Molecular Classification of Breast Carcinomas by Comparative Genomic Hybridization: a Specific Somatic Genetic Profile for *BRCA1* Tumors[1]

Lodewyk F. A. Wessels, Tibor van Welsem, Augustinus A. M. Hart, Laura J. van't Veer, Marcel J. T. Reinders, and Petra M. Nederlof[2]

Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft [L. F. A. W., M. J. T. R.], and Departments of Pathology [T. v. W., P. M. N., L. J. v. V.] and Radiotherapy [A. A. M. H.], the Netherlands Cancer Institute, Amsterdam, the Netherlands

ABSTRACT

In ~70% of the families with a high frequency of early-onset breast and/or ovarian cancer, *BRCA1* or *BRCA2* germline mutations cannot be identified with the current screening regime. Therefore, we used data mining to identify a somatic genetic signature to differentiate *BRCA1* mutation carriers from non-*BRCA1* carriers based on the genetic characteristics of their breast carcinomas.

For this purpose, we developed a molecular classifier, which assigns a given tumor to either the *BRCA1* or control group based on somatic genetic profiles as revealed by comparative genomic hybridization. This was performed on breast tumors selected from two groups of patients: 28 proven *BRCA1* germline mutation carriers; and a control group consisting of 42 breast tumors from patients with unknown *BRCA1* or *BRCA2* status.

We show that *BRCA1* breast carcinomas exhibit specific somatic genetic aberrations and can be distinguished from control tumors with an accuracy of 84% (sensitivity of 96% and specificity of 76%). Chromosomal bands used by this classifier include regions on chromosomes 3p, 3q, and 5q. The classifier miss-assigned one patient with a *BRCA1* mutation to the non-*BRCA1* class. The germline mutation in this patient is a 62bp deletion in the last exon of *BRCA1* (5622del62). Possibly, this mutation may give a different phenotypic effect than do mutations in other regions of the gene.

Validation on an independent set of *BRCA1* and sporadic tumors showed that the *BRCA1* classifier correctly identified all 6 *BRCA1* tumors and assigned 4 of the 19 control patients to the *BRCA1* class. The resulting accuracy on the validation set is 84%.

INTRODUCTION

Germline mutations in *BRCA1* or *BRCA2* confer an estimated cumulative risk of developing breast cancer by age 70 years of 56–84% (up to 50% by age 50 years) and a 15–65% increased cumulative lifetime risk of ovarian cancer. The range in risk estimates depends on whether the group analyzed is a sample of community-based breast cancer patients with only a limited number of (Ashkenazi Jewish) mutations evaluated or is composed of large selected high-risk breast cancer families extensively tested for *BRCA1* and *BRCA2* (1, 2). Their risk for contralateral breast cancer is ~1–4% per annum depending on the age at the primary diagnosis, as compared with 0.5–1% observed for breast cancer patients without a breast cancer family history (3, 4).

Candidates for genomic screening for *BRCA1* and *BRCA2* germline mutations are usually selected based on a set of criteria including family history, age of onset, and occurrence of ovarian carcinoma in the family. At present, in ~30% of the families that fulfill these criteria (~2200 families tested in the Netherlands), germline mutations are identified by the currently used genomic screening techniques (5, 6). The actual prevalence of *BRCA1*/*BRCA2* mutation carriers within this group, however, is expected to be higher, based on *BRCA* linkage analysis. It is likely that particular mutations (*e.g.*, in the regulatory region) are undetected because not the complete gene but only the coding sequence is usually analyzed.

Breast carcinomas from patients with a germline mutation in *BRCA1* or *BRCA2* show some typical histopathological characteristics (7–10). They are generally high grade and more frequently show an atypical medullar-like appearance and "pushing margins," and are often estrogen receptor negative. However, no unique clinical and/or histological profile by which *BRCA1* or *BRCA2* tumors can be identified has emerged.

In addition to the histopathological differences between *BRCA1* tumors and sporadic breast tumors, genetic differences have been reported that indicate that *BRCA1* tumors may develop via a distinct developmental pathway. More frequent loss of heterozygosity at chromosomes 2q, 4q, 5q, and 12q and more frequent and specific somatic mutations in p53 have been reported (11).

These differences between *BRCA1* tumors and sporadic tumors motivated the development of a genome wide screening technique at the somatic molecular level. In this paper, we describe how such a screening technique can be designed by training a classifier to identify potential *BRCA1* mutation carriers within a group of breast carcinomas based on specific CGH[3] marker regions. We used CGH to build the classifier, because this technique allows the use of archival, routinely formalin-fixed paraffin-embedded material, often the only available material of affected deceased family members. Although several recent CGH studies report on the genetic aberrations in sporadic breast tumors (12–19), only a single CGH study involved both sporadic and *BRCA1* tumors (20). Although a *BRCA1* profile was proposed, no classifier was constructed to distinguish between sporadic and *BRCA1* tumors, and consequently no classification accuracy was attached to the profile.

MATERIALS AND METHODS

Tumor Series

Breast tumors from patients were selected from the pathology paraffin archive or the tissue-bank containing snap-frozen material, of the Netherlands Cancer Institute; 28 breast carcinomas were from proven *BRCA1* germline mutation carriers, and 42 control breast carcinomas were from patients with unknown *BRCA1* or *BRCA2* status. Five *BRCA1* patients had bilateral breast cancer (of which three were synchronous tumors) of whom both tumors were analyzed. From the remaining seven *BRCA1* patients with bilateral breast cancer, we analyzed one tumor only. The control set includes 13 pairs of primary tumors from patients with bilateral breast cancer; of these, 6 were synchronous bilateral tumors. The seven asynchronous tumors had a mean interval of six years between diagnosis. Seven patients had unilateral breast cancer (Table 1). We chose a control group with a relatively large proportion of bilateral tumors because we want to develop a classifier that will perform well on tumors from high-risk families, where bilateral tumors are frequent.

Received 2/18/02; accepted 9/25/02.

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.

[1] This work was partly funded by The Dutch Cancer Society NKB (T. v. W.) and the Intelligent Molecular Diagnostic System Program of the Delft Inter-Facultary Research Center at the Technical University of Delft (L. F. A. W. and M. J. T. R.).

[2] To whom requests for reprints should be addressed, at Department of Pathology, Netherlands Cancer Institute, Plesmanlaan 121, 166 CX Amsterdam, the Netherlands.

[3] The abbreviations used are: CGH, comparative genomic hybridization; VAF, variance accounted for; SBC, simple Bayesian classifier; LOOCV, leave-one-out-cross-validation; ROC, receiver operator characteristic.

