Genome-wide Copy Number Alterations in Subtypes of Invasive Breast Cancers in Young White and African American Women

Lenora WM Loo1*, Yinghui Wang2, Erin M Flynn1, Mary Jo Lund4,8, Erin J Aiello Bowles5, Diana SM Buist5, Jonathan M Liff4, Elaine W Flagg6, Ralph J Coates7, J William Eley8, Li Hsu2, Peggy L Porter1,2,3

1 Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA
2 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA
3 Department of Pathology, University of Washington, Seattle, WA
4 Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA
5 Group Health Research Institute, Group Health Cooperative, Seattle, WA
6 Division of STD Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA
7 National Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA
8 Hematology and Medical Oncology, Emory University School of Medicine, Atlanta, GA
*Present address: Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, HI

Running Title: Genomic alterations in breast cancer subtypes of young women
Key Words: breast cancer; triple negative; genomic alteration; array comparative genomic hybridization; young women

Send all correspondence to:
Peggy L. Porter, MD
Member, Human Biology and Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N C1-015
Seattle, WA  98109
Ph: 206-667-3751
FAX: 206-667-5815
Email: pporter@fhcrc.org
Abstract:

Genomic copy number alterations (CNA) are common in breast cancer. Identifying characteristic CNAs associated with specific breast cancer subtypes is a critical step in defining potential mechanisms of disease initiation and progression. We used genome-wide array comparative genomic hybridization (aCGH) to identify distinctive CNAs in breast cancer subtypes from 259 young (diagnosed with breast cancer at <55 years) African American (AA) and Caucasian American (CA) women originally enrolled in a larger population-based study. We compared the average frequency of CNAs across the whole genome for each breast tumor subtype and found that estrogen receptor (ER)-negative tumors had a higher average frequency of genome-wide gain \( (p<0.0001) \) and loss \( (p=0.02) \) compared to ER-positive tumors. Triple negative (TN) tumors had a higher average frequency of genome-wide gain \( (p<0.0001) \) and loss \( (p=0.003) \) than non-TN tumors. No significant difference in CNA frequency was observed between HER2-positive and negative tumors. We also identified previously unreported recurrent CNAs (frequency >40%) for TN breast tumors at 10q, 11p, 11q, 16q, 20p and 20q. In addition, we report CNAs that differ in frequency between TN breast tumors of AA and CA women. This is of particular relevance because TN breast cancer is associated with higher mortality and young AA women have higher rates of TN breast tumors compared to CA women. These data support the possibility that higher overall frequency of genomic alteration events as well as specific focal CNAs in TN breast tumors might contribute in part to the poor breast cancer prognosis for young AA women.
Introduction:

Breast cancer is a heterogeneous disease consisting of five major breast tumor subtypes, basal, human epidermal growth factor receptor 2 (HER2)-expressing/estrogen receptor (ER)-negative, luminal A, luminal B and normal-like. These subtypes have been shown to have distinct expression patterns, based on microarray profiling, and clinical outcomes [1-4]. Protein expression of ER, progesterone receptor (PR), and HER2, detected by standard immunohistochemical (IHC) assays, has been used to approximate the gene expression subtypes. Triple negative (TN) breast cancer (ER-, PR-, HER2-) overlaps with basal cancer, ER-positive cancers closely define luminal cancers and the ER-negative, PR-negative and HER+ cancers approximate the HER2-expressing subtype.

The TN breast cancer subtype is not synonymous with the basal phenotype and expression of basal markers such as epidermal growth factor receptor (EGFR) and/or basal cytokeratins highlight the heterogeneity of the TN grouping [5]. However, the TN subset of breast cancers is highly enriched for the basal phenotype [6,7,1,3]. As is the case for basal cancers, TN tumors arise at an earlier age than non-TN cancers, and are almost exclusively high grade tumors [8].

One of the most striking findings related to TN breast cancer is that African American (AA) women exhibit an almost two-fold higher prevalence of this breast cancer subtype than Caucasian American (CA) women. The prevalence of TN breast cancer has been reported as high as 40% in AA women [9,1,10]. Additionally, TN incidence rates have now been reported and are also nearly twice as high among AA compared to CA women [11]. Even though AA have an overall lower incidence of breast cancer than CA women, the higher likelihood of developing TN breast cancer might contribute to the higher mortality from breast cancer experienced by AA women compared with CA women [12-17,10].

