



**HAL**  
open science

## **Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes**

Marcel Smid, Marlous Hoes, Anieta M. Sieuwerts, Stefan Sleijfer, Yi Zhang, Yixin Wang, John A. Foekens, John W. M. Martens

► **To cite this version:**

Marcel Smid, Marlous Hoes, Anieta M. Sieuwerts, Stefan Sleijfer, Yi Zhang, et al.. Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. *Breast Cancer Research and Treatment*, 2010, 128 (1), pp.23-30. 10.1007/s10549-010-1026-5 . hal-00615363

**HAL Id: hal-00615363**

**<https://hal.science/hal-00615363>**

Submitted on 19 Aug 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes.**

Marcel Smid<sup>1,2</sup>, Marlous Hoes<sup>1</sup>, Anieta M. Sieuwerts<sup>1,2</sup>, Stefan Sleijfer<sup>1</sup>, Yi Zhang<sup>3</sup>, Yixin Wang<sup>3</sup>, John A. Foekens<sup>1,2</sup> and John W.M. Martens<sup>1,2</sup>.

**Affiliations:**

<sup>1</sup>Department of Medical Oncology, Josephine Nefkens Institute and <sup>2</sup>Cancer Genomics Centre, Erasmus MC, 3000CA Rotterdam, the Netherlands and <sup>3</sup>Veridex LLC, a Johnson & Johnson Company, North Raritan, NJ 08869, United States of America.

Corresponding author

John W. M. Martens, PhD  
Dept. of Medical Oncology,  
Josephine Nefkens Institute  
Erasmus MC,  
PO Box 1738,  
3000 DR Rotterdam,  
The Netherlands.  
Tel ++31 10 7044372  
Fax ++31 10 7044377  
[j.martens@erasmusmc.nl](mailto:j.martens@erasmusmc.nl)

## **Abstract**

**Introduction.** One of the hallmarks of human solid tumors is chromosomal instability (CIN). We studied global patterns as well as individual levels of CIN and determined the prognostic relevance among breast cancer subtypes. **Methods.** For this, we used single nucleotide polymorphism (SNP) copy number data of 313 primary lymph-node negative breast cancers. The level of CIN for individual samples was determined by counting the total number of chromosomal segments showing a gain or loss per specimen. **Results.** Hierarchical clustering resulted in 4 groups showing distinct patterns of abnormalities, predominantly characterized by 1q gain, 8q gain, 1q&8q gain or no gain of these loci. Estrogen Receptor (ER)-positive and ER-negative samples showed an uneven distribution (statistically significant) across the cluster groups, as did the molecular subtypes and triple-negative tumors (negative for estrogen-, progesterone- and her2/neu-receptor). The CIN-score was significantly higher in ER-negative and triple-negative samples. Among luminal cancers, luminal B had a higher CIN-score than luminal A. The CIN-score was significantly associated with prognosis, measured by the time to distant metastasis, in ER-positive, luminal B and her2/neu subtypes, but not in ER-negative patients. **Conclusions.** Our study points to a multifaceted role for CIN in breast cancer. Within ER-negative samples, CIN is likely related to the onset (high CIN rate) but other factors govern the progression of the disease. In contrast, CIN is clearly associated with progression in ER-positive, luminal B and her2/neu subtypes; thus assessing CIN in these subtypes may contribute to personalized patient management.

### **keywords:**

Chromosomal instability, subtypes, prognosis, SNP copy number

## Introduction

Chromosomal instability (CIN) is a well known phenomenon observed in many types of cancer and has a pivotal role in tumor onset and progression [1]. With respect to breast cancer, literature concerning CIN is mostly based on comparative genomic hybridization (CGH) techniques, which have demonstrated specific gains and losses of chromosomal areas in diverse breast cancer subtypes [2-8]. In addition, Carter et al. [9] have shown that a gene expression profile associated with the incidence of CIN has prognostic relevance in breast cancer. However, this latter study was based on a diverse panel of cancer cell lines and by establishing CIN through a gene expression profile; thus, CIN was assessed indirectly [9]. Very recently, whole genome paired-end sequencing data of breast cancer samples showed complex somatic rearrangements with some samples showing a large number of deletions, tandem duplications or amplifications, while other samples have virtually no rearrangements [10].