Table 1 *Clinicopathological characteristics of breast carcinomas from BRCA1 mutation carriers in comparison to control cancers*

	BRCA1	Control
Clinical characteristics		
No. of tumors analyzed	28	42
Patients with breast and ovarian carcinoma	3	9
Patients with bilateral breast carcinomas	12	22
Pairs of bilateral breast carcinomas analyzed	5	13
Mean age at diagnosis (range)	40 (28–62)	51 (31–73)
Histology		
Invasive ductal carcinoma	28	35
Invasive lobular carcinoma	0	2
Other	0	5
Grade I	0	3
Grade II	6	20
Grade III	22	19
Immunohistochemistry		
ER +[a]	12% (3/24)	57% (21/37)
PR +	17% (4/23)	44% (16/36)
Neu +	0% (0/22)	11% (4/35)
Cyclin D1 +	0% (0/13)	15% (5/34)
P53 +	41% (9/22)	26% (9/34)

[a] ER, estrogen receptor; PR, progesterone receptor; Neu, HER2-neu; +, positive.

The only disadvantage of this approach is that in our control group some unidentified *BRCA1* carriers may be present. As a result, the predicted performance may be underestimated.

All tumors were reviewed for histological type and grade, and immunohistochemistry was performed with antibodies directed against the estrogen receptor, progesterone receptor, p53, Her2-neu and cyclin D1 (Table 1).

CGH Analysis

Genomic DNA was isolated from 4 × 10-μm-thick sections of 52 paraffin-embedded tumors using a commercially available genomic DNA extraction kit (Qiagen). The only adaptation was the incubation of the sections in 1 M sodium thiocyanate for 16 h at 37°C before proteinase K extraction. From 18 snap-frozen tumor samples and from peripheral blood lymphocytes of healthy women (normal reference) DNA was extracted by proteinase K according to standard protocols (14).

CGH was performed as described elsewhere (16). DNA was labeled by nick-translation with biotin-dUTP for paraffin-extracted DNA or FITC-dUTP for DNA from snap-frozen material and Texas Red-dUTP for reference DNA extracted from lymphocytes.

Labeled reference DNA (250 ng) and labeled tumor DNA (400 ng), with optimal fragment size of 300–500 bp, together with 25 μg of unlabeled human Cot-I DNA was dissolved in 10.5 μl of hybridization mixture consisting of 50% deionized formamide, 10% dextran sulfate, and 2×SSC. The DNA mixture was denatured for 5 min at 75°C and incubated for 5 min on ice and 30 min at 37°C. Slides with normal metaphase chromosomes were denatured for 3 min at 70°C-72°C in denaturation solution (70% formamide, 2× SSC, pH 7). The slides were dehydrated in 70% (−20°C), 85%, and 100% ethanol and air-dried. The probe mixture (10.5 μl) was added to the denatured metaphase spreads under a glass coverslip and allowed to hybridize in a moist chamber at 37°C for 72 h. Images were collected using a Zeiss microscope equipped with a Photometrix cooled CCD camera. The CGH analysis was performed using QUIPS software (Vysis, Inc., Downers Grove, IL) resulting in fluorescence ratios across all chromosomes and dividing the total genome into 1716 equal elements (channels). For each tumor, on average, six hybridization profiles were measured for each chromosome. Ratios were then averaged across hybridizations and used in the analysis.

Data Preparation

Discrimination between potential *BRCA1* mutation carriers and controls is based on carefully selected regions of the CGH profile. The specific pattern of aberrations exhibited by the *BRCA1* tumors on these selected regions is defined as the *BRCA1* profile. Given that raw CGH profiles consist of a large set of channels (~1700) and that the data sets contain relatively few (70) tumors, one runs the risk of "overfitting" the data. The consequence is that the resulting classifier performs very well on the collection of tumors used to develop it but may frequently misclassify unseen tumors. To avoid such a situation, the complexity of the data used by the classifier must be constrained intelligently. This is achieved by changing the resolution, the representation, as well as the number of regions used by the classifier. Suitable statistical techniques are used to determine the optimal settings.

Resolution

The data are analyzed at different levels of resolution. At the arm resolution, the ratio profile on a chromosome arm (p or q) is represented by a single value. At the most detailed level (channel resolution), the ratio profile of a particular tumor is represented as a set of 1716 values across the 23 chromosomes: 1 value per channel.

We developed an intermediate "band" resolution (different from the cytogenetic banding). To this end, we used a clustering approach to automatically group neighboring channels into nonoverlapping bands. The band boundaries are defined such that the channels in each band exhibit a high degree of correlation with each other. The degree of correlation between the channels in each band is characterized by the VAF. More specifically, the VAF reflects the fraction of information that is retained in a band if all channels in that band are represented by the mean ratio of all channels in that band. The VAF determines the number of bands and thus the resolution obtained. By lowering the minimal required level of the VAF per band, more channels can be included in a single band without violating the VAF constraint, resulting in a lower resolution, *i.e.*, fewer bands per chromosome. A particular banding, in which, for example, 85% of the information is retained, is represented by the label 0.85VAF. The band regions are represented by a tag of the form X.Y, with X denoting the chromosome, and Y the index of the band on the chromosome. The same notation is used for the channel representation, with Y denoting the index of the channel on the chromosome.

Representation

Given a particular resolution, the profiles are represented by a single variable per region (arm, band or channel). In this study, we used either discrete or continuous values. At the arm resolution, complete arms (p or q) are scored as gain (or loss) if the ratio exceeds 1.15 (or is below 0.85). Because centromeres and telomeres generally show irreproducible results in CGH analysis, they were excluded before discretization. If fewer than four consecutive channels showed a gain or loss (often the case at the telomeres), this was considered to be because of noise and was ignored. These ratio and noise thresholds were based on the results obtained with normal *versus* normal hybridization (results not shown), and similar thresholds are generally used in other published studies. Bands and channels are scored using the same thresholds. If no gain or loss occurs in a particular region, that region is scored as normal. Discretization results in three possible discrete values, gain (G), loss (L), or normal (N).

Continuous valued representations are defined as follows. The ratio values produced by the CGH analysis are log transformed to obtain the channel resolution representation. Averaging the ratio profile across the channels in a band produces the band resolution representation. No continuous valued representation was defined at the arm resolution.

Building a Classifier

Construction of a classifier involves the following steps: (*a*) collection of a suitable data set; (*b*) selection of a suitable classifier; and (*c*) optimization of the classification performance by selecting an appropriate resolution, representation, and region set (21). The first step has already been discussed in the previous sections. In the second step, we selected the SBC (22). We chose this classifier because, firstly, it performed comparably with or better than several other classifiers that were evaluated: logistic regression (see Ref. 23 for an elaborate comparison); the "See 5.0" classifier (24); and the nearest neighbor classifier (21). Secondly, it is readily interpretable (25), and finally, it can be used in both the continuous and discrete representations. In the continuous representation, a Parzen density estimator with gaussian kernels and a fixed SD of 0.2 was used. The SD of the kernel was estimated from the average SD of the hybridization profiles that were averaged to obtain the CGH profile of a patient. The prior probabilities of the *BRCA1* and control classes were set to 0.6 and 0.4, respectively. (A ROC analysis is included in "Results" (Fig. 4), from which it is clear that the optimal classifier is relatively insensitive with respect to the exact settings of these parameters.)

