agilent Technologies). RNA samples with a RNA integrity number (RIN) greater than 7 indicated sufficient quality to be used in subsequent RNA microarray analysis.

Gene expression profiling

Transcriptional profiling was performed using SurePrint G3 Human Gene Expression 8x60k V2 microarrays (Agilent Technologies, AMADID 39494) according to the manufacturer's protocol. 75 ng of total RNA was used in labeling using the Low Input Quick Amp Labeling Kit (one-color, Agilent Technologies). Raw gene expression data were extracted as text files with the Feature Extraction software 11.0.1.1 (Agilent Technologies). The expression microarray data were uploaded to ArrayExpress (www.ebi.ac.uk/arrayexpress/) and the data set is available under the accession number E-MTAB-4829. All data analysis was conducted using the R statistical platform (version 3.2.2, www.r-project.org) [51]. Data quality assessment, filtering, pre-processing, normalization, batch correction based on nucleic acid labeling batches and data

analyses were carried out with the Bioconductor R-packages limma, Agi4x44PreProcess and the ComBat function of the sva R-package [4, 21, 52]. All quality control, filtering, preprocessing and normalization thresholds were set to the same values as suggested in Agi4x44PreProcess R-package user guide [21]. Only HGNC annotated genes were used in the analysis. For multiple microarray probes representing the same gene the optimal probe was selected according to the Megablast score of probe sequences against the human reference sequence (<http://www.ncbi.nlm.nih.gov/refseq/>) [53]. If the resulted score was equal for two or more probes, the probe with the lowest differential gene expression FDR value was kept for further analyses since only one expression value per gene was allowed in subsequent GAN reconstruction analysis.

Spline regression model for two-way experimental design

A natural cubic spline regression model (NCSRM) with three degrees of freedom for an experimental two-way design with one treatment factor and time as a continuous variable was fitted to the experimental time-course data. The mathematical model is defined by the following eq (1):

$$y = y(t, x) = b_0 + b_1B_1(t - t_0) + b_2B_2(t - t_0) + \dots + b_mB_m(t - t_0) + x(d_0 + d_1B_1(t - t_0) + d_2B_2(t - t_0) + \dots + d_mB_m(t - t_0))$$

where b_0, b_1, \dots, b_m are the spline coefficients in the control group and d_0, d_1, \dots, d_m are differential spline coefficients between the control and the irradiated group. $B_1(t-t_0), B_2(t-t_0), \dots, B_m(t-t_0)$ are the spline base functions and t_0 is the time of the first measurement. For $x = 0$, $y = y_{\text{control}}$ and for $x = 1$, $y = y_{\text{irradiated}}$. For three degrees of freedom ($df = 3$), $m = 3$.

Depending on the number of degrees of freedom, two boundary knots and $df-1$ interior knots are specified. The interior knots were chosen at values corresponding to equally sized quantiles of the sampling time from both compared groups. For example, for $df = 3$ interior knots correspond to the 0.33- and 0.66-quantiles. The spline function is cubic on each defined by knots intervals, continuous at each knot and has continuous derivatives of first and second orders.

Time-course differential gene expression analysis

The time-course differential gene expression analyses were conducted between irradiated and control cells (sham-irradiated). Analyses were performed on the normalized gene expression data using NCSRM with three degrees of freedom. The splines were fitted to the real time-course expression data for each gene separately according to eq (1). The example of spline regression model fitted to the measured time-course data for one selected gene is shown on the Fig 2.

Time dependent differential expression of a gene between the irradiated and corresponding control cells was determined by the application of empirical Bayes moderated F-statistics [50] on the differential coefficients values in eq (1). In order to account for the multiple-testing error, corresponding p-values were adjusted by the Benjamini-Hochberg method for false discovery [22]. Genes with an adjusted p-value (FDR, false discovery rate) lower than 0.05 were considered as differentially expressed and associated with radiation response.

Assessment of the false positive rate of the NCSRM

Additionally, in order to assess the false positive rate (statistical type I error, also called familywise error rate or FWER) we applied differential gene expression analysis using NCSRM between two technical replicates for all treatment groups. Because only two technical replicates were generated for each time point and treatment, we could not use the same approach to assess the technical variability for the BETR method, as it requires at least two replicates in each compared groups.

Gene association network reconstruction from prior selected differentially expressed genes

Differentially expressed genes were subjected to gene association network reconstruction from time-course data using a regularized dynamic partial correlation method [54]. Pairwise relationships between genes over time were inferred based on a dynamic Bayesian network model with shrinkage estimation of covariance matrices as implemented in the GeneNet R-package available from CRAN [18]. Analyses were conducted with a posterior probability of 0.95 for each potential edge. Edge directions were not considered. In order to assess the complexity of the resulting networks, the density of each network was compared to the density of the Reactome functional interaction network [35, 55].

Identification of important nodes in the network

Graph topological analyses based on centrality measures were applied in order to determine the importance of each node in the reconstructed association networks [56]. Three most

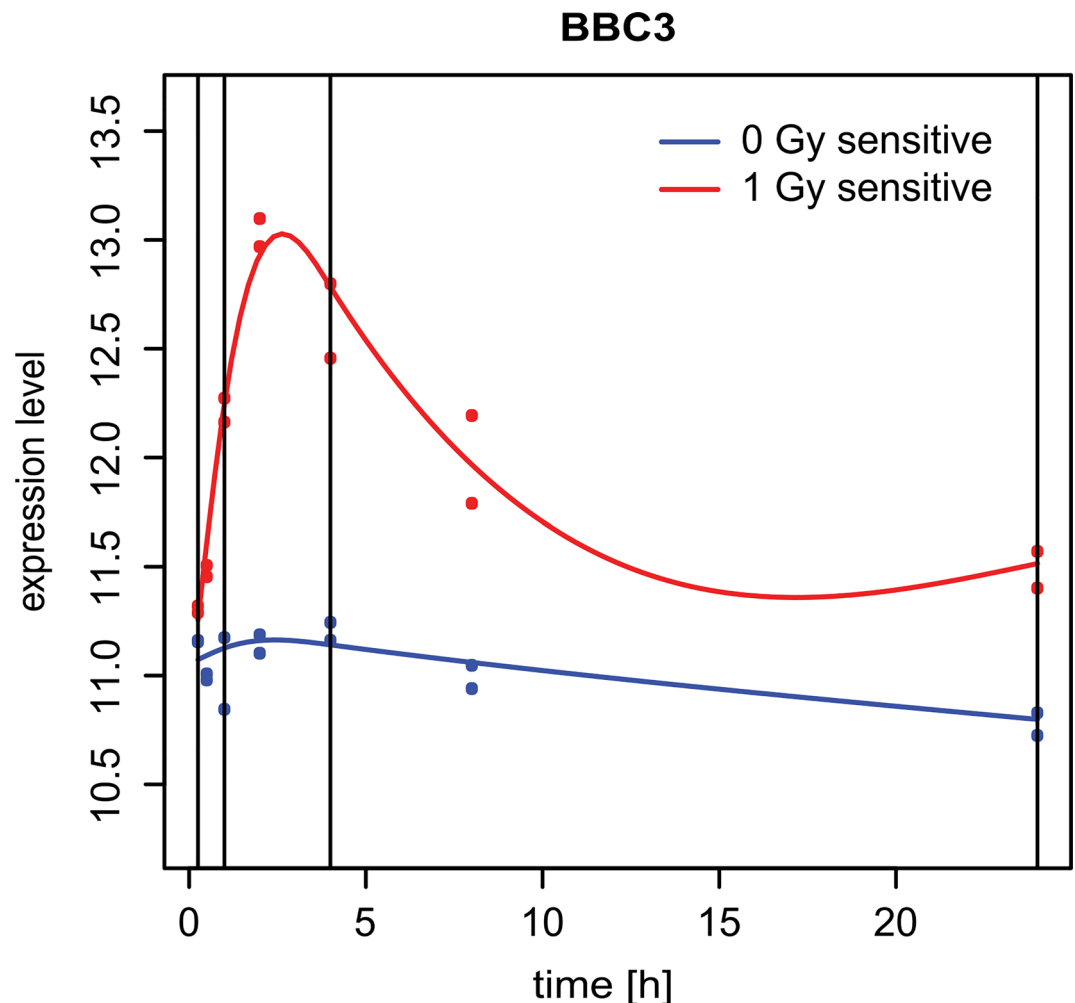


Fig 2. Example of fitted spline regression models. The plot shows spline regression models fitted to the measured time-course expression data of an arbitrary chosen gene (BBC3). The blue line represents the fitted model for the control (0 Gy) and read line that for the irradiated group (1 Gy). Blue and red dots represent the measured expression levels of the biological replicates. Vertical lines represent the endpoints and interior knots correspond to the 0.33- and 0.66-quantiles.

doi:10.1371/journal.pone.0160791.g002

commonly used centrality measures: degree, shortest path betweenness and closeness were combined into one cumulative centrality measure [34]. For each gene the three centrality values were ranked. The consensus centrality measure for each node was defined as the mean of the three independent centrality ranks. Combining centrality measures supports the identification of the nodes that are central themselves and also connected to direct central nodes, which demonstrates strategic positions for controlling the network.

Pathway enrichment analysis

The Reactome pathway database was used to conduct the pathway enrichment analysis in order to further investigate the functions of the selected sets of differentially expressed genes [35]. Statistical significance of enriched pathways was determined by one-sided Fisher's exact test. The resulting p-values were adjusted for FDR using the Benjamini-Hochberg method. Pathways with $FDR < 0.05$ were considered statistically significant and pathways were ranked according to ascending FDRs.