In this study, we investigated the patterns and prevalence of CIN in subtypes of breast cancer and the relevance of these 2 measures of CIN with respect to patient outcome. To this end, we used a high-resolution 100k single nucleotide polymorphism (SNP) microarray dataset of 313 lymph-node negative primary breast cancer patients with known outcome [11]. Since the patients in this retrospective cohort did not receive any kind of adjuvant systemic therapy after surgery, this study addresses the pure prognostic value, without the potential confounding effects of therapy. Hierarchical clustering showed 4 groups with a distinct landscape of abnormalities. We elaborated on this by devising a CIN-score for individual samples. This CIN-score was determined by enumerating the total number of genomic areas that showed a concordant loss or gain according to the copy number of the SNPs

present therein. Given the well-known clinical relevance of the diverse breast cancer subtypes such as the five molecular subtypes [12], estrogen receptor (ER)-positive and –negative disease [11, 13-15] as well as triple-negative disease (negative for ER, progesterone receptor (PR) and her2/neu), we analyzed the CIN-score and its association with prognosis in these pathological and molecular subtypes. Our data provide evidence for a highly unstable genome in ER-negative, triple-negative and basal-like tumors, whereas prognostic relevance is only present in patients with ER-positive breast cancer and of the luminal B and her2/neu subtype. Furthermore, to understand the biology related to CIN, we tested whether biologically meaningful pathways are related to CIN in ER-positive breast cancer, the cancer subtype in which CIN was strongly prognostic.

## **Materials and Methods**

### *Patient and tumor material*

Primary tumors of lymph node–negative breast cancer patients were stored in liquid nitrogen at the Erasmus Medical Center. 344 tumors of which Affymetrix U133a gene expression data were available [15-16] were initially selected for subsequent analysis of DNA copy number alteration using Affymetrix GeneChip Mapping 100K Array Set (Affymetrix, Santa Clara, CA, USA). Good quality DNA was available for hybridization of 328 of the 344 tumors and reliable SNP calls could be obtained from 313 of these tumors, which were included in the previous analysis of copy number alteration [11] and in the present study on CIN. All 313 patients were treated from 1980 to 1995 with the same guidelines for local primary treatment, implying that none of these patients did receive any systemic (neo)adjuvant therapy. Patients who

first developed a local recurrence before distant metastasis were not included to avoid the possibility that the distant metastasis later on could have originated from the local recurrence. The study was approved by the medical ethics committee of Erasmus MC (MEC 02.953) and conducted in accordance to the code of conduct of Federation of Medical Scientific Societies in the Netherlands ([www.fmwv.nl](http://www.fmwv.nl)). Whenever possible, we adhered to the Reporting Recommendations for Tumor Marker Prognostic Studies REMARK [17]. Full clinical and pathological characteristics are described elsewhere [11]. Briefly, mean age at surgery was 54 years (standard deviation 12y), 49% of patients were premenopausal, 201 tumors were ER positive (according to expression measured on microarray [18]) and 112 ER negative. Of all surviving patients (n=220), median follow-up time was 99 months (range 20-169 months). Of the 313 included patients, 114 (36%) developed a distant metastasis, of whom 77 were ER-positive and 37 ER-negative.