The third step in the construction of the classifier involves the selection of an appropriate representation, resolution, and region set. If the resolution and representation are fixed, limiting the number of regions used during classification can further reduce the complexity of the classifier. The set of regions (collection of arms, bands, or channels) that is best suited for the classification task can be selected in different ways. We investigated two typical selection approaches, "filtering" and "wrapping" (26). Both approaches generate a set of regions, which maximizes classification performance. The selected set of regions represents, for the given representation and resolution, the "genetic hot-spots" on the CGH profile, *i.e.,* positions on the genome that are crucial for distinguishing potential *BRCA1* patients from non-*BRCA1* patients.

LOOCV

The selection of the representation, resolution, and region set is performed based on the LOOCV performance of the classifier. LOOCV is used during the training of a classifier to prevent "overtraining" of a classifier on the training set (*i.e.,* very good performance on the training set, whereas performance on unseen tumors is poor). The procedure is as follows: given a training set of N tumors, the first tumor in the training set, t_1, is set aside (left out). Then the classifier is trained on the remaining N-1 tumors and tested (cross-validated) on the left out sample producing a score, s_1, which is either 0 (incorrect tumor class) or 1 (correct tumor class). Then sample t_1 is inserted back into the dataset and the next tumor, t_2, is left out. A new classifier is trained on the remaining N-1 tumors and tested on t_2, producing a score, s_2. This process is repeated until every tumor in the dataset had the opportunity to be a left out sample. The LOOCV score is then defined as the average score of all of the individual classifiers: LOOCV score $= \Sigma s_i/N$. The LOOCV performance is a realistic indicator of future performance of a classifier on unseen data and is a widely used statistical technique.

Region Selection

Filtering. This approach involves two steps, and both steps are repeated for every LOOCV iteration; *i.e.,* both steps are performed on the set of tumors that remains after the left out sample has been removed. During the first step, the regions are ordered (from best to worst) based on a score that reflects how well regions distinguish *BRCA1* tumors from the control tumors. Two scoring criteria were used: mutual information (27) for discrete representations; and the Mann-Whitney U test for continuous representations.

After the regions are scored and ordered, the optimal set of regions to include in the classifier is determined as follows. The process starts by training a classifier on the best scoring region and testing on the left out tumor. Then, the next best scoring region is added, such that the classifier is trained and tested on the top two scoring regions. This process of adding the next best scoring region is continued until a definite peak in the LOOCV performance is detected. The position of the peak indicates the optimal number of top scoring regions to include in the final classifier. In other words, adding more regions to the classifier does not further improve the classification performance.

Table 2 *Significant chromosomal regions between BRCA1 and control tumors*

Percentage of tumors from *BRCA1* and control tumors with loss or gain of chromosome arms. Only chromosome arms with a significant difference (χ^2, not corrected for multiple sampling, $P < 0.05$) between the groups are listed.

Loss	*BRCA1* (%)	Control (%)	*P*	Gain	*BRCA1* (%)	Control (%)	*P*
3p	61	31	0.014	3q	79	26	0.000
4p	64	38	0.032	7p	39	12	0.008
5q	82	40	0.001	8q	96	67	0.003
12q	54	26	0.020	10p	46	14	0.003
16p	32	12	0.038	12p	46	17	0.007
18q	54	26	0.020	16p	11	31	0.048
				17q	29	52	0.049

Region scoring can also be used simply to detect regions where the *BRCA1* and sporadic classes differ significantly. In this case, no classifier is constructed. To prevent the occurrence of false positives [because multiple tests (one on each region) are performed] step-down resampling (28) was used to produce a set of corrected *P*s for the region scores.

Wrapping. In this approach, a particular set of regions is scored based on their LOOCV classification performance (rather than using the mutual information or Mann-Whitney U test scores for the regions). We used a forward selection approach (adding the single region that gives the largest immediate improvement) to search the immense space of possible combinations of regions. To avoid overfitting the data, a double loop LOOCV training approach is used. The inner LOOCV loop is used to select and evaluate the regions, whereas the outer LOOCV is used to provide an independent performance estimate. This approach is computationally intensive but selects regions that optimize the LOOCV classification performance directly, rather than an indirect score.

RESULTS

Identification of Statistically Significant Regions. A series of 28 *BRCA1* and 42 control breast carcinomas were analyzed by CGH. We analyzed the obtained CGH profiles at various levels of resolution. In Fig. 1, the percentage of tumors with a gain (ratio above 1.15) or loss

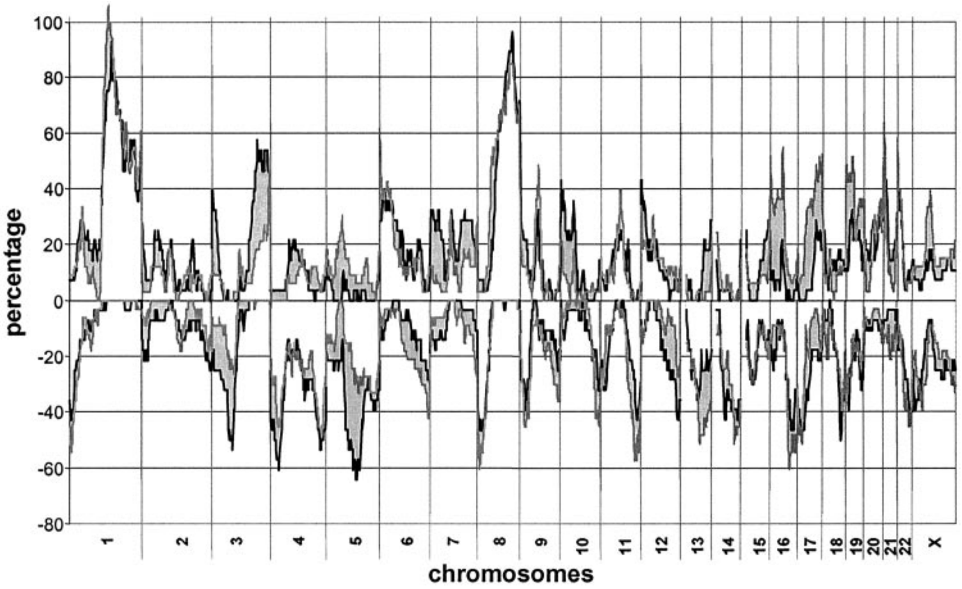


Fig. 1. Chromosomal aberrations. The percentage of tumors showing gain or loss at a specific chromosomal location (evaluated at the channel level) for the *BRCA1* group (N = 28, *black line*) and the control group (N = 42, *gray line*). *Gray shaded,* difference between the groups. Significant arms (χ^2; $P < 0.05$) are listed in Table 2.