Evaluation of NCSR approach

Since we decided to use the set of genes that appeared to be differentially expressed we assessed the performance of the herein used NCSR approach in comparison to the BETR approach implemented in the R/Bioconductor package `betr` [6]. BETR is a well-established algorithm that has been previously compared to `limma`, MB-statistic and EDGE methods and showed the best performance [6]. The results of spline and BETR methods were compared using the same initial microarray gene expression data set. The probabilities of each gene to be differentially expressed obtained with BETR method, were transformed to p-values as described in the original paper. Genes were considered significantly differentially expressed if the Benjamini-Hochberg adjusted p-value was lower than 0.05 ($FDR < 0.05$). This transformation allowed us to compare the outcomes of both methods based on the FDR values for differential expression. The resulting differentially expressed genes using BETR were analyzed and subjected to network reconstruction as described above for the differentially expressed genes obtained using NCSR. Outcomes of both obtained association networks were compared to each other and to the *a priori* known biological network provided by the Reactome database [35].

Evaluation of reconstructed gene association networks

In order to assess the quality of the *de novo* reconstructed gene association networks (GANs), we developed a novel method that compares the interactions in the reconstructed network to the experimentally validated interactions present in the Reactome interaction network. For this purpose we used the Reactome reference network, consisting of protein-protein interaction pairs stored in the Reactome database (<http://www.reactome.org/pages/download-data/>). For the comparison, sub-networks of reconstructed networks consisting only of genes overlapping with the Reactome network were built. The number of common edges between these two sub-networks was determined and referred to the total number of edges in the reconstructed network (percentage of common edges in the reconstructed network). Further, a permutation test was performed to assess whether the number of common edges in the reconstructed network was significantly higher than in randomized networks with the same genes. Random networks were generated by permutation of the node names in the network, while preserving the reconstructed sub-network topology. After each permutation ($n = 1000$) the number of common edges with the reference Reactome sub-network was determined. The reconstructed network was considered significantly better than random, if more than 90% of the random sub-networks contained lower numbers of edges common with the Reactome network than the

reconstructed sub-network (p -value < 0.1). All networks reconstructed with the genes determined as differentially expressed from the herein presented spline regression method and the BETR method were evaluated.

Supporting Information

S1 File. Reconstructed gene association networks. All obtained gene association networks are provided as R-objects of type igraph.
(RDATA)

S1 Table. Lists of differentially expressed genes. Table includes differentially expressed genes identified by spline regression and BETR methods. Additionally, a list of overlapping differentially expressed genes between both methods is included.
(XLSX)

S2 Table. Lists of significantly enriched pathways using differentially expressed genes identified by spline regression method. Four lists of significantly enriched pathways correspond to each used treatment condition. Lists include total numbers of known genes in the pathways, numbers of differentially expressed genes that belong to a single pathway (matches), percentages of differentially expressed genes in comparison to the total number of known genes in the pathway (% match), p -values, FDRs and names of pathways related differentially expressed genes.
(XLSX)

S3 Table. Lists of 5% of most important genes identified by centrality measures. Lists of 5% highest ranked genes from the reconstructed gene association networks using spline regression and BETR methods. Overlap represents common most important genes identified in networks from compared methods.
(XLSX)

S4 Table. Lists of pathways after mapping of 5% highest ranked genes from the reconstructed gene association networks. Lists include names of pathways together with names of mapped most important genes.
(XLSX)

S5 Table. Significantly enriched senescence associated pathways with corresponding differentially expressed genes. Table presents the names of significantly enriched ($FDR < 0.05$) senescence associated pathways with corresponding differentially expressed genes for all treatment conditions.
(XLSX)

Acknowledgments

We thank Aaron Selmaier from the Research Unit Radiation Cytogenetics for technical support. This study was supported by the Federal Ministry of Education and Research, ZiSS project, 02NUK024B.

Author Contributions

Conceived and designed the experiments: AD AM KU HZ NB JH MG SH.

Performed the experiments: AM AD.

Analyzed the data: AM HB MS.

Contributed reagents/materials/analysis tools: AM AD MS KU MG SH HZ.

Wrote the paper: AM KU MS.

Biological data interpretation: AM KU HZ JH MS MG SH NB. Designed the software used in analysis: AM HB. Revision of the manuscript: KU HZ JH MG SH NB AD HB. Final approval of the version to be published: AM HB MS AD JH MG SH NB HZ KU.

References

1. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*. 2012; 13:552–64. doi: [10.1038/nrg3244](https://doi.org/10.1038/nrg3244) PMID: [22805708](https://pubmed.ncbi.nlm.nih.gov/22805708/)
2. Bandyopadhyay S, Bhattacharyya M. A biologically inspired measure for coexpression analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2011; 8(4):929–42. doi: [10.1109/TCBB.2010.106](https://doi.org/10.1109/TCBB.2010.106) PMID: [21566252](https://pubmed.ncbi.nlm.nih.gov/21566252/)
3. Yuan M, Kendziorski C. Hidden Markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association*. 2006; 101(476):1323–32.
4. Smyth GK. Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor'. New York: Springer, p. 397–420; 2005.
5. Leek JT, Monsen E, Dabney AR, Storey JD. EDGE: extraction and analysis of differential gene expression. *Bioinformatics*. 2006; 22(4):507–8. PMID: [16357033](https://pubmed.ncbi.nlm.nih.gov/16357033/)
6. Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*. 2009; 10(409).
7. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*. 2006; 22(9):1096–102. PMID: [16481333](https://pubmed.ncbi.nlm.nih.gov/16481333/)
8. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*. 2006; 7(191).
9. Schliep A, Steinhoff C, Schönhuth A. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*. 2004; 20:i283–i9. PMID: [15262810](https://pubmed.ncbi.nlm.nih.gov/15262810/)
10. Magni P, Ferrazzi F, Sacchi L, Bellazzi R. TimeClust: a clustering tool for gene expression time series. *Bioinformatics*. 2008; 24(3):430–2. PMID: [18065427](https://pubmed.ncbi.nlm.nih.gov/18065427/)
11. Sivriver J, Habib N, Friedman N. An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*. 2011; 27(13):i392–i400. doi: [10.1093/bioinformatics/btr250](https://doi.org/10.1093/bioinformatics/btr250) PMID: [21685097](https://pubmed.ncbi.nlm.nih.gov/21685097/)
12. Sinha A, Markatou M. A Platform for Processing Expression of Short Time Series (PESTS). *BMC Bioinformatics*. 2011; 12(13).
13. Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci USA*. 2002; 99:9121–6. PMID: [12082179](https://pubmed.ncbi.nlm.nih.gov/12082179/)
14. Lin T, Kaminski N, Bar-Joseph Z. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*. 2008; 24:i147–i55. doi: [10.1093/bioinformatics/btn152](https://doi.org/10.1093/bioinformatics/btn152) PMID: [18586707](https://pubmed.ncbi.nlm.nih.gov/18586707/)
15. Costa IG, Schönhuth A, Hafemeister C, Schliep A. Constrained mixture estimation for analysis and robust classification of clinical time series. *Bioinformatics*. 2009; 25:i6–i14. doi: [10.1093/bioinformatics/btp222](https://doi.org/10.1093/bioinformatics/btp222) PMID: [19478017](https://pubmed.ncbi.nlm.nih.gov/19478017/)
16. Hafemeister C, Costa IG, Schönhuth A, Schliep A. Classifying short gene expression time-courses with Bayesian estimation of piecewise constant functions. *Bioinformatics*. 2011; 27(7):946–52. doi: [10.1093/bioinformatics/btr037](https://doi.org/10.1093/bioinformatics/btr037) PMID: [21266444](https://pubmed.ncbi.nlm.nih.gov/21266444/)
17. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. *PNAS*. 2005; 102(36):12837–42. PMID: [16141318](https://pubmed.ncbi.nlm.nih.gov/16141318/)
18. Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*. 2007; 1:37. PMID: [17683609](https://pubmed.ncbi.nlm.nih.gov/17683609/)
19. Rosenberger A, Rossler U, Hornhardt S, Sauter W, Bickeboller H, Wichmann HE, et al. Validation of a fully automated COMET assay: 1.75 million single cells measured over a 5 year period. *DNA Repair (Amst)*. 2011; 10(3):322–37.