In addition to the differentiation of breast cancer subtypes by gene expression profiling,
genome-wide array comparative genomic hybridization (CGH) techniques have demonstrated that breast cancer subtypes are associated with characteristic copy number alteration (CNA) profiles [18-21]. These characteristic genomic alterations serve as useful markers for subtype classification and can be applied towards the identification of critical genes that are consistently lost or gained in the specific subtypes to impact tumor biology.

The acquisition of genomic alterations is a critical event in cancer formation, potentially impacting the expression patterns of individual genes or entire biological pathways. For this study of young women we used genome-wide array CGH to identify CNAs that are high frequency and/or different in frequency between breast cancer subtypes defined by protein expression of ER, PR and HER2. We also compared CNA profiles of TN breast cancers in young CA and AA women to identify biological factors that might contribute to the poor breast cancer prognosis for young AA women.

**Materials and Methods:**

**Tumor Specimens**

Primary breast tumors were from AA or CA women ages 20-54 years previously enrolled in the population-based Atlanta Women’s Interview Study of Health (WISH) cohort, which included 950 women diagnosed with breast cancer between 1990 and 1992 [22]. When diagnosed, these women were residents of a 3-county metropolitan region of Atlanta, Georgia. The 259 breast tumors analyzed in the present study are the subset of the women in the WISH study with sufficient tumor tissue for testing. All tumors were reviewed and scored by an expert pathologist (P.L.P.) for tumor grade, ER, PR, and HER2 status [23].

**Flow Cytometry Cell Sorting and DNA Extraction and Labeling**
Flow cytometry was performed on macro-dissected and dissociated formalin-fixed, paraffin-embedded tumor samples to enrich for tumor cells by removing contaminating stromal and lymphocytic cells as previously described [24].

Genomic DNA was extracted from the flow sorted tumor cells (minimum of 100,000 cells per tumor) as previously described [25]. We quantified tumor genomic DNA by real-time PCR (Applied Biosystems, Foster City, CA) using two chromosome 2 specific probes at 2p25.3 (29,907–30,162) and 2q31.1 (21,407,882–21,408,181) with normal human female genomic DNA (Promega, Madison, WI) as the reference.

Whole-genome amplified and labeled samples were prepared as previously reported [21]. Briefly, ten nanograms each, of tumor and DpnII digested normal female reference DNA (Promega, Madison, WI), were random amplified and labeled with a Cy5- or Cy3-labeled primer, respectively, according to the method of Lieb et al., with modifications [26]. Labeled PCR products purified and combined with blocking agents (50 µg of human Cot-1 DNA and 100 µg of yeast tRNA) and hybridized to the microarray.

Array CGH

The array consists of 4320 human bacterial artificial chromosomes (BAC) with a median spacing of 413 kb when pericentric heterochromatic regions and the short arms of acrocentric chromosomes are excluded [21]. BAC clone locations are based on NCBI Build 36.1 of the human genome.

The Cy3- and Cy5- labeled genomic DNA and blocking agents were hybridized to the BAC array as described previously [21]. Arrays were scanned with a GenePix 4000A scanner (Axon Instruments Inc., Union City, CA); fluorescence data were processed with GenePix 3.0 image analysis software (Axon Instruments, Inc.). For each spot, log₂ratio=log₂(Cy5/Cy3) and average log₂intensity=[log₂(Cy5) + log₂(Cy3)]/2 were calculated, where Cy5 and Cy3 refer to the median foreground fluorescent signals of the tumor and reference DNA, respectively. The
log₂ ratios on each array were normalized and corrected for intensity-based location adjustment with a block-level lowess algorithm [27].

**Statistical Analysis**

Normalized aCGH data were processed using wavelet along the genome [28]. The processed aCGH values were then categorized into copy number loss, no change, and gain events using the cut-off log₂ ratio -0.34 and 0.38, for loss and gain respectively, where the cut-off values were chosen based on X-chromosome titration experiments, previously reported [21].