#### *DNA copy number analysis and CIN-score*

All tumors were confirmed to have sufficient (>70%) tumor content. The CIN-score was established using SNPs present on the Affymetrix GeneChip Mapping 100K Array Set (Affymetrix, Santa Clara, CA, USA). Copy number data was obtained using previously described methods [11] and is available in the National Center for Biotechnology Information/Gene Expression Omnibus database (GEO) entry GSE10099. For each SNP, the estimate diploid copy number was established [11] and the distance of the actual copy number for that SNP to the estimate was calculated for each tumor. Hierarchical clustering of copy number data was performed using Nexus (BioDiscovery, El Segundo, CA, USA). For the CIN-score, a SNP was called gain or loss, if the copy number was 0.4 higher or lower than the

estimate diploid copy number, respectively. This cut-off was established using Nexus by visual inspection of many samples showing a gain of a complete chromosomal arm. Next, for each patient, the number of genomic segments showing CIN was counted. A CIN-area met these criteria; presence of at least 2 SNPs and sequential SNPs lied within 100 kb distance. Then the status of the SNPs was classified as either 'gain', 'loss' or 'normal'. Due to the inherited noise within these arrays, one normal-call was allowed within two SNPs having a gain (or loss) call. Thus, two sequential SNPs with a 'normal' call signaled the end of that area. Likewise if a 'gain' SNP followed a 'loss' SNP, or vice versa. The total number of affected areas in a tumor genome was considered a reliable measure for CIN. We tested if the CIN-score was influenced by the selected criteria. First, the maximum of 100 kb distance between 2 sequential SNPs was omitted. The resulting CIN-scores had a Pearson correlation of 0.946 with the original score. Similarly, when including regions which had only a single SNP showing gain or loss, the Pearson correlation with the original score was 0.998. Thus, irrespective of the inclusion criteria, highly consistent CIN-scores were retrieved. In our view, the CIN-score based on including the 100 kb distance criterion and exclusion of single SNP regions, is less likely to be overly optimistic and was therefore used in this study.

#### *Gene expression data*

mRNA expression data available for this entire cohort [15-16] were used to establish ER, PR and ERBB2 status, to determine the intrinsic molecular subtypes [12, 19], to associate genes with CIN and to conduct pathway analysis. Gene Set Enrichment Analysis (GSEA) [20] was used to identify which of the public gene profiles as

described in the database (<http://www.broad.mit.edu/gsea/msigdb/index.jsp>) complemented with the prognostic breast cancer gene sets [9, 12, 14-15, 21-26], are associated with the CIN-scores. Using default parameters, ER-positive samples with the top 25% CIN-scores were compared to the bottom 25% CIN-scores of ER-positive samples (50 samples in each group). Next, the genes within the public profiles were assessed to see if they were significantly enriched in the top (or bottom) ranked order of significant genes. The Global Test [27] program was used (version 4.2.0) to associate Biocarta pathways (<http://www.biocarta.com>) to the same sample-groups as were used in the GSEA analysis. SAM analysis [28] was used to identify differentially expressed genes between the samples with the highest and lowest CIN-scores (50 samples in each group for ER-positive tumors and 28 samples in each group for ER-negative tumors). Independent datasets [25, 29] were downloaded from GEO (GSE11121 and GSE2990) and processed with the same parameters as our own dataset [15-16]. The top 40 significant genes (20 genes higher expressed in high CIN and 20 genes higher expressed in low CIN tumors) were selected and were used in BRB-ArrayTools (v3.7.1 developed by Dr. Richard Simon and the BRB-ArrayTools Development Team) using the Compound Covariate Predictor to predict the status of the samples in the external datasets as “high CIN” or “low CIN”. Our samples were used as training data.

STATA, release 10.1 (StataCorp, Texas, USA) was used to calculate the Cox proportional HR. MFS was used as endpoint and was defined as the first detection of a distant metastasis that was confirmed after symptoms were reported by the patient or when clinical signs of metastasis were detected at regular follow-up. Each of the subgroups showed a sufficient number of events (more than 20) to use the HR tests,



except for the normal-like subtype (7 events). Log-rank tests for trend were used to test the equality of survivor functions across more than two groups. A two-sided  $P$  value of  $< .05$  was considered statistically significant.