Table 3 *Significant regions at different levels of resolution*

Summary of the chromosomal regions where the filtering criteria (mutual information and *U* test) detected significant (adjusted for multiple sampling, $P < 0.1$) correlations with the class labels (*BRCA1* and control).

	Total no. of regions	Discrete (Mutual information)		Continuous (*U* test)	
		Region	Adjusted *P*	Region	Adjusted *P*
Arm	41	3q	0.013	Not applicable	
		10p	0.080		
Channel	1716	5.26–5.30	<0.10	3.83–3.98	<0.10
				5.27–5.50	<0.10
Band 0.85VAF	81	3.1	0.015	3.2	0.053
		3.5	0.002	3.5	0.005
		5.2	0.009	5.2	0.003
				5.3	0.078

(ratio below 0.85) for each channel (1720 in total) along the chromosomes is depicted. The percentages for the *BRCA1* (Fig. 1, *black line*) and the control group (Fig. 1, *gray line*) are depicted with the *light gray areas* representing the differences between the groups. When tested individually, several chromosomal arms show a significant difference (χ^2; $P < 0.05$) between the groups (Table 2). At the chromosome arm level, *BRCA1* tumors showed more frequent loss of 3p, 4p, 5q, 12q, 16p, and 18q and gain of chromosome 3q, 7p, 8q, 10p, and 12p. Control tumors showed more frequent gain of 16p and 17q.

As is apparent from Fig. 1, the loss or gain of chromosomal material does not always span the complete chromosome arm; sometimes smaller regions, such as on 5q, are affected. For this reason, we developed a method to increase the resolution (number of regions) by creating bands consisting of channels with similar fluorescence ratios. In Table 3, the chromosomal regions that harbor significant differences (adjusted for multiple testing; $P < 10\%$) are given. Results for both discrete data and continuous valued representations are presented, where mutual information and the *U* test were respectively used as filtering criteria. No effort was made to classify with these regions; this is described in "Building a Classifier." Table 3 only lists results for the 0.85VAF banding, because it was found to be the best scoring banding. From the results, we note that for all levels of resolution, for both the continuous and discrete representations, the significant regions are on chromosomes 3 and 5, except for the arms representation, where 10p is also significant. In Fig. 2 (*top*), the positions of the regions on chromosomes 3 and 5 are indicated for the arm, 0.85VAF banding, and channel resolutions.

Building a Classifier. Two approaches to select marker regions were evaluated: wrapping and filtering. As filtering criteria we used the *U* test and mutual information for continuous and discrete representations, respectively. In all cases, the SBC was used to perform the classification on the selected regions. Table 4 lists the classifiers with the best LOOCV classification performance and the associated region sets. Results are listed for the different resolutions and the two alternative representations. The region sets that were selected with mutual information filtering gave consistently better results than the regions produced by wrapping. The only exception is for the discrete chromosomal arms resolution, where wrapping results in a classification performance of 79% whereas filtering reaches 73%. Secondly,

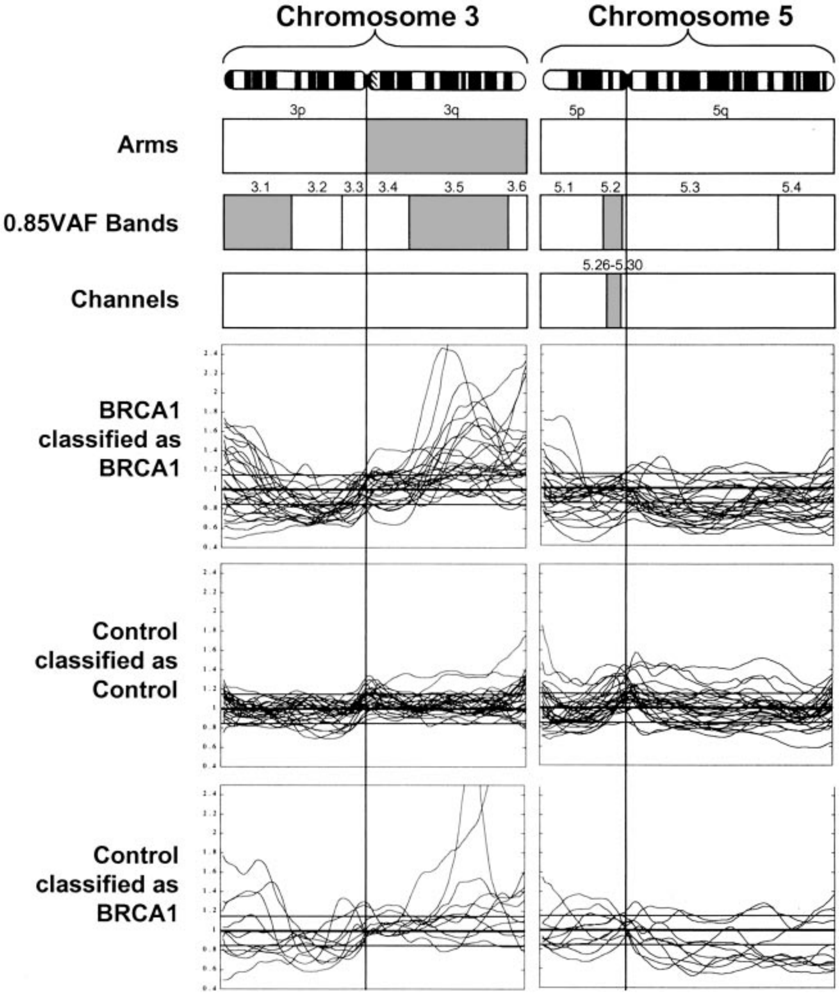


Fig. 2. Cytogenetic banding and 0.85VAF banding of chromosome 3 and 5. *Top,* statistically significant (adjusted $P < 0.1$) chromosomal regions for the mutual information criterion indicated as *gray shaded areas* for the arm (only 3q is shown and not 10p), 0.85VAF band, and channel resolution. *Bottom,* CGH ratio profiles after classification of the 28 *BRCA1* and 42 control tumors using the *BRCA1* classifier based on the 0.85VAF banding and discrete representation. Ten controls were misclassified as *BRCA1* (*bottom*). *Two horizontal lines* at ratios of 0.85 and 1.15 delineate the range associated with no genetic change. *Vertical line,* positions of centromeres.

Table 4 *Performance of BRCA1 classifiers*

Summary of the best LOOCV performance and associated region sets for classification of *BRCA1* tumors. Results are shown for the wrapped discrete SBC, as well as for the mutual information filtered SBC for both discrete and continuous representations. Only region sets where limited variation occurred in the selected sets (across the LOOCV loops) are listed.