Sampling weights were incorporated into the analysis based on the larger cohort of 950 cases for analysis of the 259 cases that were analyzed by aCGH. (See Supplemental Data 1 for additional details of statistical analysis; Supplemental Data Table 1). We calculated the weighted average overall genome-wide frequencies of copy number gain or loss by race (AA/CA) and the following tumor subtypes: ER status (positive/negative), triple negative (yes/no), and HER2 status (positive/negative), where the genome-wide copy number gain (or loss) for a tumor was defined as number of clones showing gains (or losses) divided by the total number of clones. To adjust for possible confounding effects of age and stage, weighted multivariable logistic regression was performed to examine whether each comparison group differs in gains and losses at each of the 4320 clones, respectively. Given some clones may have no or few events of gains or losses, the \( p \)-values based on asymptotic distributions of the test statistics would be biased. To correct for this bias, the bootstrap method was used to obtain exact \( p \)-values. A total of 1000 bootstrap samples were used for each comparison.

Hierarchical clustering was performed using clones that show statistical significance in any of the comparisons to identify whether subtypes of tumors would cluster based on the profiles of copy number alterations. For the heatmap clustering, we used the euclidian distance as the dissimilarity function and complete linkage.
All the analyses were done using statistical software R version 2.6.0. The wavelet smoothing required package of 'waveslim'; the weighted logistic regression required package of 'survey' (http://www.r-project.org/). Throughout the paper, a $p$-value < 0.05 is considered statistically significant.

**Results:**

**Breast Tumor Characteristics of Young African American and Caucasian American Women**

The individual and tumor characteristics of the 259 women (AA n=53 and CA n=206) such as age, vital status, ER, PR, HER2 expression status, grade, and stage are shown for all tumors and separately by race (Table 1). The samples in the present study, as observed in previous reports of this population-based Atlanta WISH cohort, show a significant racial difference in the distribution of tumors based on ER, PR, and HER2 expression status, as well as vital status, stage and grade for young AA compared to CA women [10,23,29] and, as described in the methods and Supplemental Data, a weighted analysis was performed to more accurately reflect the make-up of the original cohort.

**Genome-wide Copy Number Alterations in Subtypes of Breast Cancer and Racial Groups**

We compared the average genome-wide frequency of copy number gain and loss for individual tumors of IHC-defined breast cancer subtypes (Table 2). In the comparison between ER-positive and ER-negative tumors, ER-negative tumors had a significantly higher frequency of both copy number gain (gain: 6.9% versus 4.9% $p<0.0001$) and loss (8.5% versus 7.3% $p=0.02$) events. TN breast tumors similarly had a significantly higher overall frequency of both genome-wide copy number gain (7.3% versus 5.2%; $p<0.0001$) and loss (8.9% versus 7.4%; $p=0.003$) than non-TN tumors, respectively. There was no statistically significant difference in the overall genome-wide frequency of copy number gain (5.6% versus 5.8%; $p=0.68$) and loss
(7.3% versus 7.9%; \(p=0.37\)) when comparing HER2-positive to HER-2-negative tumors, respectively. When we compared overall frequency of genomic alterations in breast tumors of AA to CA women, we observed higher frequencies of both gain (6.9% versus 5.2%; \(p=0.0002\)) and loss (9.0% versus 7.3%; \(p=0.003\)) in AA than CA women, respectively. This difference is consistent with the higher percentages of ER negative and TN breast tumors in AA compared to CA women (Table 1), suggesting that breast tumor subtypes differ with respect to frequency of genome-wide alteration events.

**Frequent Copy Number Alterations in Breast Cancer Subtypes and Racial Groups**

To identify the specific genomic regions that were more frequently altered (copy number gain or loss) in the breast cancer subtypes, we classified CNAs as a “high-frequency” event if the gain or loss events for the specific probe on the array occurred in over 40% of the tumors in the specified subtype. We mapped these probes indicating high-frequency CNAs for the specific breast cancer subtypes to chromosome arms (Table 3) and cytogenetic bands (Supplemental Data Table 2) to profile the distribution of high frequency events. We identified high-frequency CNAs that were shared by the subtypes, as well as high-frequency CNAs that differed by subtype. The high frequency events common to all subtypes included gain events on chromosomes 1q and 8q, and high-frequency loss events on chromosomes 8p, 10q, 11q, 12q, and 16q. Subtype specific high-frequency CNAs included copy number loss events on 4p and 11p, which were observed in ER-negative, but not ER-positive tumors. Interestingly, HER2-positive tumors had the widest distribution of high-frequency loss events across the genome, i.e., the most chromosome arms with high-frequency loss events. We also observed high-frequency loss on 13q and 20q in TN breast cancers that were not observed in other non-TN tumors. Taken together, these results support the presence of high-frequency subtype-specific CNA events.