20. Guertler A, Kraemer A, Roessler U, Hornhardt S, Kulka U, Moertl S, et al. The WST survival assay: an easy and reliable method to screen radiation-sensitive individuals. *Radiat Prot Dosimetry*. 2011; 143(2–4):487–90. doi: [10.1093/rpd/ncq515](https://doi.org/10.1093/rpd/ncq515) PMID: [21183542](https://pubmed.ncbi.nlm.nih.gov/21183542/)
21. Lopez-Romero P. Agi4x44PreProcess: PreProcessing of Agilent 4x44 array data. R package version 1.16.0.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995; 57:289–300.
23. Alsner J, Andreassen CN, Overgaard J. Genetic markers for prediction of normal tissue toxicity after radiotherapy. *Seminars in Radiation Oncology*. 2008; 18:126–35. doi: [10.1016/j.semradonc.2007.10.004](https://doi.org/10.1016/j.semradonc.2007.10.004) PMID: [18314067](https://pubmed.ncbi.nlm.nih.gov/18314067/)
24. Andreassen CN. Can risk of radiotherapy-induced normal tissue complications be predicted from genetic profiles? *Acta Oncologica*. 2005; 44:801–15. PMID: [16332587](https://pubmed.ncbi.nlm.nih.gov/16332587/)
25. Liu H, Andrews DW, Evans JJ, Werner-Wasik M, Yu Y, Dicker AP, et al. Plan quality and treatment efficiency for radiosurgery to multiple brain metastases: non-coplanar RapidArc vs. Gamma Knife. *Frontiers in Oncology*. 2016; 6(26).
26. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research*. 2012; 22:577–91. doi: [10.1101/gr.133009.111](https://doi.org/10.1101/gr.133009.111) PMID: [22110045](https://pubmed.ncbi.nlm.nih.gov/22110045/)
27. Lienau J, Schmidt-Bleek K, Peters A, Weber H, Bail HJ, Duda HN, et al. Insight into the molecular pathophysiology of delayed bone healing in a sheep model. *Tissue Engineering Part A*. 2010; 16(1):191–9. doi: [10.1089/ten.TEA.2009.0187](https://doi.org/10.1089/ten.TEA.2009.0187) PMID: [19678759](https://pubmed.ncbi.nlm.nih.gov/19678759/)
28. Schell H, Thompson MS, Bail HJ, Hoffmann J-E, Schilla A, Duda GN, et al. Mechanical induction of critically delayed bone healing in sheep: radiological and biomechanical results. *Journal of Biomechanics*. 2008; 41(14):3066–72. doi: [10.1016/j.jbiomech.2008.06.038](https://doi.org/10.1016/j.jbiomech.2008.06.038) PMID: [18778822](https://pubmed.ncbi.nlm.nih.gov/18778822/)
29. Gürtler A, Hauptmann M, Pautz S, Kulka U, Friedl AA, Lehr S, et al. The inter-individual variability outperforms the intra-individual variability of differentially expressed proteins prior and post irradiation in lymphoblastoid cell lines. *Arch Physiol Biochem*. 2014; 120(5):198–207. doi: [10.3109/13813455.2014.953548](https://doi.org/10.3109/13813455.2014.953548) PMID: [25174346](https://pubmed.ncbi.nlm.nih.gov/25174346/)
30. Azzama EI, Jay-Gerinb J-P, Pain D. Ionizing radiation-induced metabolic oxidative stress and prolonged cell injury. *Cancer Letters*. 2012; 327(1–2):48–60. doi: [10.1016/j.canlet.2011.12.012](https://doi.org/10.1016/j.canlet.2011.12.012) PMID: [22182453](https://pubmed.ncbi.nlm.nih.gov/22182453/)
31. Li L, Story K, Legerski RJ. Cellular responses to ionizing radiation damage. *International Journal of Radiation Oncology, Biology, Physics*. 2001; 49(4):1157–62. PMID: [11240259](https://pubmed.ncbi.nlm.nih.gov/11240259/)
32. Jung M, Dritschilo A. Signal transduction and cellular responses to ionizing radiation. *Seminars in Radiation Oncology*. 1996; 6(4):268–72. PMID: [10717184](https://pubmed.ncbi.nlm.nih.gov/10717184/)
33. Koschützki D. Network Centralities. In: Junker BH, Schreiber F, editors. *Analysis of Biological Networks*: Wiley; 2007. p. 65–84.
34. Abbasi A, Hossain L. Hybrid Centrality Measures for Binary and Weighted Networks. In: Menezes R, Evsukoff A, González MC, editors. *Complex Networks*. 424. Berlin: Springer; 2013. p. 1–7.
35. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*. 2009; 37(Database issue):D619–D22. doi: [10.1093/nar/gkn863](https://doi.org/10.1093/nar/gkn863) PMID: [18981052](https://pubmed.ncbi.nlm.nih.gov/18981052/)
36. Sabin RJ, Anderson RM. Cellular Senescence—its role in cancer and the response to ionizing radiation. *Genome Integr* 2011; 2: 7 2011; 2(1):7. doi: [10.1186/2041-9414-2-7](https://doi.org/10.1186/2041-9414-2-7) PMID: [21834983](https://pubmed.ncbi.nlm.nih.gov/21834983/)
37. Meng Y, Efimova EV, Hamzeh KW, Darga TE, Mauceri HJ, Fu YX, et al. Radiation-inducible immunotherapy for cancer: senescent tumor cells as a cancer vaccine. *Molecular Therapy*. 2012; 20(5):1046–55. doi: [10.1038/mt.2012.19](https://doi.org/10.1038/mt.2012.19) PMID: [22334019](https://pubmed.ncbi.nlm.nih.gov/22334019/)
38. Freund A, Orjalo AV, Desprez PY, Campisi J. Inflammatory networks during cellular senescence: causes and consequences. *Trends Mol Med*. 2010; 16(5):238–46. doi: [10.1016/j.molmed.2010.03.003](https://doi.org/10.1016/j.molmed.2010.03.003) PMID: [20444648](https://pubmed.ncbi.nlm.nih.gov/20444648/)
39. Nelson G, Wordworth J, Wang C, Jurk D, Lawless C, Martin-Ruiz C, et al. A senescent cell bystander effect: senescence-induced senescence. *Aging Cell*. 2012; 11(2):345–9. doi: [10.1111/j.1474-9726.2012.00795.x](https://doi.org/10.1111/j.1474-9726.2012.00795.x) PMID: [22321662](https://pubmed.ncbi.nlm.nih.gov/22321662/)
40. Wu PC, Wang Q, Grobman L, Chu E, Wu DY. Accelerated cellular senescence in solid tumor therapy. *Experimental Oncology*. 2012; 34(3):298–305. PMID: [23070015](https://pubmed.ncbi.nlm.nih.gov/23070015/)
41. Tsai KK, Stuart J, Chuang YY, Little JB, Yuan ZM. Low-dose radiation-induced senescent stromal fibroblasts render nearby breast cancer cells radioresistant. *Radiation Research*. 2009; 172(3):306–13. doi: [10.1667/RR1764.1](https://doi.org/10.1667/RR1764.1) PMID: [19708779](https://pubmed.ncbi.nlm.nih.gov/19708779/)

42. Amundson SA, Grace MB, McLeland CB, Epperly MW, Yeager A, Zhan Q, et al. Human in vivo radiation-induced biomarkers: gene expression changes in radiotherapy patients. *Cancer Research*. 2004; 64(18):6368–71. PMID: [15374940](#)
43. Badie C, Dziwura S, Raffy C, Tsigani T, Alsbeih G, Moody J, et al. Aberrant CDKN1A transcriptional response associates with abnormal sensitivity to radiation treatment. *British Journal of Cancer*. 2008; 98(11):1845–51. doi: [10.1038/sj.bjc.6604381](#) PMID: [18493234](#)
44. Shah PP, Donahue G, Otte GL, Capell BC, Nelson DM, Cao K, et al. Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes & Development*. 2013; 27:1787–99.
45. Lee M, Urata SM, Aguilera JA, Perry CC, Milligan JR. Modeling the influence of histone proteins on the sensitivity of DNA to ionizing radiation. *Radiation Research*. 2012; 177(2):152–63. PMID: [22103271](#)
46. di Masi A, D'Apice MR, Ricordy R, Tanzarella C, Novelli G. The R527H mutation in LMNA gene causes an increased sensitivity to ionizing radiation. *Cell Cycle*. 2008; 7(13):2030–7. PMID: [18604166](#)
47. Xu M, Myerson R, Hunt C, Kumar S, Moros E, Straube B, et al. Treatment of cells with Mre11 siRNA increases radiation sensitivity and reduces heat induced radiosensitization. *International Journal of Radiation Oncology Biology Physics*. 2003; 57(2):144–5.
48. Xu M, Myerson RJ, Hunt C, Kumar S, Moros EG, Straube WL, et al. Transfection of human tumour cells with Mre11 siRNA and the increase in radiation sensitivity and the reduction in heat-induced radiosensitization. *International Journal of Hyperthermia*. 2004;20(2).
49. Söderlund K, Stål O, Skoog L, Rutqvist LE, Nordenskjöld B, Askmalm MS. Intact Mre11/Rad50/Nbs1 complex predicts good response to radiotherapy in early breast cancer. *International Journal of Radiation Oncology Biology Physics*. 2007; 68(1):50–8.
50. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3(1):Article 3.
51. R Core Team. R: A language and environment for statistical computing 2013.
52. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):R80. PMID: [15461798](#)
53. Li Q, Birnbak NJ, Györfy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011; 12(474).
54. Opgen-Rhein R, Strimmer K. Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. *The 4th International Workshop on Computational Systems Biology, WCSB*. 2006.
55. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*: Cambridge: Cambridge University Press; 1994.
56. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*. 2008; 2:193–201. PMID: [19787083](#)

Copy number aberrations from Affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches?

Adriana Pitea, Ivan Kondofersky, Steffen Sass, Fabian J. Theis, Nikola S. Mueller and Kristian Unger

Corresponding author: Nikola S. Mueller, Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, 85764, Germany. Tel: +49-89-3187-1174; Fax: +49-89-3187-3369; Email: nikola.mueller@helmholtz-muenchen.de; Kristian Unger, Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, Neuherberg, 85764, Germany and Clinical Cooperation Group Personalized Radiotherapy in Head and Neck Cancer, Helmholtz Zentrum München, Neuherberg, 85764, Germany. Email: unger@helmholtz-muenchen.de

Abstract

Copy number aberrations (CNAs) are known to strongly affect oncogenes and tumour suppressor genes. Given the critical role CNAs play in cancer research, it is essential to accurately identify CNAs from tumour genomes. One particular challenge in finding CNAs is the effect of confounding variables. To address this issue, we assessed how commonly used CNA identification algorithms perform on SNP 6.0 genotyping data in the presence of confounding variables. We simulated realistic synthetic data with varying levels of three confounding variables—the tumour purity, the length of a copy number region and the CNA burden (the percentage of CNAs present in a profiled genome)—and evaluated the performance of OncoSNP, ASCAT, GenoCNA, GISTIC and CGHcall. Furthermore, we implemented and assessed CGHcall*, an adjusted version of CGHcall accounting for high CNA burden. Our analysis on synthetic data indicates that tumour purity and the CNA burden strongly influence the performance of all the algorithms. No algorithm can correctly find lost and gained genomic regions across all tumour purities. The length of CNA regions influenced the performance of ASCAT, CGHcall and GISTIC. OncoSNP, GenoCNA and CGHcall* showed little sensitivity. Overall, CGHcall* and OncoSNP showed reasonable performance, particularly in samples with high tumour purity. Our analysis on the HapMap data revealed a good overlap between CGHcall, CGHcall* and GenoCNA results and experimentally validated data. Our exploratory analysis on the TCGA HNSCC data revealed plausible results of CGHcall, CGHcall* and GISTIC in consensus HNSCC CNA regions. Code is available at <https://github.com/adspit/PASCAL>.

