

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 ($1 - \text{NPV} = 7/23 = 0.304$, 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*	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, **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

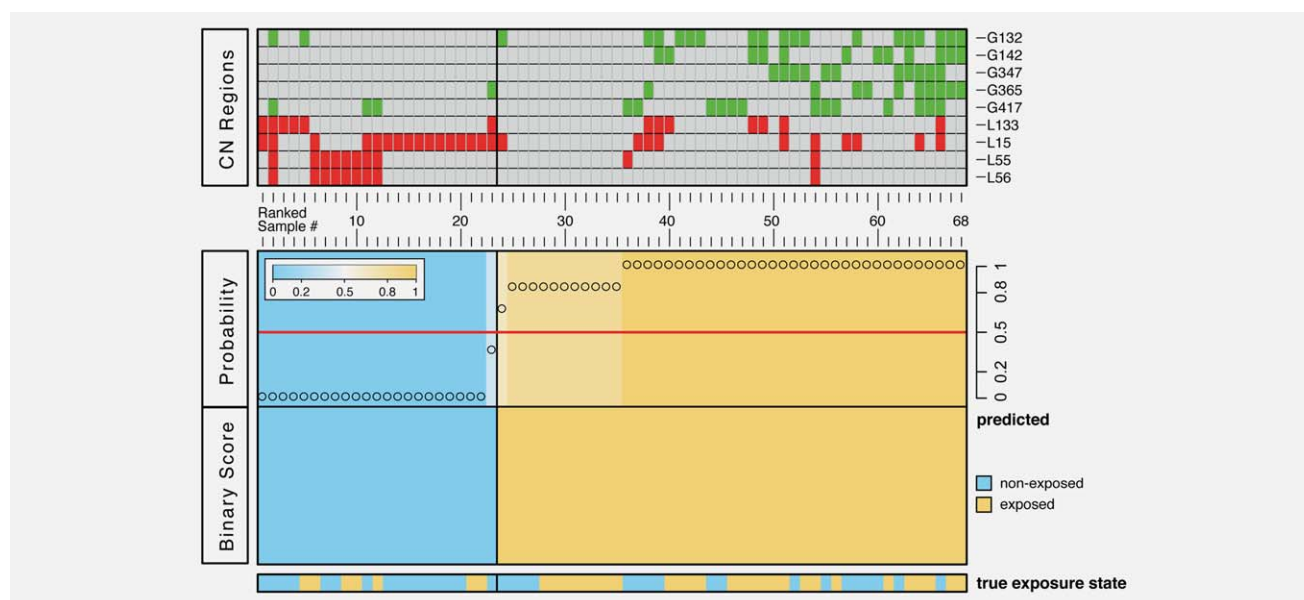


Figure 1. Heatmap of the 9-CNA-signature of 68 breast cancer patients of the validation set composed of 34 exposed and 34 non-exposed cases. Copy number gains are represented by green color, losses by red color (top panel). The middle panel shows the risk score on the probability scale calculated according to the formular 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.

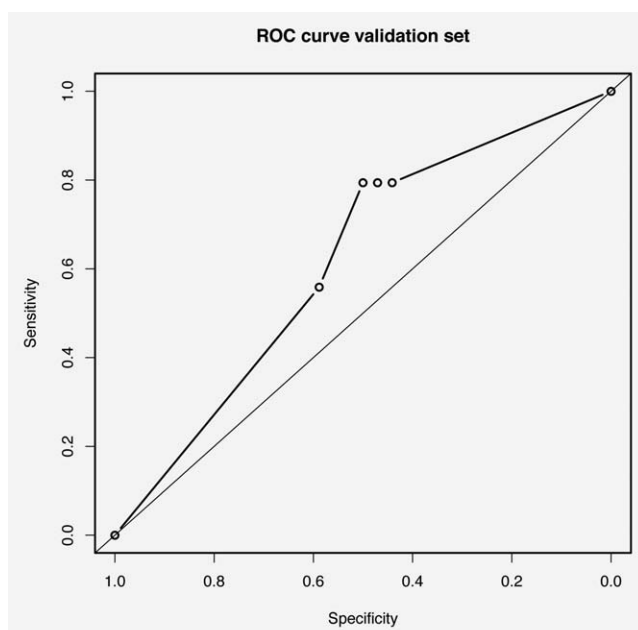


Figure 2. ROC curve calculated by applying a logistic regression model fitted on the training set and evaluated on the validation set. Each point (circles) corresponds to a probability cutoff level decreasing from left to right, given by the steps visualized in Figure 1. Points are connected by straight 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 heterogenous conditions in contrast to the rather homogenous 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, TTLL4, 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.*^{29,30} 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.
- Prisyazhnyuk AY, 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, Cyrillic 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 radioresistance 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.