## Results

### CIN distribution among breast cancer subtypes

Unsupervised hierarchical clustering of SNP copy number data of 313 breast cancer patients identified 4 main groups, of which the frequency of copy number alterations are shown in Figure 1. Most striking features of the clustered groups were the large percentage of samples showing a gain of either chromosome (chr) 8q (first cluster), chr 1q (cluster 4) or both (second cluster-group). Association of these cluster-groups with diverse subtypes shows a highly variable division; e.g. 90% (57 out of 63) of samples in cluster 4 are ER-positive, or 73% (16 out of 22) of the normal-like subtype are present in cluster 3, the cluster with the samples showing very few abnormalities (full data in Table 1). Figure 1 also makes clear that a large variety in chromosomal instabilities is present in the different samples (see supporting information; Online Resource Fig. 1 for 2 extreme examples). We found no significant differences in prognosis between the 4 clustergroups (Online Resource Fig. 2). We next established a CIN-score for each individual sample. The CIN-score in our cohort ranged from 66 to 3380 (median 907). ER-negative tumors, as well as the triple-negative tumors, had significantly higher CIN-scores than those that were positive for one or more of these receptors (Mann–Whitney test  $P < 0.0001$  in both analyses) (Fig. 2A and B). Among the intrinsic subtypes, basal and luminal B cancers had higher CIN-scores (Fig. 2C). A Kruskal-Wallis test for difference in

median levels between all the molecular subgroups was significant,  $P < 0.0001$ . Furthermore, post-hoc pairwise comparisons (Bonferroni corrected for multiple testing) of the subtypes showed a significant difference in median CIN-scores of basal vs. luminal B samples ( $P = 0.001$ ). In fact, all pairwise comparisons were significant ( $P = 0.005$  or lower), except for her2/neu vs. luminal A, her2/neu vs. luminal B, and luminal A vs. normal-like.

### CIN and prognosis

Next, we related the CIN-score to prognosis in the various subtypes. Of the 313 patients, 114 (36%) developed a distant metastasis and were counted as failures in the analysis of distant metastasis-free survival (MFS). Log-transformed CIN-scores as continuous variable in Cox regression analysis were associated with MFS and showed a HR of 2.16 (95% Confidence Interval (CI): 1.53-3.04,  $P < 0.0001$ ) for ER-positive and a HR of 1.05 (95% CI: 0.70-1.58,  $P = 0.81$ ) for ER-negative patients. Kaplan-Meier plots, using three equally sized quantiles of the CIN-score, showed a significant difference in MFS-curves in ER-positive patients (Fig. 3a), but not in ER-negative patients (Fig. 3b). Analysis within other subgroups of patients showed an association of CIN-scores as continuous variable with MFS in luminal B (HR: 2.59, CI: 1.37-4.88,  $P = 0.003$ ) and in the her2/neu subtype (HR: 2.47, CI: 1.15-5.30,  $P = 0.02$ ) while within basal, luminal A, normal-like and triple-negative breast cancer no such relation was seen (HRs of the latter 4 subgroups: 1.06, 1.55, 0.99 and 1.01, respectively; all statistically not significant). Analysis of the survival curves within the subtypes using the CIN-quantiles suffered from low numbers of patients but similar trends as in the ER subgroups were seen (Fig. 3c and 3d). Further analysis also revealed that with regard to prognosis, CIN probably

explains the difference between the luminal A and luminal B breast cancer since the difference in MFS between these groups ( $P = 0.01$ ) was no longer present after correcting for CIN-score (Online Resource Fig. 3) ( $P = 1.00$ ). To validate our association of CIN with prognosis in ER-positive breast cancer, we employed gene expression data of two external datasets [25, 29] with in total 254 ER-positive and 57 ER-negative tumors, from patients who were all lymph-node negative and did not receive systemic treatment after surgery. We developed a 40-gene classifier in our own ER-positive cohort, which, using a Compound Covariate Predictor [30] showed a 86% sensitivity and 78% specificity in correctly assigning “low CIN” samples. Next, this same classifier was used to predict the status of the ER-positive samples in the external datasets as “high CIN” or “low CIN”. The CIN-signature significantly stratifies the ER-positive samples in the independent datasets according to outcome ( $P < 0.001$ , Fig. 3e and 3f). The same procedure for the ER-negative samples yielded no significant log-rank test ( $P = 0.23$ , Online Resource Fig. 4).