Key words: copy number calling algorithm; performance assessment; cancer genomics; copy number aberrations

Adriana Pitea is a PhD student at the Computational Cell Maps, Institute of Computational Biology and at the Integrative Biology Group, Research Unit of Radiation, Helmholtz Zentrum München.

Ivan Kondofersky is a Postdoctoral Fellow at the Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.

Steffen Sass is former Postdoctoral Fellow at the Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.

Fabian J. Theis is Head of the Institute of Computational Biology and Group Leader Machine Learning, Helmholtz Zentrum München and associate professor holding the chair of 'Mathematical modeling of biological systems', Department of Mathematics, Technical University of Munich.

Nikola S. Mueller is Group Leader of Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München.

Kristian Unger is Head of the Integrative Biology Group and deputy Head of the Research Unit of Radiation, Helmholtz Zentrum München.

Submitted: 21 May 2018; Received (in revised form): 11 August 2018

© The Author(s) 2018. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Copy number aberrations (CNAs) are present in all known cancer genomes [1–3]. Unlike copy number variations (CNVs) which occur naturally and originate in germline cells [4–6], CNAs accumulate somatically, emerge after many selection events and have been associated with development and progression of human disease, especially with carcinogenesis: Bardeesy *et al.* showed that the deletion of the tumour suppressor gene SMAD4 plays a critical role in progression and tumour biology of pancreatic cancer [7], Witkiewicz *et al.* showed that amplification of the gene MYC is uniquely associated with poor outcome in pancreatic ductal adenocarcinoma [8], Leucci *et al.* showed that the long non-coding RNA (lncRNA) gene SAMMSON is consistently co-gained with MITF in more than 90% of human melanomas [9], while Wells *et al.* showed that deletion of the gene PTGHD1 in the thalamic reticular nucleus only leads to attention deficiency and hyperactivity [10]. Identifying CNAs that are affecting oncogenes or tumour suppressor genes provides knowledge required for the development of new targeted cancer therapies or patient stratification. It is thus of great importance to accurately estimate CNAs from tumour genomes. However, one particular challenge in the accurate estimation of cancer-related CNAs is the presence of confounding variables such as tumour purity and length of CNAs.

The tumour purity represents the ratio between cancerous cells and all the cells present in a tumour sample—comprising both of cancerous and non-cancerous cells. The mixture of cancerous and non-cancerous cells affects the expected allelic fraction between germline and somatic variants and thus influences the accuracy of CNA calling [11]. In simple terms, the higher the non-tumour cell content within the assessed tissue sample, the lower the sensitivity of the copy number calling algorithm gets. Previous studies have shown that the length of a CNA region, i.e. the number of covered base pairs by a genomic region, affects the sensitivity of CNA calling, with longer CNA regions being easier to find [12, 13].

Within this study we focus on algorithms that call CNAs from single-nucleotide polymorphism (SNP) arrays. Nowadays, SNP arrays typically comprise approximately 1.8 million probes and return allele-specific signals at each marker of genetic variation. Affymetrix SNP 6.0 data also come with the great advantage that they can be used for both genotype and copy number analysis. Another advantage of this technology is that it allows us to characterise both copy number changes and allelic imbalances of a sample. To achieve this, the signals resulting from the array genotyping need to be processed and analysed by specific methods. Although numerous methods have been proposed, reliably uncovering cancer-associated CNAs from SNP array data still represents a challenge [3, 14, 15]. One difficulty is that CNA calling algorithms fail to address the effect of known biological confounding variables [16, 17], i.e. the tumour purity of the analysed tissue and the length of underlying CNA regions. GenoCN represents a statistical framework that simultaneously searches for CNAs and CNVs while taking into account the tumour purity but does not account for a chromosomal background that is not diploid [18]. OncoSNP represents a unified Bayesian framework based on a cancer-specific statistical model that classifies SNP array signals into 21 states and accounts for tumour purity, ploidy and intra-tumour heterogeneity [19]. ASCAT focuses on analysing allele-specific copy numbers in solid tumour initially but requires a threshold-based, model-free segmentation of the SNPs into regions of equal copy number [6]. Another method that is used for finding cancer-related CNAs is CGHcall. CGHcall

makes use of breakpoint information from segmentation across all samples and includes information as tumour purity for finding CNAs [20].

The Cancer Genome Atlas (<https://cancergenome.nih.gov>) (TCGA) is one of the largest resources providing molecular omics data on multiple levels. TCGA covers various cancer types and aims to improve general knowledge about cancer development and treatment. The commonly used method to estimate copy number states from SNP genotyping data in TCGA studies is GISTIC 2.0 (GISTIC) [21]. GISTIC was designed to primarily estimate significant relative CNAs across a set of patients and not on single patient level. GISTIC eliminates common chromosome arm-level events which are not cancer-specific and focuses on focal events. However, GISTIC does not address the effect of confounding variables on the resulting CNA regions.

Within this study we assessed the performance of the following common-used CNA calling algorithms on Affymetrix SNP 6.0 array data: OncoSNP [19], ASCAT [6], CGHcall [20], genoCNA [18] and GISTIC [21]. All algorithms are commonly used for estimating copy number states in tumour samples and, except for GISTIC, correct for tumour purity, intra-tumour heterogeneity and tumour cell ploidy (ASCAT and OncoSNP). Unlike previous studies that evaluated CNV detection—and not cancer-specific CNAs—for an SNP platform [13, 22] or used a model with 24 parameters for which it is difficult to find a combination that provides realistic data [23, 24], we focused on five different algorithms designed to specifically find CNAs and, moreover, evaluated them on synthetic data derived from Affymetrix SNP 6.0 data. Our contribution consists of

- a pipeline that uses realistic Affymetrix SNP 6.0 array-like synthetic DNA copy number profiles for evaluating the performance of OncoSNP, ASCAT, CGHcall, genoCNA and GISTIC CNA calling algorithms, under the influence of tumour purity, length of CNA and CNA burden (the percentage of CNAs present in the profiled genome, [25])
- the implementation of an adjusted version of the CGHcall algorithm that allows the estimation of CNAs in highly variant genomes.

We applied our pipeline on two real data sets derived from patient samples: a cohort of 522 head and neck squamous cell carcinoma (HNSCC) samples from TCGA [26] and a set of 81 Haplotype Map samples [4]. The pipelines consist of R, Python and shell scripts and can be accessed at <https://github.com/adspit/PASCAL>. Finally, we provide an appropriate framework to compare CNAs calling algorithms with the aim of finding the algorithm that classifies genomic regions correctly independent of tumour purity, length of a CNA region and CNA burden. Moreover, we developed an improved version of CGHcall that we refer to as CGHcall* and included it in our comparison.

Methods and materials

Preliminaries

The data resulting from Affymetrix SNP 6.0 arrays experiments comprised of fluorescence intensity values of hybridised A and B allele probes for each genetic marker on the array [27]. We obtained and used the following measures from the data:

- (i) the log R ratio (LRR) – a log₂-transformed value of the total intensity for allele A and allele B for more than 1.8 million markers of genetic variation.

- (ii) the B allele frequency (BAF) – the ratio of bases genotyped as variant allele (B allele). BAF ranged from 0 to 1, where 0 represented the AA/A– genotype, 0.5 represented the heterozygous AB genotype and 1 represented the BB/B– genotype [28].

Realistic synthetic data

We used the jointseg R package [24] to generate realistic Affymetrix SNP 6.0 array-like synthetic tumour data consisting of 400 samples. Each sample comprised of 1.844.399 markers of genetic variation. Jointseg was built to generate realistic synthetic DNA copy number profiles. The framework resamples signals corresponding to genomic regions with manually annotated copy number states from the publicly available lung cancer NCI-H1395 SNP microarray data [24, 29]. We generated 100 samples with each of the following tumour purity levels: 30, 50, 70 and 100%. The tumour purity levels corresponded to the experimental settings of the [29] study. We randomly placed between 1 and 8 breakpoints within each sample. A breakpoint represented a loci where one of the two parental copy number changed. For the resulting regions we sampled the copy number states from a predefined set of copy number states: (0,1), (0,2), (1,1), (1,2), (1,3), (2,2) and (3,2), where (0,1) represented the loss of a single copy, (0,2) and (1,1) represented normal and (1,2), (1,3), (2,2) and (3,2) represented the gain of one, two or three copies.

Haplotype Map data

We started the analysis with 98 Affymetrix 6.0 SNP array profiles of healthy patients from the publicly available Haplotype Map (HapMap) repository: <ftp://ftp.ncbi.nlm.nih.gov/hapmap/> [4]. We preprocessed the data with the Aroma Affymetrix Power Tools package [30] and the PennCNV-Affy pipeline [31]. In the preprocessing step, we performed quantile normalisation and generated genotype calls from the Affymetrix spot intensity readout files (CEL format) as output by the Affymetrix microarray scanner files using the Birdseed algorithm [32]. Next, we extracted allele-specific signals, and we calculated the canonical clustering parameters for each marker of genetic variation. We then calculated probe-wise LRR and BAF for each patient sample. Further, we split the signal file into individual files for each patient. We then selected 81 patients that were further experimentally profiled by Redon *et al.* [4].