**Differential Copy Number Alterations in Breast Cancer Subtypes**
In addition to the CNA frequency differences at the genome-wide and chromosomal arm level, we identified subtype-specific CNA events (gain and loss) at individual BAC clones. Differential CNAs (CNAs that differ in frequency between groups (p<0.05)) by individual BAC clones were identified based on comparisons of age- and stage-adjusted tumors in each comparison and restricted to clones with the difference of CNA frequency ≥0.20 (Table 4 and Supplemental Data Table 3). In the comparison between ER-positive (n=154) and ER-negative (n=105) breast cancers, we found 90 (79 loss and 11 gain) differential CNAs. ER-positive tumors were characterized by differential gain events on 1q, and loss events on 11q and 16q, whereas, ER-negative tumors were characterized by differential gain events on 8q and 10p, and loss events on 3p, 4p, 5q, 9q, 10q, 12q, 13q, and 14q (Figure 1). In HER2-positive (n=35) vs. HER2-negative (n=224; including both ER-positive and ER-negative) tumors, a total of 49 (25 loss and 24 gain) CNAs were identified to be differential. HER2-positive tumors were characterized by differential gain on 17q and 20q, and loss on 4p, 8p, and 13q. HER2-negative tumors exhibited differential loss only on 16q (Figure 1). When we compared TN (n=65) to all non-TN tumors (n=194), we found a total of 203 (145 loss and 58 gain) differential CNAs. Triple-negative tumors were characterized by differential gain events on 1q, 2p, 8q, 10p, 12p, and 18p, and loss events on 3p, 4p, 5q, 9q, 10q, 12q, 13q, 14q, 15q, and 20q. Non-TN tumors had differential gain on 1q, and loss on 11q and 16q (Figure 1).

When we compared CNA frequencies by race in breast tumors from AA (n=53) and CA (n=206) women, we found a total of 22 (20 loss and 2 gain) differential CNAs. Breast tumors from AA women, regardless of subtype, were characterized by differential gain at 8q, and loss at 5q, 9q, 10q, 14q, 15q. In this comparison, there were no differential gain or loss events characteristic for CA women. When we compared CNA events in TN only breast tumors of AA (n=22) and CA women (n=43), we found 216 CNAs (130 loss and 86 gain) with a frequency difference >20% by race (see Materials and Methods; Table 4). Overall, these data support the
observation that there are differences in frequency and location of CNA events for TN tumors in AA and CA women.

Hierarchical Clustering of Breast Cancer Subtypes Based on Differential Copy Number Alterations

Hierarchical clustering was performed with the 320 statistically significant differential BAC clones that were identified in the subgroup comparisons. We were interested in determining if breast cancer subtypes would cluster based on patterns of their characteristic CNAs. The dendrogram shows that the 259 tumors clustered into two major groups, one enriched for ER-negative tumors (Figure 2. Group A) and the other enriched for ER-positive tumors (Figure 2. Group B), regardless of HER2 status. The Group A cluster contained the majority of TN breast tumors, 52 of the 65 TN tumors (80%) TN tumors were characterized by gains in 1q, 8q, and 10p, and loss in 4p, 5q, 14q, and 15q. Gains at 10p15-10p12 and loss at 14q32 were the most predominant differentiating regions of CNA associated with the TN tumors. Group B on the dendrogram contained the majority of ER-positive tumors (115/154, 75%) and HER2-positive tumors (24/53, 69%).

Discussion:

In the current study, we used a genome-wide aCGH approach to profile CNAs from breast cancers for 259 young women from a previously reported population-based case-control study in Atlanta, GA [22]. We identified characteristic CNAs associated with breast cancer subtypes (ER+/-, HER2+/-, TN) and found statistically significant differences in the average overall frequency of genome-wide CNAs in subtype comparisons, as well as frequency differences in CNAs occurring at specific genomic sites. We also observed differences in the frequency (>20%) and genomic locations of CNA events for TN tumors in AA and CA women.
Our results demonstrate that TN tumors had marked genomic instability with the highest average frequency of genome-wide CNAs compared to the other breast cancer subtypes. Chin et al. and Bergamaschi et al. also reported similar findings, with TN/basal tumors having the highest frequency of both copy number gains and losses compared to other subtypes [18,19]. Fridlyand et al. observed a subset of ER- tumors associated with poor outcomes and extensive genomic instability, classifying this molecular subtype as the “complex” subtype. These “complex” tumors were found to have a high degree of similarity for CNA profiles when compared to BRCA1 hereditary tumors [20]. We also observe our TN tumor samples to have CNAs in genomic regions that are characteristically altered in BRCA1 hereditary tumors specifically at 5q, 10p, 12p, 12q, and 20q [30,31].

