Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

Next Generation Sequencing of Serum Circulating Nucleic Acids from Patients with Invasive Ductal Breast Cancer Reveals Differences to Healthy and Nonmalignant Controls

Julia Beck[1], Howard B. Urnovitz[1], William M. Mitchell[2], and Ekkehard Schütz[1]

Abstract

Circulating nucleic acids (CNA) isolated from serum or plasma are increasingly recognized as biomarkers for cancers. Recently developed next generation sequencing provides high numbers of DNA sequences to detect the trace amounts of unique serum biomarkers associated with breast carcinoma. Serum CNA of 38 women with ductal carcinoma was extracted and sequenced on a 454/Roche high-throughput GS-FLX platform and compared with healthy controls and patients with other medical conditions. Repetitive elements present in CNA were detected and classified, and each repetitive element was normalized based on total sequence count or repeat count. Multivariate regression models were calculated using an information-theoretical approach and multimodel inference. A total of 423,150 and 953,545 sequences for the cancer patients and controls, respectively, were obtained. Data from 26 patients with stages II to IV tumors and from 67 apparently healthy female controls were used as the training data set. Using a bootstrap method to avoid sampling bias, a five-parameter model was developed. When this model was applied to a validation data set consisting of patients with tumor stage I (n = 10) compared with healthy and nonmalignant disease controls (n = 87; 1,261,561 sequences) a sensitivity of 70% at a specificity of 100% was obtained. At a diagnostic specificity level of 95%, a sensitivity of 90% was calculated. Identification of specific breast cancer–related CNA sequences provides the basis for the development of a serum-based routine laboratory test for breast cancer screening and monitoring. *Mol Cancer Res; 8(3); 335–42. ©2010 AACR.*

Introduction

Breast cancer is one of the most frequent malignancies among women and accounts for one in four cancer occurrences in adult women (1). Screening methods and adjuvant therapy following surgery led to a remarkable decrease in mortality between 1975 and 2000 (2). Nevertheless, ~185,000 new cases and 41,000 deaths from breast cancer are estimated to have occurred in 2008 (3). In addition to physical examination, detection of breast cancer has relied on imaging technologies for screening that have reached the apparent limits of their capacity and/or have low societal cost-benefit ratios (i.e., magnetic resonance imaging). The prime example is the decreased sensitivity of standard mammography in young women and women with dense parenchyma (4). Results from seven population-based community screening programs in the United States revealed an overall sensitivity of 75% and a specificity of 92.3%, although sensitivity in women with extremely dense breasts dropped to 63% (5). This is further influenced by a decrease in the use of mammography from 2000 to 2005 (6), with pain caused by breast compression during the procedure cited as a common complaint (7, 8). Development of minimally invasive molecular techniques to supplement imaging procedures for the early diagnosis and treatment of breast cancer and the consequent reduced mortality has remained elusive despite the recent remarkable advances in molecular biology.

The risk for breast cancer in humans has been associated with a limited number of gene mutations and gene expression signatures. High-risk single-gene alleles of *BRCA1, BRCA2, PTEN*, and *TP53* are relatively rare in the general population and account for less than 25% of the familial clustering observed in breast cancer (9). Other reported breast cancer–associated single nucleotide polymorphisms include *FGFR2, TNRC9, MAP3K1, LSP1*, and *CASP8*, which in combination might distinguish women at high risk for the development of breast cancer (10). Tissue samples of established breast carcinomas can be evaluated for metastatic potential by gene expression signatures (11). Unfortunately, none of these

Authors' Affiliations: [1]Chronix Biomedical GmbH, Goettingen, Germany and [2]Department of Pathology, Vanderbilt University, Nashville, Tennessee

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

Corresponding Author: Ekkehard Schütz, Chronix Biomedical GmbH, Goetheallee 8, 37073 Goettingen, Germany. Phone: 49-551-3707-5722; Fax: 49-551-3913-968. E-mail: esc@chronixbiomedical.de

doi: 10.1158/1541-7786.MCR-09-0314

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

known genetic risk associations is of sufficient diagnostic value for clinical use.