HNSCC data

We used Level 1 Affymetrix SNP 6.0 array data generated by the TCGA research network (<http://cancergenome.nih.gov/>) consisting of 522 samples collected from patients suffering from HNSCC [26]. We preprocessed the tumour and normal matched raw HNSCC CEL files with the Affymetrix Power Tools package [30] and the PennCNV-Affy pipeline [31] as described in the previous section.

Genomic copy number calling algorithms

We selected five CNA calling algorithms for comparison: CGH-Call (release 3.6), OncoSNP (version 2.1), ASCAT (version 2.4), genoCNA and GISTIC (version 2.0).

OncoSNP

OncoSNP was built upon a statistical model that classifies SNP array signals—both LRR and BAF, from cancer genomes into 21 states covering different combinations of allele loss and ampli-

fication. The model includes effects of polyploidy, tumour purity and intra-tumour heterogeneity [19]. We applied OncoSNP on the synthetic data with the arguments specific for Affymetrix SNP array, together with the predefined number of training states and tumour states. We used the intratumour mode and set the tumour purity parameter to 30, 50, 70 and 100%. For the HapMap data, we used the same parameter settings, except for the tumour purity which was set to 0.

ASCAT

ASCAT was designed to perform allele-specific CNA analysis in tumour samples. The algorithm corrects for the effects of tumour purity and tumour aneuploidy and infers copy number classes, loss of heterozygosity and homozygous deletions. ASCAT estimates the number of copies for both alleles at all SNP marker positions together with the tumour purity of each sample [6].

We preprocessed the synthetic data and generated the ASCAT-format input tumour LRR and BAF files. Afterwards, we generated corresponding germline genotypes with the `ascat.predictGermlineGenotypes` R function with the platform parameters set to 'AffySNP6'. Finally, we segmented the data with the ASPCF segmentation algorithm and applied the ASCAT copy number calling function. Next, we applied the same steps to the HapMap data.

GenoCNA

GenoCN was built as a statistical framework that simultaneously searches for CNAs and CNVs while inferring the tumour purity. In this study we used the genoCNA component, which was specifically designed for CNA finding. Applying genoCNA required the following information for each of the genetic markers: name, chromosome, position and population frequency (PFB). We used the genetic marker information as provided by the Affymetrix PFB file corresponding to the human genome assembly hg18. Each input file contained LRR, BAF and PFB values for each genetic marker. We selected the output format 2 which returned the most likely copy number and genotype state of all the genetic markers.

GISTIC

GISTIC was designed to find regions of the genome that are significantly amplified or deleted across a set of samples. The significance measure is based on the amplitude of the CNA, on how frequently the CNA occurs across samples and a user-defined threshold for the discovery rate. GISTIC required as input a segmentation file, a reference genome file and the LRR signals. GISTIC does not use the BAF signals. For all data sets we used the hg18 reference genome and segmentation files obtained by applying circular binary segmentation—further referred to as CBS [33]. For the TCGA HNSCC analysis we used the GISTIC results provided by TCGA as level 3 data.

CGHcall

CGHcall was originally designed for array Comparative Genomic Hybridization (aCGH) data. The algorithm uses breakpoint information from CBS [33] and classifies raw \log_2 -ratios between reference and tumour DNA into five discrete states: double loss-homozygous (biallelic) deletion, loss-hemizygous deletion (loss of one of the alleles), normal-two copies, gain-three to four copies and amplification—more than four copies [20]. We \log -transformed the total copy numbers and we applied the CGHcall

pipeline on resulting signals with adjustment for tumour purity. For the HNSCC TCGA data set, we implemented a Python script to calculate \log_2 -ratios between tumour and normal matched patient samples. As the HapMap cohort included only healthy patients, we calculated \log_2 -ratios between each LRR signal and the mean LRR signal of the 81 selected samples.

CGHcall*

We developed an adjusted version of CGHcall to prevent shifts of the baseline level after global normalisation: CGHcall*. We adjusted the normalisation and post-segmentation normalisation for samples in which the CNA burden exceeded 50% of the sample profile, by considering only the signals included in the $[-0.1, 0.1]$ interval (see Section 3.1). We applied the CGHcall* pipeline on the synthetic data and on the HapMap as described in the previous section for CGHcall. Further, we applied CGHcall* on the \log_2 -ratios between tumour and normal matched TCGA HNSCC samples. When running CGHcall and CGHcall* on the TCGA HNSCC data, we set the tumour purity parameter to the consensus measurement of TCGA HNSCC estimations derived by Aran et al. [34]. For samples with missing derived consensus measurement estimations, we used the immunohistochemistry measurements as tumour purity values.

Performance analysis of genomic copy number calling algorithms

For evaluating the performance of the selected algorithms, we collapsed the resulting calls to three states: loss, normal and gain. For CGHcall, CGHcall* and GISTIC the double loss and loss were collapsed to loss, while the gain and amplification were collapsed to gain. For OncoSNP we collapsed the homozygous and the hemizygous deletion states to loss, and all the states that were defined by more than two copies were considered gain. For ASCAT and genoCNA, the probes with less than two copies were defined as lost, while the probes with more than two copies were defined as gained. We calculated the sample-wise confusion matrix, precision, recall and balanced F-score [35] as follows:

$$\text{precision}_c = \frac{TP}{TP + FP} \quad (1)$$

$$\text{recall}_c = \frac{TP}{TP + FN} \quad (2)$$

$$F_c = 2 \cdot \frac{\text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}, \quad (3)$$

where c represented the class: loss, normal or gain. True positives (TP) represent the number of probes that were classified correctly for each class c , while false positives (FP) are the probes classified incorrectly as class c . False negatives (FN) represent the number of probes that belong to class c but were classified as belonging to another class. To test for statistically significant shifts between F-score distributions of the algorithms, we performed non-parametric pairwise comparison Wilcoxon tests [36]. We adjusted the resulting P -values for multiple testing error through Bonferroni correction [37].

Next, we assessed the performance of the CNA calling algorithms on the Affymetrix SNP 6.0 HapMap samples with matched experimentally genomic copy number validated results. Finally, we analysed the results of the CNA calling algorithms on the TCGA HNSCC Affymetrix SNP 6.0 samples in

HNSCC consensus regions with focus on the Cyclin D1 (CCND1) and the cyclin dependent kinase inhibitor 2A (CDKN2A) genes.

Results and discussion

Characterising molecular phenotypes in cancer research requires the accurate identification of DNA copy number changes. Although genomics increasingly deploys genome sequencing, there is still a wealth of cost-effective SNP array data available. Thus, making use of these data is important and requires best possible analysis approaches that, among other features, are able to correct for cancer-specific confounding variables such as tumour purity and a wide range of CNA lengths. To benchmark commonly used CNA calling approaches in the presence of such confounding variables, we developed an evaluation pipeline.

To evaluate the CNA algorithms, tumour samples with known true states are required. Since the true copy number states for real cancer data are unknown and experimental validation on genome-wide level is not feasible (the human genome size is about 3.0×10^9 bp and is affected by CNVs), we assessed the performance of the algorithms using synthetic data mimicking Affymetrix SNP 6.0 array experiments (see Methods for details). To make the samples as similar as possible to the real Affymetrix SNP 6.0 array samples, we simulated data for 1.844.399 markers of genetic variation—number of probes comparable to the one present on an Affymetrix SNP 6.0 array. Subsequently, we evaluated the performance of OncoSNP, ASCAT, GenoCNA, CGHcall and GISTIC at SNP level resolution.

When conducting a benchmarking study, in addition to realistic synthetic data, we need to use an appropriate measure for the performance of copy number calling algorithms. In general, to show how prediction algorithms perform, receiver operating characteristics (ROC) curves are commonly used [38]. However, when the distribution of the classes is imbalanced, as in our case (Figure S1), ROC curves can present an over-optimistic view on how an algorithm performs, while the recall and the precision have been shown to give a more informative view [39, 40]. Since the F-score represents the balance between the precision and the recall of an algorithm, we selected it as an appropriate criteria and used it to evaluate the performance of the copy number algorithms for each class. The F-score allowed us to determine the algorithm that classified correctly genomic regions independently of the CNA type. This is of great importance, since for a putative future use in personalised medicine, classifying correctly regions overlapping oncogenes or tumour suppressor genes may affect the diagnosis and, thus, the treatment of a patient.

We were interested whether the investigated algorithms can classify precisely the LRR and the BAF signals on probe level into three classes: loss, normal and gain. Therefore, we split the multi-class classification problem into three binary classification problems.