Copy number gain at 10p has been reported to be a distinguishing CNA for TN/basal tumors compared to other breast cancer subtypes [32,18,33,19]. We observed a copy number gain in the region of 10p spanning 10p15-10p12 in our set of TN tumors. There are numerous genes spanning this region, with several confirmed to have increased protein expression associated with TN/basal tumors. Up-regulation of gene expression for several genes in this region (10p13), specifically, C10orf7, UPF2, HSPA14, RPP38 and CAMK1D has been confirmed to correlate with copy number gain [32,34]. The region of 10p13 also contains the gene for vimentin (VIM) that has been associated with increased expression with TN/basal tumors and plays a role in the epithelial-mesenchymal transition [35]. Although we do not present corresponding gene expression data for our samples, we see a significantly higher frequency of copy number gain at 10p13, corresponding to the genomic region containing the gene for VIM in TN tumors.

Amplification at 8q24 is common in breast cancer and has been previously observed in TN/basal and BRCA1 breast tumors and associated with poor outcomes [36,37]. We observed a significant difference in the frequency of gain events in the genomic region (8q24) containing the
C-MYC gene in ER-negative, primarily TN tumors, with >50% of the TN tumors compared to non-TN tumors (30%) having copy number gain in this region. We also compared the frequency of CNAs in TN tumors of AA and CA at 8q24, and observed a negligible CNA frequency difference between AA women (54%) and CA women (52%) for gains in this region, indicating that copy number gain in this region, containing the C-MYC, is not a distinguishing feature between tumors of AA and CA women.

There was twice the frequency of copy number gain in 13q31-13q34 for TN tumors for AA (20%) versus CA (9%) women. Amplification in the region of 13q31-13q34 has been previously reported to be associated with TN/basal tumors (20%) and BRCA1-associated breast tumors (8.1%) in a study reported by Melchor et al. [38]. They identified two “driver” genes in 13q34 that facilitate tumor progression, cullin4A (CUL4A) and transcription factor Dp-1 (TFDP1). Both were demonstrated to have increased protein expression in tumors with amplification at 13q34. Both CUL4A and TFDP1 overexpression in breast cancers have been associated with shorter overall and disease-free survival [39,40]. The study conducted by Melchor et al. included a total of 188 familial and 277 sporadic breast cancer samples, most of which came from cancer centers of predominantly Latino/Hispanic patients in Spain and Ecuador. Both AA and Hispanic women with TN/basal tumors have poorer outcomes compared to CA women [9,41]. Additional studies are needed to evaluate events associated with amplification of 13q31-13q34 in relation to race and ethnicity and clinical outcome for AA, Latino/Hispanic, and Caucasian women.

ER-negative, specifically TN tumors, had a statistically significant differential frequency of copy number loss at 14q32.2 (p=0.001) when compared to the ER-positive and non-TN tumors, respectively and rarely occurred in HER2-positive tumors (5%) (Table 4). In addition, this CNA occurred more than twice as often in TN tumors of AA women compared to TN tumors of CA women (59% vs. 21%, respectively). The 14q32.2 region contains the gene for the
microRNA, miR-342. MiR-342 has a critical role in proliferation, differentiation, development, and metabolism (reviewed in [42]) and functions as a pro-apoptotic tumor suppressor in colon tumors [43]. For breast cancer, a recent study demonstrated that miR-342 expression was highest in ER-positive and HER2-positive breast tumors and lowest in TN/basal tumors. This expression pattern is consistent with the CNA profiles they found at 14q32.2 for the breast tumor subtypes, suggesting that copy number loss at 14q32.2 in TN tumors may lead to the downregulation of miR-342 expression, particularly in tumors of AA women.