Cell-free circulating nucleic acids (CNA) isolated from blood plasma or serum have become increasingly recognized as potential biomarkers for the early detection and clinical monitoring of various human cancers. The first report that DNA could be detected in the blood of cancer patients occurred in 1948 (12), although efforts to explore this observation for diagnostic purposes did not occur until the last decade. Measurements of absolute levels of CNAs have been suggested for the diagnosis (13) and prognosis (14) of breast and lung cancer (15). The general diagnostic value of simple quantitative measures of CNA, however, is controversial because nonspecific CNA elevations are seen in patients with benign diseases (16–19). Cancer-specific DNA perturbations such as microsatellite instability, mutations, sequence length, and promoter methylation patterns detected in serum/plasma have been proposed for the diagnosis and clinical assessment of cancer treatment (20).

In this report, we show that the quantities of specific cell-free transposable elements and endogenous retroviral DNA sequences in blood could distinguish all stages of invasive ductal breast cancer from normal controls and nonmalignant controls with sensitivities and specificities that are promising as a useful clinical tool.

Materials and Methods

Specimen

Serum samples were obtained from 38 women with breast cancer and 67 healthy female donors. All patients were histologically diagnosed with infiltrating ductal carcinoma. The group was comprised of 10, 20, 2, and 6 patients with clinical stages of I, II (including IIA and IIB), III, and IV, respectively. Tumor staging and histology was used as given in the patient's records. Blood was drawn preoperatively from treatment-naïve patients. Thirty-two of the patient samples were collected at the Ryazan Central Oblast Hospital, Russia. In the same hospital, an additional 39 apparently healthy control samples were collected. Six breast cancer samples were collected at the Cleveland Clinic satellite facility in Florida, USA. A detailed patient description is given in Supplementary Table S1. Sequence data from 28 apparently healthy female controls were previously generated and stored in a CNA database that was described recently (21). Additional 87 collected samples of patients and controls were analyzed as an independent noncancerous validation set and 31 samples collected from a patient with multiple myeloma and liver sarcoma served as a malignant control (Table 1). All donors and patients provided their informed consent.

Sampling

Serum samples were collected as described (22) and stored at −80°C prior to further processing. Briefly, frozen serum was centrifuged at 4,000 × *g* for 20 min and 200 μL of the supernatant was used in the High Pure Nucleic Acids Extraction Kit (Roche) according to the instructions of the manufacturer.

Generation of Random DNA Libraries, Sequencing, and Sequence Analysis

The serum DNA was processed, sequenced, and analyzed as described previously (21). Briefly, extracted DNA was amplified using the GenomePlex Single Cell Whole Genome Amplification Kit (WGA4, Sigma-Aldrich) in duplicates, which were pooled for subsequent sequencing. Although probably not necessary for CNA, the shearing procedure as included in the WGA4 protocol was done for comparability with, e.g., genomic results. Previously, we have tested the WGA4 procedure using either unsheared or sheared genomic DNA and could not detect essential differences in sequencing results (21). Using the adapters that are attached during the whole genome amplification procedure, a sample-specific molecular barcode was attached to the amplicons by PCR. The resulting DNA preparations were pooled and sequenced using a Roche/454 GS-FLX high-throughput sequencer. After sequencing, the sequences were assigned to the individual patients and control samples according to the 10-bp tag sequence. An overview of the further sequence analysis pipeline is depicted in Supplementary Fig. S1.

Statistical Procedure

Normalization for each sample and each repetitive element (per assigned nucleotide) was done either on the basis of the total assigned nucleotides of human genomic origin or of human repetitive element origin (Supplementary Fig. S1). The normalized values were used to construct receiver operator characteristics (ROC) curves for each repetitive element present in at least 95% of the serum samples. The C value was calculated as area under the receiver operating characteristic (ROC) curve by applying the linear trapezoidal rule (23). Repetitive elements with a C value of at least 0.62 were considered as independent variables for the subsequent calculation of binominal multivariate regression models. The general linear model was fitted using a training set comprising normal female controls (n = 67) and patients with tumor stages II to IV (n = 28). An information-theoretical model selection with subsequent multimodel inference approach was taken (24). First, the corrected Akaike's information criterion values were calculated for all possible models of two to five predictors (e.g., 174,500 when 30 independent variables were considered). Akaike's information criterion values were calculated, according to Burnham and Anderson (25), as transformation of the residual sum of squares of the linear multivariate regressions. These values were used to calculate normalized Akaike weights (ω) in such a way that they sum up to 1 for all models as described (25).