An improved algorithm for copy number calling from Affymetrix SNP 6.0 data: CGHcall*

During manual inspection of the CGHcall pipeline we observed that the normalised signals before and after segmentation in the synthetic samples with more 50% non-normal states covering the sample profiles were incorrectly shifted (either to -1, either to 1). This led to defining an incorrect baseline level in these samples and thus, calling the wrong copy number state. Since cases in which more than half of the genotyped probes are in

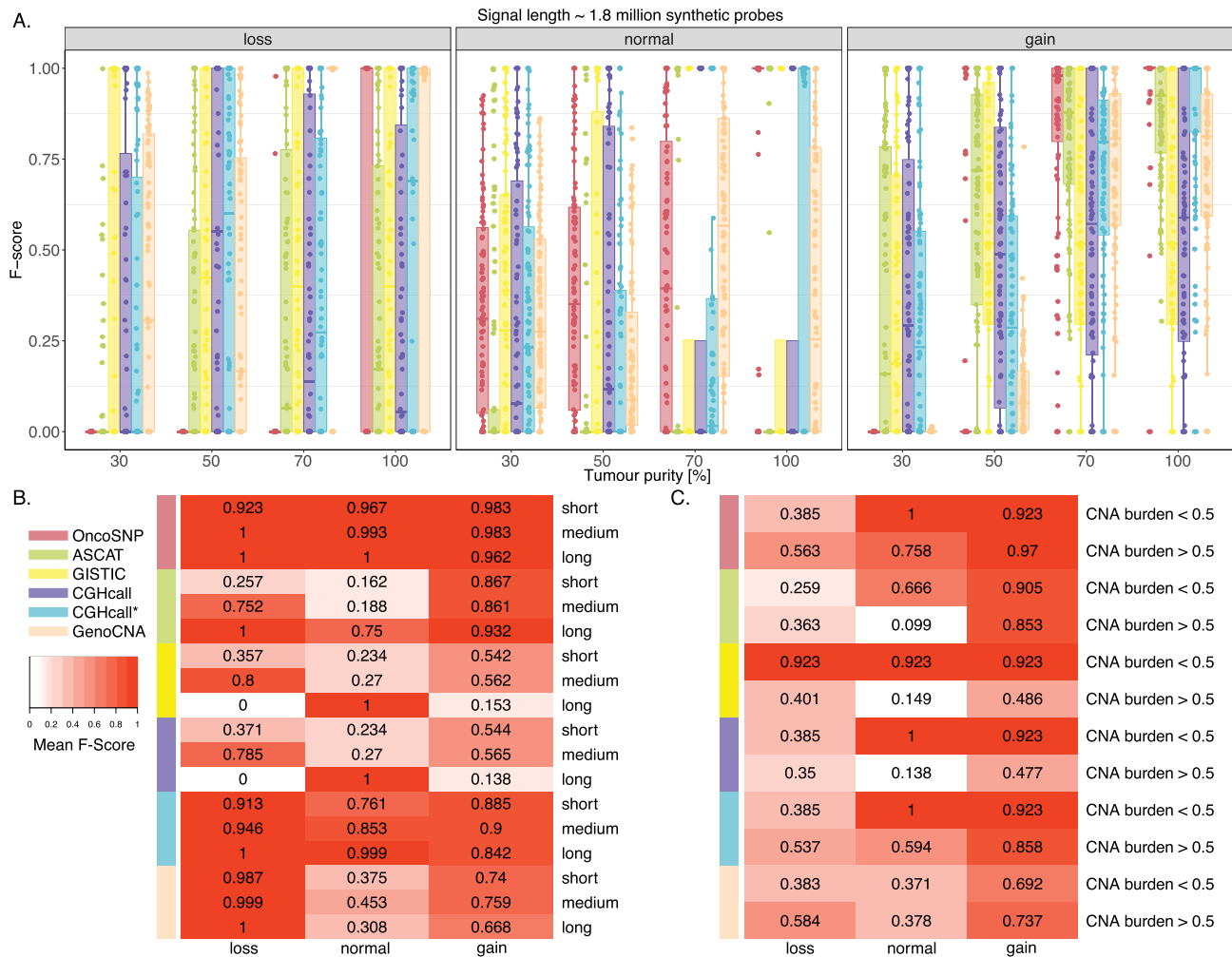


Figure 1. Performance of CNA calling algorithms on synthetic data. We evaluated the performance of six algorithms which are colour-coded as it follows: OncoSNP, coral red; ASCAT, light green; CGHcall, purple; CGHcall*, cyan; GenoCNA, pale pink brown; and GISTIC, yellow. (A). The y-axis represents the F-score and x-axis represents the tumour purity level in %. The three facets represent the different classes: loss, normal and gain. Each boxplot consists of F-scores for 100 synthetic samples. The total number of genetic markers covered by the synthetic signal was approximately 1.8 bp. (B). Heatmap of mean F-scores for different lengths of copy number regions. (C). Heatmap of mean F-scores for samples with CNA burden ratio < 0.5 versus samples with CNA burden ratio > 0.5.

a non-normal state have already been reported in a pan-cancer study on somatic genomic CNAs [14], we set up to correct for the CNA burden effect.

The problem arose from the LRR levels being normalised to the median level over a sample. If more than half of the genome is changed in one direction (loss or gain), CGHcall is unable to correctly estimate the baseline level and assigns the 0 level to what is actually lost or gained. We observed the same behaviour when we applied the post-segmentation normalisation, which assigns the baseline segment to a segment that is either lost or gained. To correct for this effect, we selected three different intervals as constrains for the LRR signals, $[-0.1, 0.1]$, $[-0.05, 0.05]$ and $[-0.2, 0.2]$, and analysed how the performance of the algorithm changes in samples with 100% tumour purity. The resulting F-scores suggested that the LRR signals within the $[-0.1, 0.1]$ interval provided the optimal mean for normalisation and post-segmentation normalisation (Figure S2). As a result, we proposed a solution in which, instead of performing normalisation and post-segmentation normalisation based on all LRR signals, we limit ourselves to LRR signals that fall in the $[-0.1, 0.1]$ interval.

Tumour purity showed strong influence on performance

We first analysed how different tumour purities influenced the performance of the algorithms on synthetic data. We compared the algorithms based on their F-score distributions (Figure 1A). We first showed how the six algorithms (OncoSNP, red; ASCAT, neon green; GISTIC, yellow; CGHcall, purple; CGHcall*, cyan; and GenoCNA, pale orange) identified losses at tumour purity levels (depicted on the x-axis) varying from 30 to 100% (Figure 1A, left panel). OncoSNP was not able to identify losses in samples with tumour purity < 100% (mean F-score = 0.03). ASCAT, GISTIC, CGHcall and CGHcall* showed poor performance when calling losses independent of the tumour purity level (mean ASCAT F-score = 0.26, mean GISTIC F-score = 0.34, mean CGHcall F-score = 0.39, mean CGHcall* F-score = 0.51). GenoCNA showed good performance for correctly calling losses in samples with tumour purities > 50% (mean F-score = 0.68). Thus, the performance of CGHcall* and GenoCNA for calling losses increased with the tumour purity.

OncoSNP showed increasing performance for calling normal states as the tumour purity level increased (Figure 1A, middle panel). This may be caused by the log₂ ratios being pushed towards the 0 baseline in the presence of normal DNA. Moreover, since the normal state represented the majority class, the improved F-score for OncoSNP when calling normal states suggested that the algorithm may not be able to tackle the imbalance of the classes—represented by the copy number states. ASCAT was unable to classify correctly normal states independent of the tumour purity. GISTIC, CGHcall and GenoCNA showed poor performance when trying to classify normal states (mean GISTIC F-score = 0.28, mean CGHcall F-score = 0.29, mean GenoCNA F-score = 0.35). CGHcall* showed overall good performance in correctly finding the normal state when compared to the other three algorithms in samples with tumour purity 100% (mean F-score = 0.70, Figure 1A, middle panel).

Next, we compared how the algorithms performed when trying to identify gains (Figure 1A, right panel). OncoSNP showed good performance when the tumour purity was > 50%. This suggests that OncoSNP is not able to correct the effect of tumour contamination > 50% on the signals in gained genomic regions. The performance of all algorithms for calling gains increased as the tumour purity increased. ASCAT was the only algorithm able to correctly call gains in samples with tumour purities > 30% (mean F-score = 0.76). Overall, our adjusted version of CGHcall—CGHcall* showed improved performance with regard to all copy number states and all tumour purities when compared to CGHcall. GISTIC and CGHcall showed comparable results. This can be explained by the fact that both algorithms use CBS segmentation results and do not make use of the BAF. Our analysis suggested that OncoSNP and CGHcall* handled calling CNAs better than the other algorithms in samples with high tumour purities. The main message of this analysis is that tumour purity strongly influences the results of the CNA calling algorithms. This is an important information to be considered in designing a CNA study, since samples with tumour purities markedly below 50% should not be included in the analysis or at least, profiles resulting from such samples should be handled with care.

The effect of copy number region length

Next, we aimed to understand how the length of a copy number region influenced the performance of the calling algorithms. For this purpose, we examined the difference between the mean F-scores of samples with region lengths of $\leq 10^5$ probes (short), between 10^5 and 10^6 probes (medium) and region lengths $> 10^6$ (long) (Figure 1B). In order to eliminate the effect of reduced tumour purity, we selected only samples with 100% tumour purity. The region length was equal to the number of genetic markers with the same copy number state within a chromosomal segment. One chromosomal segment covered from 3 kilo base pairs (kbp) to 1.8 million base pairs (Mbp).

We observed that OncoSNP, GenoCNA and CGHcall* showed little sensitivity to the length of copy number regions. While CGHcall* and OncoSNP performed well for all three states, GenoCNA had difficulty in correctly identifying normal genomic regions. ASCAT performed worse in samples that included short- and medium-length CNA regions than in samples containing long CNA regions. GISTIC was not able to correctly find lost or amplified genomic regions independent of the length. We observed the same behaviour for CGHcall. One reason that may lay at the core of this problem is the fact that both CGHcall and GISTIC use the CBS algorithm. In all, OncoSNP and CGHcall*

showed consistency and performed well for all three copy number states across the investigated ranges of copy number region lengths.

The effect of CNA burden

Since we observed that the percentage of aberrated regions in a tumour sample—CNA burden—affected the normalisation of the log₂ ratios in the CGHcall pipeline, we investigated whether we observe a similar effect when applying the other copy number calling algorithms.

We therefore grouped the synthetic data into samples with CNA burden > 50% and samples with CNA burden < 50% and calculated the mean F-scores statewise (Figure 1C). We observed that both CGHcall and GISTIC performed poorly for samples with CNA burden > 50%. ASCAT also showed decreased performance for the same scenario, but only for the normal state. The performance of CGHcall* increased in samples with CNA burden > 50% when compared to CGHcall, confirming that we corrected the inaccuracy from CGHcall, especially for predicting normal and gained genomic regions. OncoSNP and CGHcall* were again the best performing algorithms included in this study.