	Wrapped discrete SBC		Mutation information filtered SBC	
	Regions	Performance (%)	Regions	Performance (%)
Discrete representation				
Arms	3q and 10p	79	3q and 10p	73
Bands[a]	3.5 and 5.2 and (2.1 or 10.2 or 3.1)	77	3.1 and 3.5 and 5.2	84
Channels	11 regions	64	16 regions	67
Continuous representation				
Arms	Not applicable		Not applicable	
Bands[a]	9 regions	67	3.5 and 5.2	76
Channels	Too computationally expensive		5.31 and 5.32	77

[a] 0.85VAF bands.

the discrete representation consistently resulted in better performance than the continuous representation, except for the channel resolution. Finally, the band resolution (relative to channel and arms resolution) results in an improvement of the classification performance for the discrete representation (0.85VAF being the best resolution) whereas banding results in a decrease in performance for the continuous representation.

The BRCA1 Profile. The best classifier is the SBC trained on regions with a discrete representation selected from the 0.85VAF banding with mutual information filtering. The best 0.85VAF regions are bands 3.1, 3.5; and 5.2. (Fig. 2; Table 4). The *BRCA1* profile yielding the best classification performance is: [±3.1, +3.5, −5.2]. Fig. 3 depicts the *a posteriori* probabilities of these regions as they are used in the discrete SBC. These histograms represent the probability that a tumor is *BRCA1* or control given the particular aberrations observed in the tumor.

These histograms provide insight into the behavior of the derived CGH profiles. More specifically, samples from the *BRCA1* class typically show more frequent gain or loss than samples from the control class on band 3.1 (Fig. 3*A*). In addition, they show more frequent gain on band 3.5 (Fig. 3*B*) and more frequent loss than the controls on band 5.2 (Fig. 3*C*). This is also visible in the profile plots in Fig. 2. Here, the individual profiles of the correctly classified tumors for the *BRCA1* and control groups are depicted, as well as the 10 control tumors that show a *BRCA1* profile.

False Positive Analysis. For the best *BRCA1* classifier, the ROC was constructed. The result is depicted in Fig. 4. This curve is obtained by varying the prior probability of the *BRCA1* class (P_B) from 0 to 1 in small steps. The prior probability of the control class is then given by $P_S = 1 - P_B$. The classifier associated with the settings used in the study ($P_B = 0.6$ and $P_S = 0.4$) is indicated with a circle in Fig. 4. The encircled position of the graph is, in fact, associated with a whole set of classifiers obtained for values of P_B ranging from 0.58 to 0.89. This indicates that the optimal classifier is relatively insensitive to the exact setting of P_B. The encircled position is characterized by a relatively high false positive rate and a low false-negative rate. In our particular screening application, this is a desirable situation, because very few patients with the *BRCA1* mutation will then be missed when the preselected set of patients is screened. Consequently, a rigorous genetic screening (sequencing) can be used to determine the carrier status.

Applying the *BRCA1* classifier resulted in 10 false positives (control tumors classified as *BRCA1*) and one false negative (*BRCA1* tumor classified as control). These are listed in the first column of Table 5. The third column lists the *a posteriori* score produced by the classifier, *i.e.*, the probability that a tumor is *BRCA1* given the observed profile. The larger this score, the greater is the "confidence" of the classifier in the class assignment. All false positives with a score larger than 0.9 have at least two regions that correspond with the *BRCA1* profile, either a gain or loss at 3.1 and a gain at 3.5. None of these cases shows loss of 5.2. The clinicopathological parameters of the false positive and false negative samples are also listed in Table 5. These parameters show little correlation with the false positive results.

Validation of the Classifier. To further validate the classifier, we applied the classifier on a new set of tumors. We analyzed 6 *BRCA1* and 19 control breast carcinomas. All tumors were grade 3 invasive ductal carcinomas; mean age at diagnosis was 45 years (range, 29–51) for the sporadic (all unilateral patients) and 47 years for the *BRCA1* (range, 31–62) group.

The classification scores for the training and the validation sets are depicted in Fig. 5. All six *BRCA1* tumors are classified correctly. The control group shows 4 of 19 tumors (21%) with a high score implying a potential *BRCA1* mutation. This is a similar percentage of false positives as was found in the control training set (8/40 = 20%). The overall classification accuracy for the validation set was 84%.

DISCUSSION

A Classifier Based on a Unique *BRCA1* Profile. We showed that *BRCA1* breast carcinomas exhibit specific somatic genetic aberrations and can be distinguished from control tumors with an accuracy of 84% on both the training and validation sets. Our goal is to use the classifier to prescreen high-risk patients, who meet the inclusion

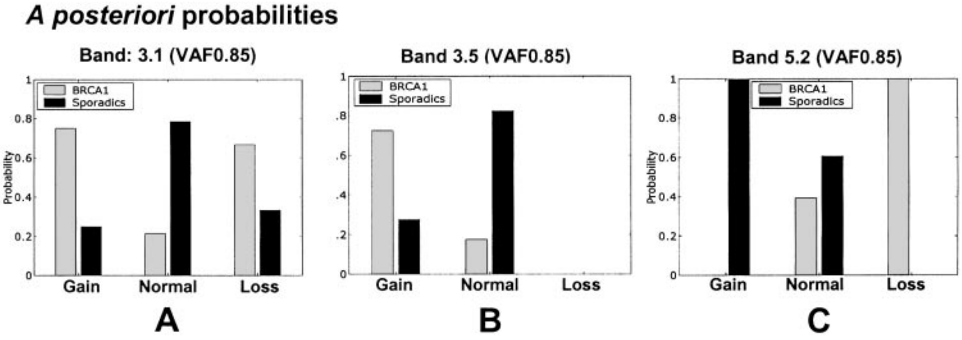


Fig. 3. *A posteriori* densities for the different regions for the best classifier. The distributions for the selected regions, bands 3.1, 3.5 and 5.2 (0.85VAF) are depicted in *A*, *B*, and *C*, respectively. The *a posteriori* density histograms represent the probability that a tumor is *BRCA1* or control given the particular aberration observed in the tumor. For example, for a tumor with a gain of band 3.1, the *a posteriori* distribution (*A*) indicates that the probability that the tumor belongs to the *BRCA1* class is high (0.75) and that it is low for the control class (0.25).