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

Table 1. Sample sets used for model development and validation

Training/model development

Name of sample set	*n*	Gender (female/male)	Description
Normal female training set	67	67/0	Samples were obtained from apparently healthy women
Stage II to IV training set	28	28/0	Contains 20 samples of tumor stage II, 2 samples of tumor stage III, and 6 samples of tumor stage IV

Validation

Name of sample set	*n*	Gender (female/male)	Description
Normal female training set	67	67/0	
Breast cancer validation set	10	10/0	Contains only patients with tumors in stage I, which were not included in the training set
Noncancerous validation set	87	34/53	Combination of independent healthy controls and patients with other noncancerous medical conditions. Other medical conditions include multiple sclerosis (n = 34), diabetes mellitus (n = 1), and autism (n = 1)
Multiple myeloma	31	0/31	Samples were obtained from one male patient at different time points

Those models accounting for the lower 97.5% of the 5% trimmed ω sum were discarded. For each predictor (independent variable) Akaike ω of all remaining models containing that predictor were summed ($\Sigma\omega$). The six variables with the highest Akaike $\Sigma\omega$ were used to constitute the final model. For these six variables, the respective estimates were averaged (25). In an additional approach, to further reduce the chance of random group effects on modeling, a random resampling strategy was applied. On average, each sample was resampled into 172 bootstrap rounds (SD, 9), in which 46 normal controls and 19 stage II to IV patient samples were used in each round. In total, 255 rounds were done that yielded 37.6×10^6 models, of which for each calculation the best (highest ω) 1%, but not less than 40 models, were selected and the predictor estimates were recorded as given above. This resulted in ~1.1×10^5 models, from which the sum of $\Sigma\omega$ per independent variable was calculated. Again, the six predictors with the highest weight sums were used for model definition, in which average estimates of independent variables were calculated as above.

For each study sample, the calculated intercept and the predictor estimates (slope) were used to calculate the model value of the samples. To assess the validity of the achieved models, samples originating from patients with tumor stage I (n = 10) were used as a validation set, together with samples from individuals with other and nonmalignant medical conditions. The performance of the obtained models was assessed by the generation of ROC curves from those calculated model values. Sensitivity and specificity were derived at the point of best separation, which was defined as the maximized Youden's index value (26).

Results

Sequencing Results

A total of 423,150 sequences for cancer patients and 953,545 sequences for apparently healthy female controls were obtained, totaling 7.5×10^7 and 1.61×10^8 nucleotides, respectively. Together with the additional controls, a total of 2.8×10^6 sequences were analyzed herein. Supplementary Table S2 gives a summary of the average fragment and nucleotide counts in each group and the average number of nucleotides of human genomic origin. Nucleotides of human origin are defined as nucleotides that were either identified as human repetitive elements by RepeatMasker (27, 28) or produced significant hits after BLAST (29) searching of the human genome.

Representation of Repetitive Elements in the CNA

Generally, the serum CNA pool of patients with breast cancer contains significantly more repetitive elements (52%; SD, 0.8%) than the serum CNA of normal control subjects (51%; SD, 1%). Although small, the difference is significant (P = 0.0013; median test) when all patients with invasive carcinomas in stage II to IV were compared with the healthy female control cohort.

Two different normalization approaches were taken to compare repetitive element representation. The normalization on total repetitive elements instead of total genomic

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

Table 2. Model-averaged regression coefficients for the six predictors in $model_{normRE}$, $model_{normGO}$, and $model_{RRnormRE}$

$Model_{normGO}$

Predictor	Σω*	Estimate	C value†
Total RE content	4.98	16	0.73 (0.51-0.87)
L1PA12	3.36	–199	0.70 (0.49-0.85)
DNA/MER1-type	3.27	86	0.70 (0.49-0.85)
AT-rich	2.90	170	0.74 (0.54-0.87)
MLT2A2	2.09	134	0.66 (0.45-0.82)
MIRb	1.89	44	0.69 (0.47-0.84)

$Model_{normRE}$

Predictor	Σω*	Estimate	C value
L1PA12	2.00	–192	0.70 (0.50-0.85)
AT-rich	2.00	164	0.73 (0.53-0.87)
MER2B	1.30	–232	0.64 (0.43-0.80)
L1P4a	1.10	–130	0.67 (0.45-0.83)
FLAM-C	0.90	89	0.63 (0.42-0.79)
L1MC2	0.40	–28	0.67 (0.45-0.83)