Performance of the copy number calling algorithms on SNP 6.0 array profiles of healthy patients (HapMap)

To assess how OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC perform on real data, we would need a gold standard. Due to the size of human genome – 3.0×10^9 bp, we lack a complete Affymetrix SNP 6.0 array gold standard. Since the HapMap project subsequently experimentally validated the CNAs determined from Affymetrix SNP 6.0 data, we defined the copy number profiles annotated by Redon et al. [4] as our ‘gold standard’. OncoSNP, ASCAT, CGHcall, CGHcall* and genoCNA returned predictions for over 14,500 regions that overlapped the ‘gold standard’. When analysing the F-scores of the algorithms corresponding to 81 profiles with matched annotated copy number profiles (Figure 2), we first observed that OncoSNP, ASCAT, CGHcall, CGHcall* and genoCNA performed well for the normal class (mean F-score = 0.91). Unlike the other algorithms, GISTIC returned predictions for only 381 regions overlapping the ‘gold standard’ and performed poorly for all the classes (mean F-score

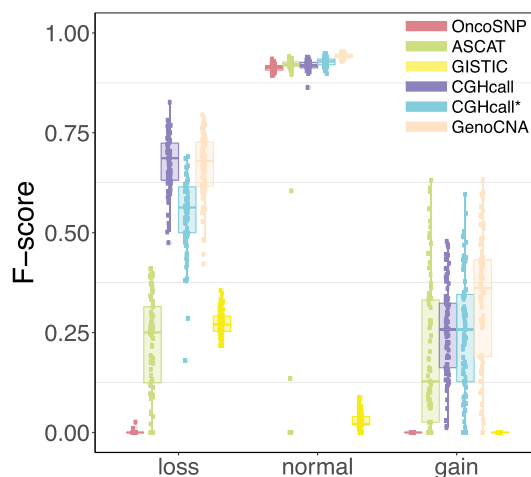


Figure 2. Distribution of F-scores for OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC in 81 healthy HapMap subjects.

= 0.10). OncoSNP could not identify any germline alterations. ASCAT showed a poor performance for identifying gains and losses (mean F-score = 0.20). CGHcall showed a mean F-score of 0.67 for identifying losses, but performed poorly for identifying gains (mean F-score = 0.25). CGHcall* showed a significant improvement only for the normal class compared to the other algorithms. GenoCNA performed best for identifying losses and gains, mean F-score = 0.50. ASCAT, just as OncoSNP and GISTIC, was implemented to find somatic CNAs in cancer samples and was not designed to find germline alterations in the first place. We hypothesise that this might be the reason why OncoSNP, ASCAT and GISTIC perform poorly on healthy patient data.

We are aware that tumour data tailored genomic copy number algorithms are designed to consider CNAs deriving from tumour cell populations. However, HapMap data were generated from blood cells. The genomic copy number changes to be expected from these samples are germline. Therefore, all cells analysed should contain the same alterations. We assume that it would be 'easier' for a tumour data tailored algorithm to pick

up copy number changes. The genomic copy number changes present in the HapMap samples were comprehensively experimentally validated. Thereby, HapMap provides added value since the 'gold standard' with regard to genomic copy number is known for these samples and allowed us to calculate the performance of the CNA calling algorithms on real data. Based on the resulting F-scores, genoCNA, CGHcall and CGHcall* were the best performing algorithms.

CNAs in HNSCC

To test the plausibility of CNA calling results in tumour samples, we explored the concordance between raw LRR signals from TCGA HNSCC samples and the CNA calls of the six algorithms. Additionally, we compared the results with the HNSCC-specific CNA regions defined in Gollin *et al.* [41]. We focused on two genes: one known to be amplified in HNSCC—CCND1 and one that is known to be lost in HNSCC—CDKN2A (Figures 3 and 4).

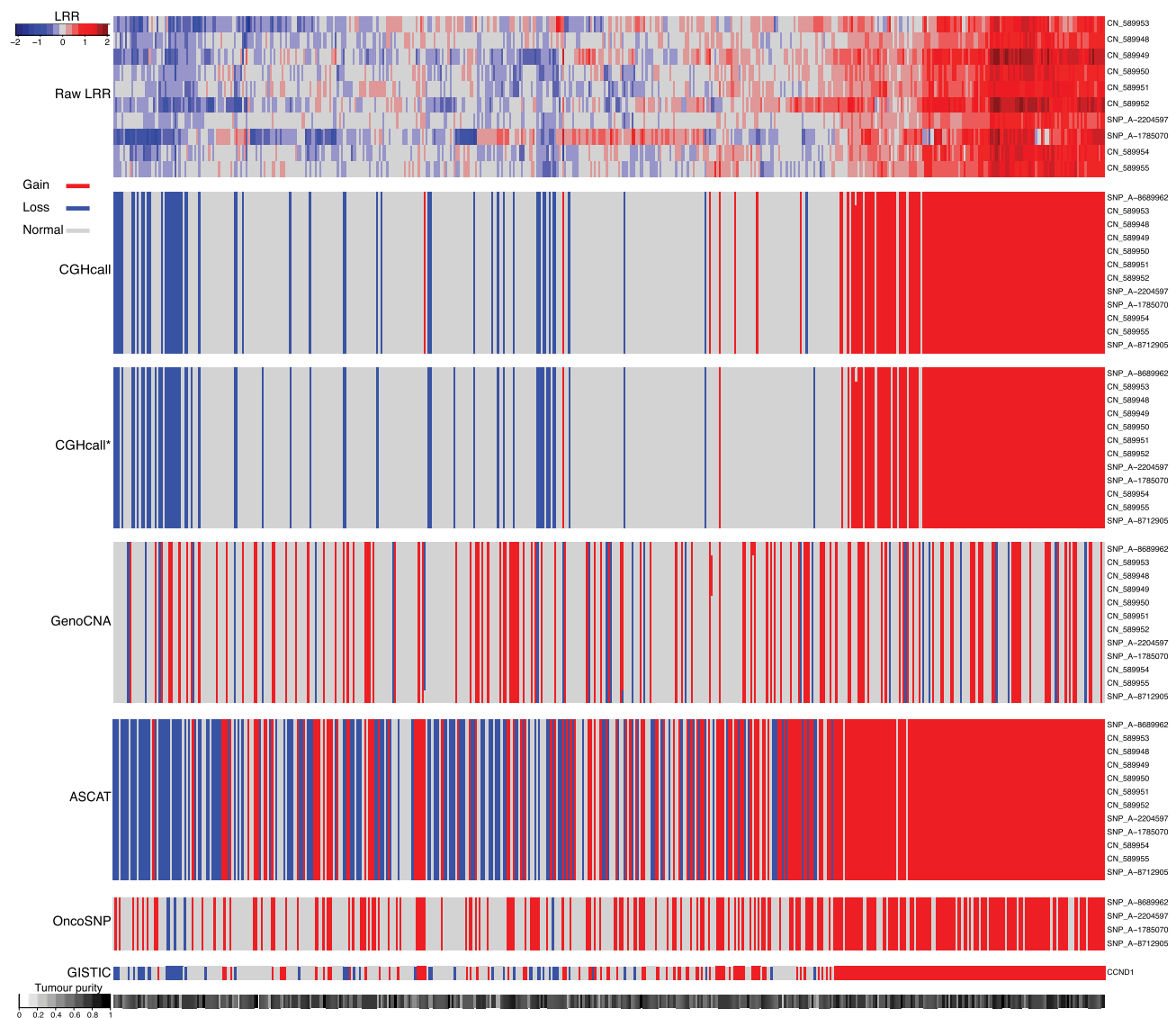


Figure 3. CCND1: Concordance between raw data and algorithm calls in TCGA HNSCC. The heatmap columns represent patients clustered by raw LRR signals. The rows represent the Affymetrix SNP 6.0 probes that overlap the CCND1 region. For CGHcall*, CGHcall, GenoCNA, ASCAT and OncoSNP we also include the neighbouring probe sets of the overlapping region. The lower bar represents the tumour purity of each sample.