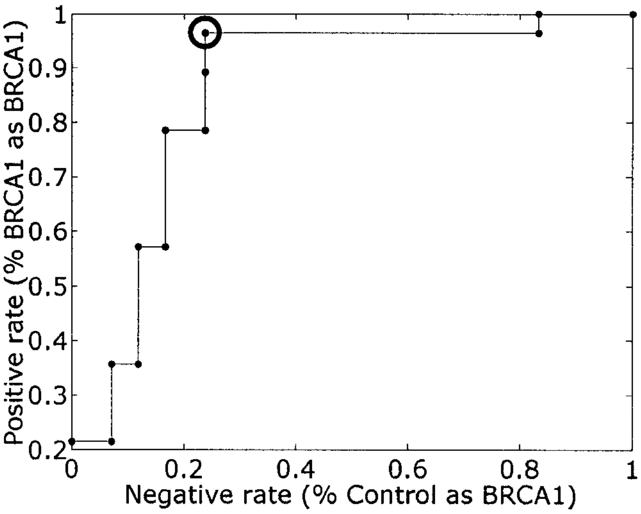


Fig. 4. ROC. The ROC for the best classifier (discrete SBC on selected 0.85VAF regions) is given. The curve depicts, for every combination of prior probabilities of the *BRCA1* and control groups, the positive rate and the negative rate. The positive rate is defined as the number of correctly classified *BRCA1* tumors; the negative rate represents the number of control tumors classified as *BRCA1* tumors. ○, position of the optimal classifier on the curve.

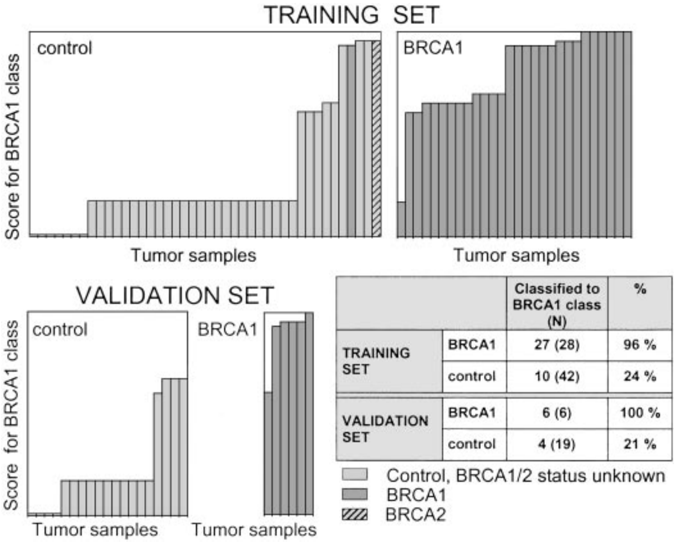
criteria for genetic screening in the context of counseling for hereditary breast cancer. In this group, we currently identify a *BRCA1* mutation in ∼30% of the patients, and we expect that an additional 10% of the patients may be *BRCA1* mutation carriers. Because the best classifier was trained on a population consisting of 40% *BRCA1* patients, the makeup of the training data matches the makeup of the population on which the classifier will be used.

The classification rule with the best performance was constructed using a discrete representation and the 0.85VAF band level resolution. Regions used by this classifier are situated on chromosomes 3p, 3q, and 5q. More specifically, the classification rule uses the following *BRCA1 profile* to identify *BRCA1* tumors: [±3.1; +3.5; −5.2]. The classifier classified 10 non-*BRCA1* (24%) tumors as *BRCA1* tumors. New information on the *BRCA1/2* status of two of the five false positives with the highest score (>0.9) became available and showed that one patient indeed has a *BRCA1* mutation. Two other high-scoring tumors are bilateral tumors from the same patient who also developed an ovarian tumor and has a family history of breast cancer; thus, although no mutation is known, this patient is highly likely to be a *BRCA1* carrier.

One high-scoring control patient had a *BRCA2* mutation. This is not completely surprising because *BRCA1* and *BRCA2* are interacting proteins involved in the same pathway. Preliminary results show that *BRCA2* tumors indeed show specific genetic similarities with *BRCA1* tumors on the regions used by the classifier (data not shown).

Even if all of the remaining false positives have no *BRCA1* mutation, a classifier with a relatively high false positive rate (and low false negative rate) is preferred, because it reduces the risk of missing a patient with a *BRCA1* mutation during a prescreening step. The set of patients labeled by the classifier as potential *BRCA1* mutation carriers can then be subjected to a final phase of rigorous genetic screening.

The classifier misassigned one patient with a proven *BRCA1* mutation to the non-*BRCA1* class. Also, the immunohistochemical characteristics for this tumor are different (positive staining for estrogen and progesterone receptors) compared with the majority of the *BRCA1* tumors. The germline mutation in this patient is a 62-bp loss in the last exon of *BRCA1* (5622del62; involving the last 1% of the protein). One possibility would be that this mutation gives a different phenotypic effect compared with mutations in other regions of the gene. However, a tumor from another patient with the same germline mutation showed the characteristic *BRCA1* profile. This could implicate that the tumor in this patient is actually a sporadic tumor in a *BRCA1* patient, which developed independent of the germline mutation. Interestingly, the mRNA expression profile of this tumor was also distinct from the



		Classified to BRCA1 class (N)	%
TRAINING SET	BRCA1	27 (28)	96 %
	control	10 (42)	24 %
VALIDATION SET	BRCA1	6 (6)	100 %
	control	4 (19)	21 %

Fig. 5. *BRCA1* classification score. For each individual tumor of the training set and the validation set (*BRCA1* and control), the *BRCA1* classification score as calculated using the simple Bayes rule, is depicted. A high score indicates a high probability to belong to the *BRCA1* class. The percentage of tumors from each group that are classified to the *BRCA1* class is indicated.

Table 5 *False positive analysis*

Profile, classifier scores, and clinical parameters for the false positives and false negatives identified by the best classifier.

Tumor	Features 3.1, 3.5, 5.2	Classifier score	Age (1st, 2nd)	Ovarian tumor	Histological grade	PR status	ER status	Tumor type
False positives								
1	[G,G,N][a]	0.95	58	No	3	−	−	IDC
2	[G,G,N]	0.95	45,47	Yes	3	−	−	IDC
3	[G,G,N]	0.95	45,47	Yes	3	−	−	MP
4	[L,G,N]	0.93	33,39	No	3	ND	ND	IDC
5	[L,G,N]	0.93	52,52	No	2	+	+	ILC
6	[N,G,N]	0.65	49,61	Yes	3	−	+	IDC
7	[N,G,N]	0.65	48,53	No	3	−	−	MP
8	[L,N,N]	0.61	39,42	No	3	−	+	IDC
9	[L,N,N]	0.61	44,57	No	3	−	−	IDC
10	[L,N,N]	0.61	45,54	No	2	+	−	IDC
False negatives								
11	[N,N,N]	0.17	35,46	No	2	+	+	IDC

[a] [G,N,L], gain, normal, loss; ND, not done; ER, immunohistochemistry for estrogen receptor; PR, immunohistochemistry for progesterone receptor; IDC, invasive ductal carcinoma; MP, metaplastic; ILC, invasive lobular carcinoma; Grade, histological grade.

BRCA1 signature as determined by expression arrays (31), which supports the hypothesis of a non-*BRCA1* development pathway.