To further explore the biology associated with the CIN process, we correlated transcriptome expression also available from this cohort [15-16] with the CIN-score. For this, gene expression patterns were compared between ER-positive samples with the highest CIN-score versus those with the lowest CIN-scores. Pathway analyses were performed using GSEA [20] and global testing [27]. These analyses showed association of CIN with gene sets related to the genomic grade of breast cancer, a CIN-profile derived from a diverse set of cancer cell lines, gene sets related to outcome in breast, multiple myeloma and gastric cancer, a wound response signature and gene sets more generally related to cell cycle progression and those involved in pyrimidine metabolism (see Online Resource Table 1 and 2 for

full details). In addition, genes linked to CIN were enriched for E2F transcription factor binding sites. Pathways negatively associated with CIN were related to cellular differentiation, to extrinsically induced cell death and to T-cell function. The strong relation between CIN and the genomic grade in our cohort prompted us to investigate, on a beforehand unplanned analysis, the influence of CIN after grouping our samples according to the genomic grade signature. It appeared that the CIN-score and genomic grade were independently significant in the analysis for MFS in ER-positive patients in a multivariate analysis (HR: 2.29, 95% CI: 1.33-3.90,  $P = 0.002$  and HR: 1.90, 95% CI: 1.19-3.01,  $P = 0.007$  for CIN and genomic grade, respectively). The CIN-score was significantly associated with MFS in the good-grade group (logrank  $P = 0.004$ ), but not in the poor-grade group, although the MFS curves based on CIN also in this latter group tended to separate (Online Resource Fig. 5).

## **Discussion.**

Chromosomal instability is a long described concept in tumor biology [1]. In breast cancer, more recent publications describe results based on aCGH copy number data [2-3, 8, 31]. As shown in Figure 1, we observed four major groups of instability profiles; two of these were formerly described by Andre et al. [8] who used 106 samples. They described a group with very few abnormalities (our clustergroup 3) and two groups of samples, which both had the characteristic gains of chr 1q and 8q (our second clustergroup). We identified two additional groups, one characterized by many samples with a gain in chr 1q but without a gain of chr 8q and a group with the opposite pattern. Although the ER-status of the tumors or the intrinsic subtypes do

not fall exclusively in one of the cluster groups, there are clear non-random patterns. For example group 4 contains 63 samples of which 57 are ER-positive, in which 90% of which have a gain of the q-arm of chr 1, and almost 50% of samples a gain of 16p. This suggests a very different etiology for these cancers than for ER-negative tumors, where only 19% of cases fall in cluster 2 or 4 (which show the chr 1q gain). Since other studies already described specific genomic aberrations which are implicated in the different subtypes of breast cancer [2-3, 5, 8, 31], and we knew from our recent work that relatively few specific regions with altered copy number were found to be prognostic (12), our prime objective based on the large variability among samples was to assess the clinical significance of the number of chromosomal abnormalities in individual tumors as well as the biology related to this phenomenon. The 4 cluster groups already showed a trend with 118 tumors showing very few abnormalities to tumors having genomes that were greatly rearranged. These observations are backed by the results of a recent publication [10], which described genome-wide completely paired-end sequenced breast cancer genomes of cell lines and primary breast tumors; some samples show hardly any events, while others show many aberrations. Thus, notwithstanding specific abnormalities, CIN in itself is not necessarily a common denominator in all breast cancer cases, confirming the notorious heterogeneous nature of breast cancer. The measure for CIN we used here for individual samples showed highly variable CIN-scores and we observed a clear uneven distribution within the different predefined breast cancer subtypes. Significantly higher CIN-scores for ER-negative and triple-negative samples were found, indicating that CIN is a more common event in receptor negative breast cancers. This suggests that this subtype is more prone to accumulation of genomic instabilities. Higher CIN-scores were also observed in basal and luminal B samples,