$Model_{RRnormRE}$

Predictor	Σω‡	Estimate	C value
AT-rich	2.97	107	0.73 (0.53-0.87)
L1PA12	1.95	–80	0.70 (0.50-0.85)
MER2B	1.17	–134	0.64 (0.43-0.80)
L1P4a	0.92	–133	0.67 (0.45-0.83)
L1MC2	0.71	–72	0.67 (0.45-0.83)
MLT2A2	0.57	38	0.66 (0.45-0.81)

*Σω shows the sum of Akaike weights for each predictor across all models that contain the predictor (%). This sum reflects the relative importance of the respective predictor and was used as selection criteria.
†C value gives the area under the ROC curve as calculated for each single predictor (95% confidence intervals).
‡Total sum in resampled model. The estimates are defined as slopes of each predictor.

hit length identifies which repetitive elements are overrepresented or underrepresented within the group of repetitive elements, in which AT-rich sequences show the largest deviation in both normalizations. AT-rich sequences were found in higher amounts in patients with breast cancer yielding z values of 1.04 and 0.96 when normalized to total genomic hit length and repetitive elements hit length, respectively. A ROC curve was calculated for each repetitive element. A ROC curve plots the true positive rate against the false positive rate for the different diagnostic cutpoints and the area under the curve (C value) is a measure of test accuracy. Supplementary Table S3 summarizes the ROC curve data of those repetitive elements that yielded C values of >0.62 for the two different normalizations. The data shows that there is no single repetitive element that can be used alone to obtain good separation between patients with breast cancer and controls.

Binominal Multivariate Regression Modeling

The repetitive elements listed in Supplementary Table S3 were subsequently used for developing binominal multivariate models. Models were calculated for normalization by the number of repetitive nucleotides ($model_{normRE}$) and the total hit length of genomic origin ($model_{normGO}$).

The models were fitted using data from the stages II to IV training set versus the normal female training set (Table 1). Table 2 gives the predictors, weight sums, and respective averaged estimators for both models, together with the C statistic from ROC curves.

The stage I patient data was subsequently used as an independent data set (breast cancer validation set) to validate the model performance. Table 3 provides the diagnostic performance characteristics of $model_{normRE}$ and $model_{normGO}$ in the different patient groups. The $model_{normRE}$ yielded a higher C value than $model_{normGO}$ when applied to the breast cancer validation set (Table 3). Further validation of the model predictors employed a random resampling approach ($model_{RRnormRE}$), generated from the initial training data set. As illustrated in Table 2, the predictors of $model_{normRE}$ with the exception of FLAM-C were validated by the $model_{RRnormRE}$. FLAM-C was replaced by MLT2A2, which had a higher Σω in the random resampling modeling. Following the same approach as conducted for the first two models, the $model_{RRnormRE}$ was applied to the breast cancer validation set (stage I) and to the combined data set (stage I-IV) using ROC curve analysis. The performance characteristics for the three different groupings of the cancer samples compared with the normal female training set, are given in Table 3. Supplementary Fig. S2 shows the calculated model values and the ROC curves for this model. Although the C value of $model_{RRnormRE}$ (0.9; confidence interval, 0.76-0.97) in the stage II to IV training set was lower than that of $model_{normGO}$ (0.94; confidence interval, 0.8-0.99), the former yielded a slightly higher C value when the breast cancer validation set was compared with the normal female training set.

Model Validation Using an Independent Control Group

The $model_{RRnormRE}$ was analyzed additionally in a cohort of 87 individuals that had noncancerous diseases (1.26×10^6 sequences with 2.1×10^8 nucleotides) and in a multievent time course of a patient suffering from multiple myeloma ($n = 31$; 320,342 sequences with 5.8×10^7 nucleotides). A detailed description of the samples is given in Table 1. The distribution of the obtained scores are shown in Fig. 1. The two different breast cancer patient groups are displayed for comparison. It is obvious that neither the multiple myeloma nor the independent nonmalignant control set is associated with serum DNA sequences

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

similar to invasive ductal carcinoma. From ROC curve analysis using the noncancerous validation set versus the breast cancer validation set, a C value of 0.986 (0.927-0.998), with a specificity of 95% and a sensitivity of 90%, was calculated. At 100% specificity, a corresponding sensitivity of 70% was obtained.