The Control Group. The selection of the patient and control group has important consequences for the results obtained. The control group used in this study contains a many patients with bilateral tumors and five patients with additional ovarian carcinoma (Table 1). On the one hand, one would like to match the control group as much as possible to the group on which the classifier eventually will be used. Therefore, bilateral tumors, ovarian tumors, and tumors from young patients were included in the series. On the other hand, one cannot easily exclude the possibility that the control group will contain unidentified *BRCA1* (or *BRCA2*) carriers. This indeed appeared to be the case, and two carriers were identified in retrospect.

Comparison with a Previous Study. We showed that *BRCA1* breast carcinomas exhibit distinctive somatic genetic aberrations that allow the identification of *BRCA1* carriers on the basis of this profile. There is only one other paper that compares the somatic genetic changes in *BRCA1* tumors to control tumors (20). In this publication, several chromosomal arms were indicated that show more frequent aberrations in *BRCA1* tumors. Some of these changes were confirmed by our study, which is true for loss of 5q and 12q. However, we did not find the more frequent losses of chromosome 4q and 2q. This discrepancy probably occurs because in their control group the percentage of loss of chromosome 4 is extremely low. In other studies, the percentage of loss of chromosome 4 in control breast carcinomas is much higher, comparable with the frequency we find in our control group. In contrast to this study, our goal was to go beyond the identification of specific aberrations and to develop a classifier that would allow identification of individual *BRCA1* tumors within a group of high-risk patients.

CGH Resolution. Our study indicates that the somatic genetic changes in tumors from both *BRCA1* and non-*BRCA1* carriers frequently involve large regions of the chromosome. This implicates that the analysis of somatic aberrations in breast carcinomas may not improve by increasing the resolution of the applied technique. This was confirmed by the fact that a banding with relatively broad bands (0.85VAF) performed much better than a channel level representation. We used a banding technique based on channel correlation that allows optimal banding for analysis of CGH data. This method may be applied to the CGH analysis of other tumor types as well, to allow optimal banding based on the dataset.

Gene Expression Data. Some studies have recently emerged that evaluate the expression profiles of *BRCA1* tumors compared with control tumors (29–31). Although some of these studies used a few tumors, all showed an expression pattern unique to *BRCA1* tumors, indicating a specific tumor development pathway. This is in accordance with the genetic profile we detected at the genome level. However, in clinical settings, frequently only formalin-fixed paraffin-embedded material is available for analysis, and expression arrays cannot be used. The possibility to analyze archival material has great advantages. The classical CGH technique is, however, too time consuming for routine screening of many tumors. Therefore, we are now designing a simple screening test using, *e.g.*, CGH arrays (32), based on the classifier developed in this paper.

CONCLUSION

In this paper, we addressed the problem of molecular classification of breast cancer based on somatic genetic profiles revealed by CGH. The following contributions were made: (*a*) we developed a profile and classification rule with which tumors with a *BRCA1* mutation can be distinguished from control tumors with an accuracy of 84%. This profile includes regions on chromosome 3p, 3q, and 5q; (*b*) we performed a rigorous evaluation of two data analysis approaches for selection of the optimal set of regions: wrapping and filtering. Statistical techniques such as LOOCV and step-down resampling were used to assess the accuracy of the classifiers and the quality of selected regions, respectively. We found that filtering was less prone to uninformative regions and therefore produced region sets that resulted in higher classification accuracies; (*c*) we compared continuous and discrete representations and found the latter to be superior; and (*d*) we used a simple banding algorithm for grouping correlated channels into bands and found that for the discrete representation, the resulting resolution (0.85VAF banding) resulted in a 5% improvement in classification compared with the chromosomal arms resolution.

ACKNOWLEDGMENTS

We thank Cecile Ottenheim for assistance in karyotyping, Marc van de Vijver for review of the tumor samples, and Frans Hogervorst and Senno Verhoef for critical reading of the manuscript.

REFERENCES

1. Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T., Weber, B., Lenoir, G., Chang-Claude, J., Sdobol, H., Teare, M. D., Streuwing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T. R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyford, J., Lynch, H., Ponder, B. A. J., Gayther, S. A., Birch, J. M., Lindblom, A., Stoppa-Lyonnet, Bignon. Y., Borg, A., Hamann, U., Haites, N., Scott, R. J., Maugard, C. M., Vasen, H., Seitz, S., Cannon-Albright, L. A., Schofield, A., Zelada-Hedman, M., the Breast Cancer Linkage Consortium. Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. Am. J. Hum. Genet., *62:* 676–689, 1998.
2. Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C., and Tucker, M. A. The risk of cancer associated with specific mutations in *BRCA1*and *BRCA2* among Ashkenazi Jews. N. Engl. J. Med., *336:* 1401–1408, 1997.
3. Verhoog, L. C., Brekelmans, C. T. M., Seynave, C., Meijers-Heijboer, E. J., and Klijn, J. G., Survival and tumour characteristics of breast-cancer patients with germline mutations of. *BRCA1*. Lancet, *351:* 316–321, 1998.
4. Haffty, B. G., Harold, E., Kahn, A. J., Pathare, P., Smith, T. E., Turner, B. C., Glazer, P. M., Ward, B., Carter, D., Matloff, E., Bale, A. E., and Alvarez-Franco, M. Outcome of conservatively managed early-onset breast cancer by *BRCA1*/2 status. Lancet, *359:* 1471–1477, 2002,
5. Couch, F. J., DeShano, M. L., Blackwood, M. A., Calzone, K., Stopfer, J., Campeau, L., Ganguly, A., Rebbeck T., and Weber, B. L. *BRCA1* mutations in women attending clinics that evaluate the risk of breast. N. Engl. J. Med., *336:* 1409–1415, 1997.
6. Ligtenberg, M. J. L., Hogervorst, F. B. L., Willems, H. W., Arts, P. J. W., Brink, G., Hageman, S., Bosgoed, E. A. J., Van der Looij, E., Rookus, M. A., Devilee, P., Vos, E. M. A. W., Wigbout, G., Struijcken, P. M., Menko, F. H., Rutgers, E. J. T., Hoefsloot, E. H., Mariman, E. C. M., Brunner, H. G., and Van 't Veer, L. J. Characteristics of small breast and/or ovarian cancer families with germline mutations in *BRCA1* and *BRCA2.* Br. J. Cancer, *79:* 1475–1478, 1999.
7. Lakhani, S. R., Gusterson, B. A., Jacquemier, J., Sloane, J. P., Anderson, T. J., van de Vijver, M. J., Venter, D., Freeman, A., Antoniou, A., McGuffog, L., Smyth, E., Steel, C. M., Haites, N., Scott, R. J., Goldgar, D., Neuhausen, S., Daly, P. A., Ormiston, W., McManus, R., Scherneck, S., Ponder, B. A., Futreal, P. A., Peto, J., Stoppa-Lyonnet, D., Bignon, Y. J., and Stratton, M. R. The pathology of familial breast cancer: histological regions of cancers in families not attributable to mutations in *BRCA1* or *BRCA2.* Clin. Cancer Res., *6(3):* 782–789, 2000.
8. Lakhani, S. R., Jacquemier, J., Sloane, J. P., Gusterson, B. A., Anderson, T. J., van de Vijver, M. J., Farid, L. M., Venter, D., Antoniou, A., Storfer-Isser, A., Smyth, E., Steel, C. M., Haites, N., Scott, R. J., Goldgar, D., Neuhausen, S., Daly Ormiston, W., McManus, R., Scherneck, S., Ponder, B. A., Ford, D., Peto, J., Stoppa-Lyonnet, D., Easton, D. F., *et al.* Multifactorial analysis of differences between sporadic breast cancers and cancers involving *BRCA1* and *BRCA2* mutations. J. Natl. Cancer Inst., *90:* 1138–1145, 1998.
9. Robson, M., Rajan, P., Rosen, P. P., Gilewski, T., Hirschaut, Y., Pressman, P., Haas, B., Norton, L., and Offit, K. BRCA-associated breast cancer: absence of a characteristic immunophenotype. Cancer Res, *58:* 1839–1842, 1998.
10. Breast Cancer Linkage Consortium. Pathology of familial breast cancer: differences between breast cancers in carriers of *BRCA1* or *BRCA2* mutations and sporadic cases. Lancet, *349:* 1505–1510, 1997.
11. Smith, P. D., Crossland, S., Parker, G., Osin, P., Brooks, L., Waller, J., Philp, E., Crompton, M. R., Gusterson, B. A., Allday, M. J., and Crook, T. Novel p53 mutants selected in *BRCA*-associated tumours which dissociate suppression from other wild-type p53 functions. Oncogene, *18:* 2451–2459, 1999.
12. Rooney, P. H., Murray, G. I., Stevenson, D. A., Haites, N. E., Cassidy, J., and McLeod, H. L. Comparative genomic hybridization and chromosomal instability in solid tumours. Br. J. Cancer, *80:* 862–873, 1999.
13. Richard, F., Pacyna-Gengelbach, M., Schluns, K., Fleige, B., Winzer, K. J., Szymas, J., Dietel, M., Petersen, I., and Schwendel, A. Patterns of chromosomal imbalances in invasive breast cancer. Int. J. Cancer, *89:* 305–310, 2000.