which corroborates previous data [2-3], though these were based on low-density CGH data of in total 234 patients. Collectively, accumulating evidence suggests that the breast cancer cell type influences the copy number landscape of a breast cancer. Apparently, not only specific genomic aberrations [3, 11, 32], e.g. ERBB2 amplification, are linked to a poor outcome in breast cancer, but also the sheer number of chromosomal instabilities is clearly associated with prognosis in ER-positive patients and in patients with a her2/neu or luminal B subtype (Fig. 3). These results confirm but also add substantially to the findings of Carter et al. [9] who first reported on the clinical significance of CIN in breast cancer. Literature pertaining CIN and prognosis based on copy number data is scarce. Two publicly available aCGH breast cancer datasets were investigated; the dataset by Chin et al. [3] has 174 samples, but only 11 of these were eligible (ER-positive, lymph-node negative, no adjuvant hormonal or chemotherapy treated patients), and none of these patients relapsed. The other dataset was from Chin et al. [31] describing 171 patients, of whom 60 were eligible, but only 8 patients showed a relapse. We deemed both these studies not robust enough to validate our results. To be able to confirm our observations in other datasets, we therefore employed two external gene-expression datasets with sufficient cases and appropriate follow-up [25, 29]. In these data, the pure prognostic relevance of CIN inferred from associated differentially expressed genes in ER-positive cases was clearly confirmed (Fig. 3). Pathway analysis point to an excess of proliferation, including overrepresented E2F directed transcriptional regulation and increased nucleic acid synthesis, and lack of apoptotic and immune cell responses as prominent features in ER-positive breast cancer. These pathways are considered the traditional hallmarks of unconstrained growth of cells [33].

The genomic grade signature [25] was found to be the most significant signature associated with CIN, suggesting a strong interaction between CIN and histological tumor grade. Indeed, within our cohort CIN-scores were significantly higher in the poor grade group (Online Resource Fig. 6), an observation also recently reported by Ellsworth et al. [34], based on a limited amounts of chromosomal areas (n=26). Since our cohort is a retrospective cohort obtained from various peripheral hospitals we do not have all histological grading available according to current standards. Therefore, we clustered our samples according to the genomic grade profile. Interestingly, the CIN-score and genomic grade were, in a multivariate MFS analysis, independent of each other significantly associated with MFS in ER-positive patients. Of note, since we have analyzed the genomic grade instead of the histopathological grade, these data should be interpreted with caution. The surprising observation that CIN is prognostic in good genomic grade, but not in luminal A samples, while most luminal A samples have a good grade, may be explained by the fact that luminal A samples are not exclusively of good grade; in our cohort, 19% of luminal A samples are of poor grade, while of all good-grade samples, 47% is of luminal A type. The differences in luminal A / good-grade cohorts probably explains the difference in prognostic value of CIN in these groups. Furthermore, it probably also implies that intrinsic gene expression differences between luminal A and B breast cancer capture the biological and prognostic differences in chromosomal instability quite well. The observation that CIN may be able to discern differences within the good-grade group was also reported by Carter et al. [9]. Their study showed that CIN was able to distinguish outcome in grade 2 tumors in three different datasets (total of 512 patients) and also in grade 1 tumors in a dataset of 161 patients. CIN was not prognostic in the grade 3 samples of the different datasets. This further supports the

notion that histopathological grading of well-differentiated tumors in particular can be improved by molecular tools such as gene expression signatures [9, 25] or as described here, the use of SNP copy number data. Taken together, the data suggest that an ER-positive tumor cell with a poor genomic integrity may still phenotypically appear as well differentiated. But despite that appearance, such a tumor is likely to progress. Poorly differentiated tumor cells on the other hand virtually all display an unstable genome.