Discussion

High-throughput sequencing of total serum DNA shows differential representation of certain repetitive elements in the CNA of patients with breast cancer compared with healthy controls. The comprehensive database generated allowed the simultaneous comparison of all known repetitive elements of patients with breast cancer versus normal and nonmalignant controls. The normalized nucleotide counts of repetitive elements were evaluated and those giving the best separation were included in modeling. Here, we chose model selection using an information-theoretical approach with multimodel inference, which is a relatively new concept in biological sciences providing a more robust alternative compared with traditional approaches of hypotheses testing (24, 25). As expected, due to the large number of possible predictive variables obtained with shotgun sequencing of the CNA pool, the data supported more than one possible model. In such cases, model averaging provides robust inference that is not conditional on a single "best" model. In addition, a random resampling strategy of 255 rounds was applied, which confirmed five of the six initial predictors for the final $model_{normRE}$. From the initial $model_{normRE}$, only the free left Alu monomer (FLAM-C; ref. 30) was replaced by MLT2A2, an endogenous retrovirus-related element. This suggests that the Alu monomer FLAM-C is a weaker predictor for breast ductal carcinoma than the LINE elements and AT-rich sequences. This conclusion is strengthened by the fact that MLT2A2 is included in both the $model_{RRnormRE}$ and the $model_{normGO}$.

The final models were tested for their performance with independent data obtained from patients with stage I invasive ductal carcinomas. Interestingly, both $model_{normRE}$ and $model_{RRmodelRE}$ yielded higher area under the ROC curve values in the stage I patient cohort with the latter providing the highest C value when applied to the data obtained from stage I cancers. We used the C values obtained for the validation data set as a general measure for the goodness of a model, although the obtained differences were not statistically significant, as shown by the overlapping confidence intervals. Further studies are needed to verify this trend. Nevertheless,

Table 3. Diagnostic performance of $model_{normRE}$, $model_{normGO}$, and $model_{RRnormRE}$ in different patient groups versus the normal female training set

$Model_{normGO}$

Patient group	C value*	Specificity†	Sensitivity	Sensitivity‡ (Spec. 97.5)	Maximum accuracy
Stage II to IV	0.94 (0.80-0.99)	0.90	0.93	0.61	0.91
Stage I	0.90 (0.62-0.97)	0.94	0.80	0.50	0.92
Stage I to IV	0.93 (0.80-0.98)	0.88	0.89	0.58	0.89

$Model_{normRE}$

Patient group	C value	Specificity	Sensitivity	Sensitivity‡ (Spec. 97.5)	Maximum accuracy
Stage II to IV	0.89 (0.74-0.96)	0.92	0.79	0.57	0.88
Stage I	0.94 (0.76-0.99)	0.92	0.90	0.70	0.95
Stage I to IV	0.91 (0.78-0.97)	0.92	0.82	0.50	0.89

$Model_{RRnormRE}$

Patient group	C value	Specificity	Sensitivity	Sensitivity (Spec. 97.5)	Maximum accuracy
Stage II to IV	0.90 (0.76-0.97)	0.82	0.82	0.57	0.84
Stage I	0.97 (0.89-1.00)	0.87	1.00	0.70	0.95
Stage I to IV	0.92 (0.81-0.97)	0.82	0.87	0.50	0.85

NOTE: Model performance was assessed by calculating the area under the ROC curve.
*C value (area under the ROC curve with 95% confidence intervals in brackets).
†Specificity and sensitivity were calculated at the point of best separation.
‡Sensitivity at a specificity of 97.5%.