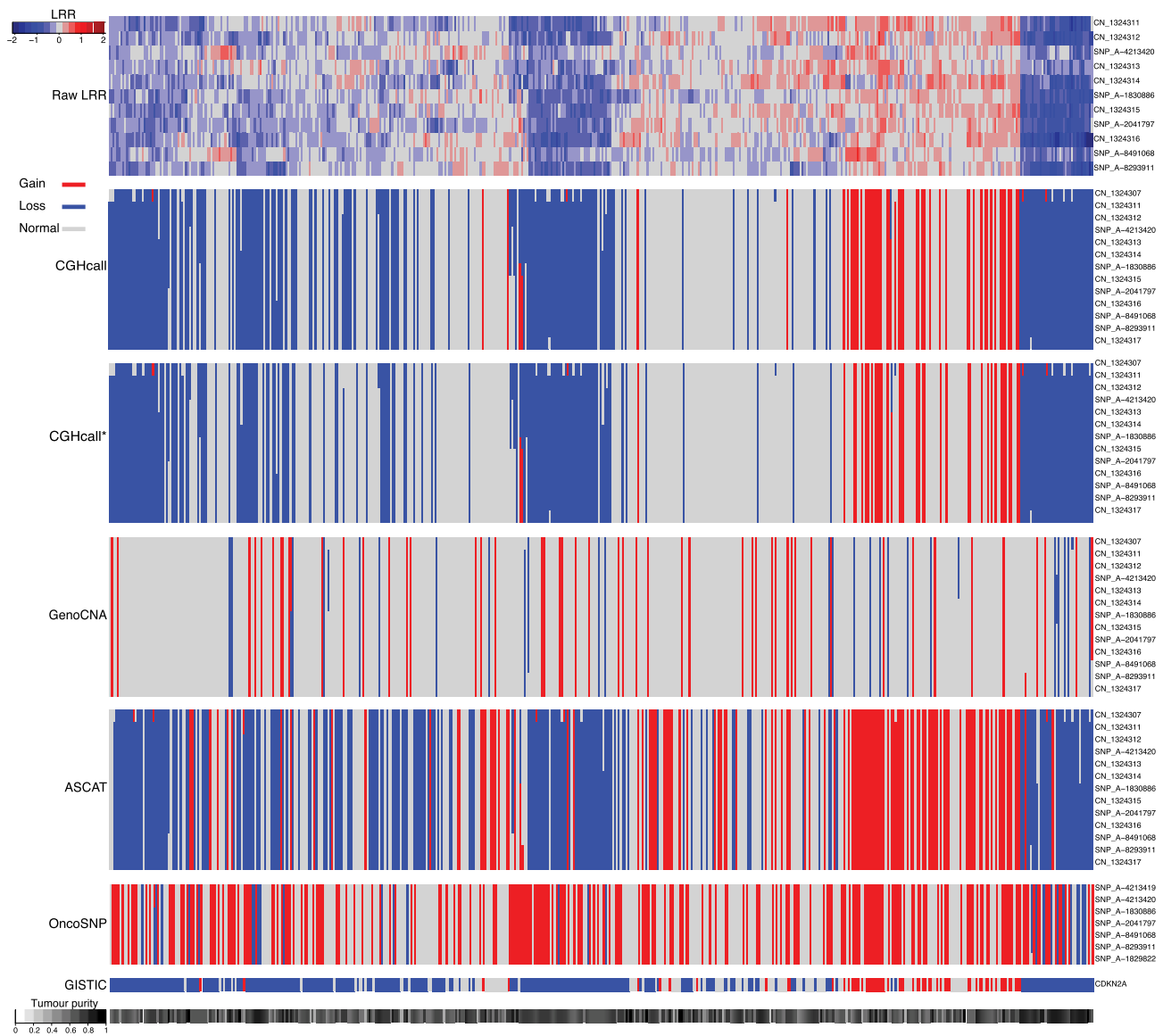


Figure 4. CDKN2A: Concordance between raw data and algorithm calls in TCGA HNSCC. The heatmap columns represent patients clustered by raw LRR signals in the probes overlapping the CDKN2A genomic region. The rows represent the Affymetrix SNP 6.0 probes that overlap the CDKN2A region. For CGHcall, CGHcall*, ASCAT and OncoSNP we also include the neighbouring probe sets of the overlapping region. The lower bar represents the tumour purity of each sample.

The data presented in Figures 3 and 4 show that genomic regions with high LRR values overlapping the CCND1 and CDKN2A genes are called as gained, while genomic regions with low LRR values overlapping the CCND1 and CDKN2A genes are called as lost. The frequencies of CCND1 gains called by CGHcall, CGHcall*, OncoSNP and GISTIC are comparable to the frequencies of CCND1 gains reported from CGH data in Gollin et. al [41], 32%; CGHcall, 26.5%; CGHcall*, 24.9%; OncoSNP, 44%; and GISTIC, 43%. CGHcall, CGHcall*, OncoSNP and GISTIC showed a good overlap in frequencies of CDKN2A losses: CGHcall, 39.8%; CGHcall*, 35.4%; and GISTIC, 59%. The tumour purity ranged from 27.9 to 97.7%. Most of the samples present tumour purity > 60%. These results indicate that in a realistic tumour purity range the algorithms that best performed on synthetic data CGHcall* and OncoSNP showed plausible results in the TCGA HNSCC data as well.

Concluding remarks

Within our study we addressed the problem of evaluating the performance of commonly used copy number calling algorithms in the presence of cancer-specific confounding variables. Since we lacked a complete Affymetrix SNP 6.0 array gold standard, we provided a pipeline to evaluate CNA calling algorithms on Affymetrix SNP 6.0 array-like synthetic data. The analysis on the synthetic data revealed that the performance of the CNA calling algorithms is strongly influenced by tumour purity. CGHcall, GISTIC and ASCAT showed high sensitivity to the length of the genomic segments. The CNA burden strongly influenced the performance of ASCAT, GISTIC and CGHcall. We proposed CGHcall*, an adjusted version of CGHcall, in which we correct for the effect of the CNA burden and we showed that indeed the performance of CGHcall* in samples with a CNA burden higher

than 50%. However, the scope of our paper was to benchmark commonly used CNA calling algorithms, and not to develop a new algorithm.

We further evaluated how the algorithms performed on a real data set comprising of 81 healthy patients HapMap samples that were subsequently experimentally validated. CGHcall and CGHcall* were able to detect germline alterations, unlike OncoSNP and ASCAT. Finally, we examined how comparable were the results of the CNA calling algorithms with the annotated CNAs in *CCND1* and *CDKN2A*, when evaluated on the TCGA HNSCC data set. The results indicated that CGHcall, CGHcall* and GISTIC return comparable calls to what has been reported so far.

In conclusion, we provided a benchmarking pipeline for CNA calling algorithms from Affymetrix SNP 6.0 array tumour profiles together with CGHcall*—an adjusted version of CGHcall for finding CNAs in highly variant genomes.

Key Points

- CNAs are tumour-specific DNA changes that play an important role in cancer research.
- The accurate identification of CNAs is affected by biological confounding variables like tumour purity, the length of a chromosomal segment and the percentage of CNAs present in a genome.
- Within this benchmarking study we provide a pipeline through which we evaluated the performance of six CNA calling algorithms (OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC) in the presence of biological confounding variables.
- We provide an adjusted version of CGHcall—CGHcall* that accounts for a high CNA burden.
- We identify tumour purity and CNA burden to significantly influence the performance of all the CNA calling algorithms.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgments

The authors would like to thank Gökçen Eraslan, Richa Batra, Linda Krause, Michael Strasser and Michael Laimighofer for helpful discussions and feedback.

Funding

German Federal Ministry of Education and Research (BMBF) (02NUK045A).

References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**(7239):719–24.
2. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;**463**(7278):191–6.
3. Beroukheim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature* 2010;**463**(7283):899–905.
4. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature* 2006;**444**(7118):444–54.
5. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet* 2015;**16**(3):172–83.
6. Van Loo P, Nordgard SH, Lingaerde OC, et al. Allele-specific copy number analysis of tumors. *PNAS* 2010;**107**(39):16910–5.
7. Bardeesy N, Cheng K-H, Berger JH, et al. Smad4 is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreas cancer. *Genes Dev* 2006;**20**(22):3130–46.
8. Witkiewicz AK, McMillan EA, Balaji U, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* 2015;**6**:6744.
9. Leucci E, Vendramin R, Spinazzi M, et al. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 2016;**531**(7595):518–22.
10. Wells MF, Wimmer RD, Schmitt LI, et al. Thalamic reticular impairment underlies attention deficit in *Ptchd1*(y/-)mice. *Nature* 2016;**532**(7597):58–63.
11. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotech* 2013;**31**(3):213–9.
12. Y, Zhao L, Wang Y, et al. SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. *BMC Bioinformatics* 2017;**18**:147.
13. Zhang X, Du R, Li S, et al. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics* 2014;**15**:50.
14. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013;**45**(10):1134–40.
15. Zhou W, Zhao Z, Wang R, et al. Identification of driver copy number alterations in diverse cancer types and application in drug repositioning. *Mol Oncol* 2017;**11**(10):1459–74.
16. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotech* 2012;**30**(5):413–21.
17. Cai TT, Jeng XJ, Li H. Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *J Roy Stat Soc Ser B Stat Methodol* 2012;**74**(5):773–97.
18. Sun W, Wright FA, Tang Z, et al. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res* 2009;**37**(16):5365–77.
19. Yau C, Mouradov D, Jorissen RN, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol* 2010;**11**:R92.
20. van de Wiel MA, Kim KI, Vosse SJ, et al. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* 2007;**23**(7):892–4.
21. Mermel CH, Schumacher SE, Hill B, et al. GISTIC 2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**(4):R41.
22. Metzger J, Philipp U, Lopes MS, et al. Analysis of copy number variants by three detection algorithms and their association with body size in horses. *BMC Genomics* 2013;**14**:487.

23. Mosén-Ansorena D, Aransay A, Rodríguez-Ezpeleta N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinformatics* 2012;13:192.
24. Pierre-Jean M, Rigail G, Neuvial P. Performance evaluation of DNA copy number segmentation methods. *Brief Bioinform* 2015;16(4):600–15.
25. Hieronymus H, Schultz N, Gopalan A, et al. Copy number alteration burden predicts prostate cancer relapse. *PNAS* 2014;111(30):11139–44.
26. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;517(7536):576–82.
27. Lin C-F, Naj AC, Wang L-S. Analyzing copy number variation using SNP array data: protocols for calling CNV and association test. *Curr Protoc Hum Genet* 2013;79:Unit–1.27.
28. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;12(5):363–76.
29. Rasmussen M, Sundström M, Kultima HG, et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* 2011;12:R108.
30. Lockstone HE. Exon array data analysis using Affymetrix power tools and R statistical software. *Brief Bioinform* 2011;12(6):634–44.
31. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007;17(11):1665–74.
32. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008;40(10):1253–60.
33. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat* 2004;5(4):557–72.
34. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;6:8971.
35. Van Rijsbergen CJ. *Information Retrieval*, 2nd edn. Newton, MA, USA: Butterworth-Heinemann, 1979.
36. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945;1(6):80–3.
37. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni R Istituto Superiore Scienze Economiche Commerciali Firenze* 1936;8:3–62.
38. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006;27(8):861–74.
39. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*. pp. 233–40. ACM, New York, NY, USA, 2006.
40. Lever J, Krzywinski M, Altman N. Points of significance: classification evaluation. *Nat Meth* 2016;13(8):603–4.
41. Gollin SM. Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: a next generation window to the biology of disease. *Genes Chromosomes Cancer* 2014;53(12):972–90.