In conclusion, we confirm that CIN is a prominent event in many breast cancer samples and our observations add to the current paradigm of the role of chromosomal instability in breast cancer. From our work it becomes clear that ER-negative tumors as well as the basal/triple-negative tumors have a highly unstable genome, but unanticipated, this instability does not seem to aggravate prognosis within ER-negative patients. This signifies an association of CIN with the processes regulating the integrity of the genome possibly in the origin, but not in the progression of the disease. In ER-positive, her2/neu and luminal B tumors the genome is to a lesser degree unstable, but if CIN is present, prognosis is poor. Obviously, the rather extensive methodology to obtain the CIN-score described in this study calls for easier assayable markers to query CIN if it were to be used in diagnostic routine. Nevertheless, assessment of CIN is a useful tool to gain insight into breast cancer biology and may further contribute to a more tailored treatment approach of breast cancer patients.



**Acknowledgments:**

The work was supported by Veridex LLC for tissue processing and isolating RNA and DNA for microarray analysis as well as by a research grant from the Netherlands Genomics Initiative (NGI) / Netherlands Organization for Scientific Research (NWO). This organization had no role in study design, analysis of data or preparation of the manuscript.

**Conflict of interest:**

Authors Y.Zhang and Y.Wang are employees of, and hold stock in Veridex LLC, which financially supported this work for tissue processing and microarray analysis. The other authors declare no conflict of interest.

## References:

1. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. *Nature* 396(6712):643-649.
2. Bergamaschi A, Kim YH, Wang P, Sorlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Borresen-Dale AL, Pollack JR (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* 45(11):1033-1040.
3. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiology. *Cancer Cell* 10(6):529-541.
4. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM (1994) Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci U S A* 91(6):2156-2160.
5. Loo LW, Grove DI, Williams EM, Neal CL, Cousens LA, Schubert EL, Holcomb IN, Massa HF, Glogovac J, Li CI, Malone KE, Daling JR, Delrow JJ, Trask BJ, Hsu L, Porter PL (2004) Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res* 64(23):8541-8549.
6. Naylor TL, Greshock J, Wang Y, Colligon T, Yu QC, Clemmer V, Zaks TZ, Weber BL (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res* 7(6):R1186-1198.
7. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23(1):41-46.
8. Andre F, Job B, Dessen P, Tordai A, Michiels S, Liedtke C, Richon C, Yan K, Wang B, Vassal G, Delalogue S, Hortobagyi GN, Symmans WF, Lazar V, Pusztai L (2009) Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin Cancer Res* 15(2):441-451.
9. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z (2006) A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet* 38(9):1043-1048.
10. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, Sieuwerts AM, Martens JW, Silver DP, Langerod A, Russnes HE, Foekens JA, Reis-Filho JS, van 't Veer L, Richardson AL, Borresen-Dale AL, Campbell PJ, Futreal PA, Stratton MR (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 462(7276):1005-1010.
11. Zhang Y, Martens JW, Yu JX, Jiang J, Sieuwerts AM, Smid M, Klijn JG, Wang Y, Foekens JA (2009) Copy number alterations that predict metastatic capability of human breast cancer. *Cancer Res* 69(9):3795-3801.
12. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747-752.

13. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 61(16):5979-5984.
14. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernardis R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530-536.
15. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365(9460):671-679.
16. Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, Klijn JG, Wang Y, Foekens JA (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer* 7:182.
17. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, Statistics Subcommittee of the NCI EWGoCD (2005) Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 97(16):1180-1184.
18. Foekens JA, Atkins D, Zhang Y, Sweep FC, Harbeck N, Paradiso A, Cufer T, Sieuwerts AM, Talantov D, Span PN, Tjan-Heijnen VC, Zito AF, Specht K, Hoefler H, Golouh R, Schittulli F, Schmitt M, Beex LV, Klijn JG, Wang Y (2006) Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J Clin Oncol* 24(11):1665-1671.
19. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, Foekens JA, Martens JW (2008) Subtypes of breast cancer show preferential site of relapse. *Cancer Res* 68(9):3108-3114.
20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.
21. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2(2):E7.
22. Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, Salim A, Wang Y, Kristensen GB, Helland A, Borresen-Dale AL, Giaccia A, Longaker MT, Hastie T, Yang GP, van de Vijver MJ, Brown PO (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med* 3(3):e47.
23. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N Engl J Med* 356(3):217-226.
24. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351(27):2817-2826.
25. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98(4):262-272.