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

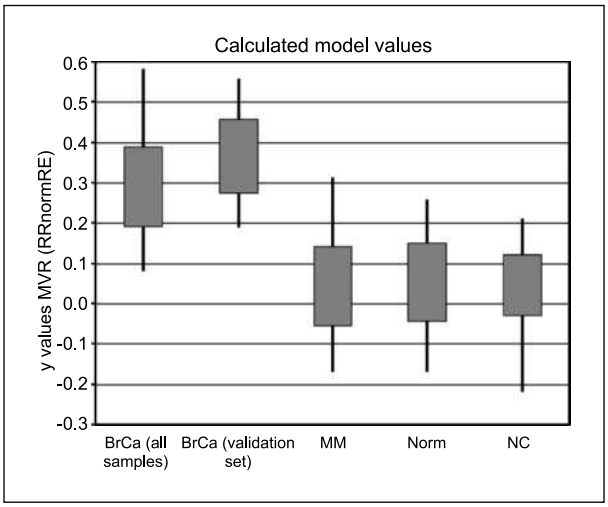

FIGURE 1. Distribution of scores calculated using $model_{RRnormRE}$ on data obtained from different patient groups. The 75th and 25th percentile are plotted as boxes, whiskers display the 97.5th and 2.5th percentiles. BrCa (all), all patients with breast cancer regardless of tumor staging; BrCa (validation set), 10 patients with tumors in stage I; NORM, normal group as used in training sets; MS, multiple sclerosis patients; NC, multiple sclerosis and normal data combined as noncancerous group. For a detailed description of samples, see Table 1.

$model_{RRnormRE}$ is especially promising for detecting stage I invasive ductal carcinoma with high sensitivity. A previously published CNA-based breast cancer analysis, which detected differences in the integrity of circulating Alu elements, showed no clear discrimination of breast cancer in stage I disease (31). Screening tests for the early detection of cancer is universally accepted to reduce cancer mortality. Nevertheless, careful optimization is required because an inordinately high false-positive rate will cause unnecessary diagnostic work-ups (32). With a sensitivity of 70% at a specificity of 100% in independent control and patient data sets, the $model_{RRnormRE}$ seems promising as a basis for the development of future screening tools. Adjustment of the specificity cutoff to 95% provides a high predicted sensitivity of 90%. The results obtained at different time points from a multiple myeloma patient that developed a second malignant neoplasm suggest that this model may not be affected by cancers in general.

The origin of serum DNA sequences is controversial. Generally, cell-free DNA in the blood is associated with histones and it is assumed that these circulating nucleosomes originate from apoptotic or necrotic cells in the body, and that the protein-bound DNA is protected from digestion by nucleases (19). Our sequencing approach revealed differences in the relative amounts of different repetitive elements that might reflect a differential nucleosome positioning in the genome of the cancerous cells. Epigenetic modifications such as methylation of DNA and acetylation or methylation of histones leading to changes in the chromatin order are known to be relevant to cancer (33). Recently, sequence-specific histone methylation was detected in the plasma of patients with multiple myeloma (34). All models presented herein contain several LINE1 elements that were lower in patients with breast cancer than in normal controls. Alterations in other LINE1 elements in serum have already been associated with breast cancer (35). Interestingly, extensive hypomethylation of LINE elements and endogenous retroviruses have been found in several cancers (36), and hypomethylation does have an effect on nucleosome positioning (33). Furthermore, LINE elements in humans are predominantly found in chromosomal G bands also known as facultative heterochromatin (37, 38). Interestingly, there is evidence of nonrandom degradation of DNA in leukemic cells in which apoptotic DNA preferentially hybridizes to heterochromatin (39). We found AT-rich sequences being a strong positive predictor in all of the calculated models. Recently, it was reported that SATB1, the special AT-rich binding protein, is highly expressed in aggressive breast cancer cells (40). SATB1 belongs to a family of nuclear matrix binding proteins, which constitute and maintain the DNA nuclear architecture, thereby regulating cell-specific gene expression. SATB1 is normally cleaved by caspases during the early stages of apoptosis (41). An overexpression of SATB1 in breast cancer cells could, therefore, be one possible reason for the relative overrepresentation of AT-rich sequences in the breast cancer–specific circulating DNA pool. Although speculative, it seems possible that apoptotic DNA degradation in cancer cells differs from nonmalignant cells and that these differences could be detected by sequencing of apoptotic DNA remnants in the serum.

Alternatively, differential repetitive element representation in the CNA pool may originate from different processes by which DNA is released from cancer and normal cells. Several forms of cell death other than apoptosis (e.g., necrosis, autophagy, or mitotic catastrophe) as well as an active release of newly synthesized DNA have been discussed as possible sources of cancer-specific CNA (31, 42). The detection of MLT2A2 as a strong predictor in two of our models also suggests that the CNA pool might also contain retrovirus-like particles that reflect distinct retroviral activity of endogenous retroviruses in cancer cells. Retroviral particles have been detected in various malignant tumors (43) and in the T47-D human breast cancer cell line (44).