14. Ried, T., Just, K. E., Holtgreve-Grez, H., Dumanoir, S., Speicher, M. R., Schrock, E., Latham, C., Blegen, H., Zetterberg, A., Cremer, T., and Auer, G. Comparative genomic hybridization of formalin-fixed, paraffin-embedded breast-tumors reveals different patterns of chromosomal gains and losses in fibroadenomas and diploid and aneuploid carcinomas. Cancer Res., *55:* 5415–5423, 1995.
15. Loveday, R. L., Greenman, J., Simcox, D. L., Speirs, V., Drew, P. J., Monson, J. R. T., and Kerin, M. J. Genetic changes in breast cancer detected by comparative genomic hybridization. Int. J. Cancer, *86:* 494–500, 2000.
16. Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., Gooden, G. C., Ethier, S. P., Kallioniemi, A., and Kallioniemi, O. P. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. Cancer Res., *60:* 4519–4525, 2000.
17. Roylance, R., Gorman, P., Harris, W., Liebmann, R., Barnes, D., Hanby, A., and Sheer, D. Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. Cancer Res., *59:* 1433–1436, 1999.
18. Hermsen, M. A. J. A., Baak, J. P. A., Meijer, G. A., Weiss, J. M., Walboomers, J. W. W., Snijders, P. J. F., and Van Diest, P. J. Genetic analysis of 53 lymph node-negative breast carcinomas by CGH and relation to clinical, pathological, morphometric, and DNA cytometric prognostic factors. J. Pathol., *186:* 356–362, 1998.
19. Tirkkonen, M., Tanner, M., Karhu, R., Kallioniemi, A., Isola, J., and Kallioniemi, O. P. Molecular cytogenetics of primary breast cancer by CGH. Genes Chromosomes Cancer, *21:* 177–184, 1998.
20. Tirkkonen, M., Johannsson, O., Agnarsson, B. A., Olsson, H., Ingvarsson, S., Karhu, R., Tanner, M., Isola, J., Barkardottir, R. B., Borg, A., and Kallioniemi, O. P. Distinct somatic genetic changes associated with tumor progression in carriers of *BRCA1* and *BRCA2* germ-line mutations. Cancer Res., *57:* 1222–1227, 1997.
21. Duda, R. O., Hart, P. E., and. Stork, D. G. Pattern Classification, Ed. 2. New York: Wiley, 2000.
22. Domingos, P., and. Pazzani, M. J. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, *29:* 103–130, 1997.
23. Wessels, L. F. A., Reinders, M. J. T., van Welsem, T., and Nederlof, P. M. Representation and classification for high-throughput data. *In:* Proceedings of SPIE, Vol. 4626 (BIOS2002), 226–237. San Jose, CA, January 2002.
24. Quinlan, J. R. C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1993.
25. Kononenko, I. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. *In:* B. Wielinga (ed.), Current Trends in Knowledge Acquisition, Amsterdam, the Netherlands: IOS Press, 1990.
26. Kohavi, R., and John, G. H. Wrappers for region subset selection. Artificial Intelligence, *97:* 273–324, 1997.
27. Butte, A., and Kohane, I. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *In:* Proceedings of the Pacific Symposium on Biocomputing, pp. 418–429. World-Scientific, Honolulu, Hawaii USA, 2000.
28. Westfall, P. H., and Young, S. S. Resampling-Based Multiple Testing: Examples and Methods for *P*-Value Adjustment. Wiley-Interscience, 1993.
29. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med., *344:* 539–548, 2001.
30. Berns, E. M., van Staveren, I. L., Verhoog, L., van de Ouweland, A. M., Meijer-van Gelder, M., Meijers-Heijboer Portengen, H., Foekens, J. A., Dorssers, L. C., and Klijn, J. G. Molecular profiles of *BRCA1*-mutated and matched sporadic breast tumours: relation clinico-pathological regions. Br. J. Cancer, *85:* 538–545, 2001.
31. Van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. Gene expression profiling of breast cancer accurately predicts clinical outcome of disease. Nature (London), *415:* 530–536, 2002.
32. Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat. Genet., *29(3):* 263–264, 2001.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

Molecular Classification of Breast Carcinomas by Comparative Genomic Hybridization: a Specific Somatic Genetic Profile for BRCA1 Tumors

Lodewyk F. A. Wessels, Tibor van Welsem, Augustinus A. M. Hart, et al.

Cancer Res 2002;62:7110-7117.