26. Winter SC, Buffa FM, Silva P, Miller C, Valentine HR, Turley H, Shah KA, Cox GJ, Corbridge RJ, Homer JJ, Musgrove B, Slevin N, Sloan P, Price P, West CM, Harris AL (2007) Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res* 67(7):3441-3449.
27. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20(1):93-99.
28. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9):5116-5121.
29. Schmidt M, Bohm D, von Torne C, Steiner E, Puhl A, Pilch H, Lehr HA, Hengstler JG, Kolbl H, Gehrman M (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* 68(13):5405-5413.
30. Radmacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. *J Comput Biol* 9(3):505-511.
31. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavaré S, Brenton JD, Ylstra B, Caldas C (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* 8(10):R215.
32. Al-Kuraya K, Schraml P, Torhorst J, Tapia C, Zaharieva B, Novotny H, Spichtin H, Maurer R, Mirlacher M, Kochli O, Zuber M, Dieterich H, Mross F, Wilber K, Simon R, Sauter G (2004) Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer Res* 64(23):8534-8540.
33. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. *Cell* 100(1):57-70.
34. Ellsworth RE, Hooke JA, Love B, Kane JL, Patney HL, Ellsworth DL, Shriver CD (2008) Correlation of levels and patterns of genomic instability with histological grading of invasive breast tumors. *Breast Cancer Res Treat* 107(2):259-265.

Legends:

**Fig. 1 Chromosomal instability patterns of 313 primary breast tumors**

Frequency of gains and losses among hierarchically clustered groups are plotted from chromosome 1 to X (left to right). Green and red indicate gain and loss, respectively.

**Fig. 2 Association of CIN with breast cancer subtypes**

CIN-scores according to ER status (A), triple-negative status (B), and molecular subtype (C) are shown. Boxes show 1<sup>st</sup>-3<sup>rd</sup> quartile with the median marked as a horizontal line, notches show the 95% confidence interval of the median.

**Fig. 3 Kaplan-Meier distant metastasis-free survival curves**

Top panel and bottom left graph: CIN-scores were divided into three equal quantiles; the red, orange and green line indicates samples with highest, intermediate and lowest CIN-scores, respectively. The bottom middle and right graph depict predicted CIN-scores in independent datasets, where red indicates a predicted high CIN-score, green a predicted low CIN-score.

**Table 1 Breast cancer subtypes according to copy number cluster groups.**

Number and percentage of samples per cluster group. Numbered groups appear top to bottom in figure 1.

All	Cluster group				P*
	1	2	3	4	

n	313	92	40	118	63	
ER						
negative	112 (36%)	49 (53%)	15 (38%)	42 (36%)	6 (10%)	
positive	201 (64%)	43 (47%)	25 (62%)	76 (64%)	57 (90%)	<0.0001
Triple-negative						
yes	68 (22%)	34 (37%)	10 (25%)	20 (17%)	4 (6%)	
no	245 (78%)	58 (63%)	30 (75%)	98 (83%)	59 (94%)	<0.0001
Intrinsic subtype						
basal	77 (25%)	37 (40%)	14 (35%)	22 (19%)	4 (6%)	
her2	62 (20%)	19 (21%)	8 (20%)	27 (23%)	8 (13%)	
luminal A	85 (27%)	16 (17%)	8 (20%)	40 (34%)	21 (33%)	
luminal B	67 (21%)	16 (17%)	9 (23%)	13 (11%)	29 (46%)	
normal	22 (7%)	4 (4%)	1 (3%)	16 (14%)	1 (2%)	<0.0001

\**P*-value according to the  $\text{CHI}^2$ -distribution.