Based on recent reports, a bias introduced by random amplification methods including the WGA4 cannot be ruled out (21, 45), in which the latter was shown to introduce the least bias (45). To avoid a significant influence on the comparative data, samples were prepared strictly the same way, two independent WGA4 reactions were pooled for each sample and breast cancer samples and normal controls were run together in mixed batches to exclude bias introduced by individual sequencing runs.

In conclusion, we report specific DNA-transposable element sequences that could discriminate all stages of invasive ductal breast carcinoma with significant specificity and sensitivity. Although not practical as such for standard

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

screening or prognostic purposes, the detected patterns provide a knowledge base for the development of a serum-based assay. Automated methods such as microarrays or quantitative multiplex PCR, once validated against the sequence data reported here, will allow further investigation in a prospective clinical setting. The data is promising for the development of a cost-effective, serum-based screening assay.

Disclosure of Potential Conflicts of Interest

All authors, ownership interest, Chronix Biomedical.

Acknowledgments

We thank Sara Henneke, Stefan Balzer, and Carsten Müller for their skilful technical assistance; Drs. Sascha Glinka and Birgit Ottenwälder at Eurofins Medigenomix GmbH for performing the GSFLX/454 sequencing; Drs. Cheryl Coffin, Robert Collins, David Page, and Charles Stratton at the Vanderbilt University School of Medicine and Dr. Ana Gonzalez-Angulo at the M.D. Anderson Cancer Center for providing critical reviews of the manuscript and their generous expertise. Myeloma samples kindly supplied by Dr. Brian G.M. Durie from Cedars-Sinai Medical Center, Los Angeles, CA. Multiple Sclerosis samples provided by Dr. Mario Clerici from the Don C. Gnocchi ONLUS Foundation IRCCS, Milan, Italy.

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

Received 07/16/2009; revised 12/21/2009; accepted 01/06/2010; published OnlineFirst 03/09/2010.