Although the current literature has been inconsistent with respect to biological differences between tumors of AA and CA women, two recent reports support the hypothesis that biological differences exist and find that in women with breast tumors of similar ER status, AA women have poorer survival than CA women, even after adjustment for socioeconomic factors [44,45]. In addition, a separate study showed that there are biological differences impacting angiogenesis, chemotaxis, and immunobiology pathways in breast tumors of AA and CA women based on the comparison of gene expression profiles of tumor and stromal tissue from breast tumors of these two racial populations [46]. Our preliminary findings of differences in CNA frequencies in TN tumors from AA and CA women support the observations that there may be biological differences in the TN tumors. It is still unknown how these differences contribute to prognosis for AA and CA women.

One potential study limitation was selection bias in the array-tested samples. Therefore, we conducted a weighted analysis to address the issue of selection bias, but cannot be certain that this weighting completely addressed that issue. In addition, although there were limitations in the use of the moderate resolution BAC array for the identification of CNAs, we successfully demonstrated that we could confirm previously identified CNAs associated with specific breast cancer subtypes and identify additional novel CNAs not previously reported, particularly for the TN/basal subtype.
In this report we found characteristic genomic alterations associated with subtypes of breast cancer. The breast cancer samples included in this study were a part of a larger cohort of young women, and included the largest aCGH study on both breast tumors from young women and on number of TN tumors analyzed by aCGH. Further replication studies will need to be performed to confirm these findings. These results can be applied to future studies to increase our understanding of the biology of the different breast cancer subtypes, particularly TN tumors, and differences by race, ultimately leading towards the identification of improved targeted therapeutic strategies and breast cancer survival.
**Figure Legend**

Figure 1. Genome-wide copy number alteration (gain and loss) frequencies by subtype. The whole-genome plots show the frequency (y-axis) of gain (gray bars above the x-axis) and loss (gray bars below the x-axis) events for breast cancer subtypes. The black circle at the top of each plot represents the genomic location of a BAC clone demonstrating a differential copy number alteration for the specific breast cancer subtype.

Figure 2. Hierarchical clustering based on differential copy number alterations. Individual tumors (columns) clustered into two major groups (A and B). Tumor characteristics such as estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor-2 (HER2) and racial background are indicated in the legend. Chromosome locations of copy number alteration events (gain in pink and loss in turquoise) are indicated to the left of the heatmap.
Reference:


invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the california cancer registry. Cancer 109 (9):1721-1728. doi:10.1002/cncr.22618


Table 1. Tumor Characteristics of Cases Analyzed with aCGH

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>African American</th>
<th>Caucasian American</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aCGH n=259</td>
<td>n=53</td>
<td>n=206</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td></td>
</tr>
<tr>
<td>Estrogen Receptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>105 (36)</td>
<td>69 (34)</td>
<td>69 (34)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>positive</td>
<td>154 (17)</td>
<td>137 (66)</td>
<td>137 (66)</td>
<td></td>
</tr>
<tr>
<td>Progesterone Receptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>99 (20)</td>
<td>67 (33)</td>
<td>67 (33)</td>
<td>0.0004</td>
</tr>
<tr>
<td>positive</td>
<td>160 (40)</td>
<td>139 (67)</td>
<td>139 (67)</td>
<td></td>
</tr>
<tr>
<td>HER2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>224 (44)</td>
<td>180 (87)</td>
<td>180 (87)</td>
<td>0.5467</td>
</tr>
<tr>
<td>positive</td>
<td>35 (17)</td>
<td>26 (13)</td>
<td>26 (13)</td>
<td></td>
</tr>
<tr>
<td>Triple Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>65 (22)</td>
<td>43 (21)</td>
<td>43 (21)</td>
<td>0.004</td>
</tr>
<tr>
<td>No</td>
<td>194 (78)</td>
<td>163 (79)</td>
<td>163 (79)</td>
<td></td>
</tr>
<tr>
<td>Tumor grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>50 (2)</td>
<td>49 (24)</td>
<td>49 (24)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Intermediate</td>
<td>96 (38)</td>
<td>81 (39)</td>
<td>81 (39)</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>113 (70)</td>
<td>76 (37)</td>
<td>76 (37)</td>
<td></td>
</tr>
<tr>
<td>AJCC Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>89 (3)</td>
<td>82 (40)</td>
<td>82 (40)</td>
<td>0.0009</td>
</tr>
<tr>
<td>II A</td>
<td>79 (3)</td>
<td>62 (30)</td>
<td>62 (30)</td>
<td></td>
</tr>
<tr>
<td>II B</td>
<td>54 (3)</td>
<td>38 (18)</td>
<td>38 (18)</td>
<td></td>
</tr>
<tr>
<td>III/IV</td>
<td>37 (2)</td>
<td>24 (12)</td>
<td>24 (12)</td>
<td></td>
</tr>
</tbody>
</table>