References

1. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin 2007;57:43–66.
2. Cronin KA, Feuer EJ, Clarke LD, Plevritis SK. Impact of adjuvant therapy and mammography on U.S. mortality from 1975 to 2000: comparison of mortality results from the CISNET breast cancer base case analysis. J Natl Cancer Inst Monogr 2006;36:112–21.
3. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin 2008;58:71–96.
4. Berg WA. Tailored supplemental screening for breast cancer: what now and what next? AJR Am J Roentgenol 2009;192:390–9.
5. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. JAMA 2005;293:1245–56.
6. Breen N, Cronin K, Meissner HI, et al. Reported drop in mammography: is this cause for concern? Cancer 2007;109:2405–9.
7. Lambertz CK, Johnson CJ, Montgomery PG, Maxwell JR. Premedication to reduce discomfort during screening mammography. Radiology 2008;248:765–72.
8. Poulos A, McLean D, Rickard M, Heard R. Breast compression in mammography: how much is enough? Australas Radiol 2003;47:121–6.
9. Easton DF. How many more breast cancer predisposition genes are there? Breast Cancer Res 1999;1:14–7.
10. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 2008;358:2796–803.
11. 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.
12. Mandel P, Metais P. Les acides nucléiques du plasma sanguin chez l'homme. C R Seances Soc Biol Fil 1948;142:241–3.
13. Gal S, Fidler C, Lo YM, et al. Quantitation of circulating DNA in the serum of breast cancer patients by real-time PCR. Br J Cancer 2004;90:1211–5.
14. Silva JM, Silva J, Sanchez A, et al. Tumor DNA in plasma at diagnosis of breast cancer patients is a valuable predictor of disease-free survival. Clin Cancer Res 2002;8:3761–6.
15. Sozzi G, Conte D, Leon M, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. J Clin Oncol 2003;21:3902–8.
16. Wu TL, Zhang D, Chia JH, Tsao KH, Sun CF, Wu JT. Cell-free DNA: measurement in various carcinomas and establishment of normal reference range. Clin Chim Acta 2002;321:77–87.
17. Boddy JL, Gal S, Malone PR, Harris AL, Wainscoat JS. Prospective study of quantitation of plasma DNA levels in the diagnosis of malignant versus benign prostate disease. Clin Cancer Res 2005;11:1394–9.
18. Boddy JL, Gal S, Malone PR, Shaida N, Wainscoat JS, Harris AL. The role of cell-free DNA size distribution in the management of prostate cancer. Oncol Res 2006;16:35–41.
19. Holdenrieder S, Nagel D, Schalhorn A, et al. Clinical relevance of circulating nucleosomes in cancer. Ann N Y Acad Sci 2008;1137:180–9.
20. Anker P, Mulcahy H, Stroun M. Circulating nucleic acids in plasma and serum as a noninvasive investigation for cancer: time for large-scale clinical studies? Int J Cancer 2003;103:149–52.
21. Beck J, Urnovitz H, Riggert J, Clerici M, Schütz E. Profile of the circulating DNA in apparently healthy individuals. Clin Chem 2009;55:730–8.
22. Schütz E, Urnovitz HB, Iakoubov L, Schulz-Schaeffer W, Wemheuer W, Brenig B. Bov-tA short interspersed nucleotide element sequences in circulating nucleic acids from sera of cattle with bovine spongiform encephalopathy (BSE) and sera of cattle exposed to BSE. Clin Diagn Lab Immunol 2005;12:814–20.
23. Kestler HA. ROC with confidence—a Perl program for receiver operator characteristic curves. Comput Methods Programs Biomed 2001;64:133–6.
24. Akaike H. Likelihood of a model and information criteria. J Econom 1981;16:3–14.
25. Burnham K, Anderson D. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 2004;33:260–304.
26. Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Crit Care 2004;8:508–12.
27. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2004 [cited 2008 April 30]; Available from: http://www.repeatmasker.org.
28. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;110:462–7.
29. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
30. Quentin Y. Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. Nucleic Acids Res 1992;20:3397–401.
31. Umetani N, Giuliano AE, Hiramatsu SH, et al. Prediction of breast tumor progression by integrity of free circulating DNA in serum. J Clin Oncol 2006;24:4270–6.
32. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst 2003;95:511–5.
33. Ballestar E, Esteller M. The impact of chromatin in human cancer: linking DNA methylation to gene silencing. Carcinogenesis 2002;23:1103–9.
34. Deligezer U, Akisik EE, Erten N, Dalay N. Sequence-specific histone methylation is detectable on circulating nucleosomes in plasma. Clin Chem 2008;54:1125–31.
35. Sunami E, Vu AT, Nguyen SL, Giuliano AE, Hoon DS. Quantification of LINE1 in circulating DNA as a molecular biomarker of breast cancer. Ann N Y Acad Sci 2008;1137:171–4.
36. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene 2002;21:5400–13.

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

37. Korenberg JR, Rykowski MC. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell 1988;53:391–400.
38. Holmquist GP, Ashley T. Chromosome organization and chromatin modification: influence on genome function and evolution. Cytogenet Genome Res 2006;114:96–125.
39. Dullea RG, Robinson JF, Bedford JS. Nonrandom degradation of DNA in human leukemic cells during radiation-induced apoptosis. Cancer Res 1999;59:3712–8.
40. Han HJ, Russo J, Kohwi Y, Kohwi-Shigematsu T. SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. Nature 2008;452:187–93.
41. Gotzmann J, Meissner M, Gerner C. The fate of the nuclear matrix-associated-region-binding protein SATB1 during apoptosis. Cell Death Differ 2000;7:425–38.
42. Gahan PB, Swaminathan R. Circulating nucleic acids in plasma and serum. Recent developments. Ann N Y Acad Sci 2008;1137: 1–6.
43. Urnovitz HB, Murphy WH. Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev 1996;9:72–99.
44. Cho K, Lee YK, Greenhalgh DG. Endogenous retroviruses in systemic response to stress signals. Shock 2008;30:105–16.
45. Ponzielli R, Boutros PC, Katz S, et al. Optimization of experimental design parameters for high-throughput chromatin immunoprecipitation studies. Nucleic Acids Res 2008;36:e144.

Published OnlineFirst March 9, 2010; DOI: 10.1158/1541-7786.MCR-09-0314

Molecular Cancer Research

Next Generation Sequencing of Serum Circulating Nucleic Acids from Patients with Invasive Ductal Breast Cancer Reveals Differences to Healthy and Nonmalignant Controls

Julia Beck, Howard B. Urnovitz, William M. Mitchell, et al.

Mol Cancer Res 2010;8:335-342. Published OnlineFirst March 9, 2010.