*p-values were determined with a chi-square test
Table 2. Comparison of the Weighted Average Frequency* of Genome-wide Copy Number Alterations

<table>
<thead>
<tr>
<th></th>
<th>ER-negative (n=105)</th>
<th>ER-positive (n=154)</th>
<th>p-value**</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gain</strong></td>
<td>0.069</td>
<td>0.049</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>0.085</td>
<td>0.073</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HER2-positive (n=35)</th>
<th>HER2-negative (n=224)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gain</strong></td>
<td>0.056</td>
<td>0.058</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>0.073</td>
<td>0.079</td>
<td>0.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Triple Negative (n=65)</th>
<th>Non-Triple Negative (n=194)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gain</strong></td>
<td>0.073</td>
<td>0.052</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>0.089</td>
<td>0.074</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>African American (n=53)</th>
<th>Caucasian American (n=206)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gain</strong></td>
<td>0.069</td>
<td>0.052</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>0.090</td>
<td>0.073</td>
<td>0.003</td>
</tr>
</tbody>
</table>

*Genome-wide gain or loss events were calculated for each tumor then averaged (weighted) across tumors in each comparison group.

**p-values were generated with a weighted t-test
Loo et al.
Genomic alterations in breast cancer subtypes of young women
BCRT submission
Supplemental Data 1.

Statistical Analysis

Normalized aCGH data were processed using wavelet along the genome [1]. The processed aCGH values were then categorized into copy number loss, no change, and gain events using the cut-off log2 ratio -0.34 and 0.38, for loss and gain respectively, where the cut-off values were chosen based on X-chromosome titration experiments, previously reported [2]. We completed an extensive evaluation of the commonly used and robust array CGH (aCGH) segmentation methods including the circular binary segmentation method [3], wavelet smoothing method [1], fused-lasso regression [4], and robust smooth segmentation method [5] and a Bayesian approach [6] using simulated data sets as well as the real data set where the ERBB2 gene was independently validated by another molecular method. In a comparison of the analysis procedures, we found that the wavelet smoothing method [1] gives the best power while maintaining the correct type I error rate in detecting the differences of various aberrational sizes (results not shown).

Sampling weights were incorporated into the analysis based on the larger cohort of 950 cases for analysis of the 259 cases that were analyzed by aCGH. Because of the limitation of the sample quality and quantity, not all 950 cases were eligible for aCGH analysis. Thus, we examined the characteristics of our sample (n=259) compared with those of the entire cohort (n=950) and weighted our analyses to match the age, race, and vital status characteristics of the original cohort population to minimize any selection bias, where the weights were calculated as the inverse probability of being sample within each stratum (Supplemental Data Table 1).

We calculated the weighted average overall genome-wide frequencies of copy number gain or loss by race (AA/CA) and the following tumor subtypes: ER status (positive/negative), triple negative (yes/no), and HER2 status (positive/negative), where the genome-wide copy number gain (or loss) for a tumor was defined as number of clones showing gains (or losses)
divided by the total number of clones. The weighted average of frequencies for copy number
gain or loss at each clone was also calculated for each subgroup comparison. We also
evaluated differences in CNA gains and losses by race among triple negative tumors only. To
adjust for possible confounding effects of age and stage, weighted multivariable logistic
regression was performed to examine whether each comparison group differs in gains and
losses at each of the 4320 clones, respectively. Given some clones may have no or few events
of gains or losses, the $p$-values based on asymptotic distributions of the test statistics would be
biased. To correct for this bias, the bootstrap method was used to obtain exact $p$-values. A
total of 1000 bootstrap samples were used for each comparison.

Hierarchical clustering was performed using clones that show statistical significance in
any of the comparisons to identify whether subtypes of tumors would cluster based on the
profiles of copy number alterations. For the heatmap clustering, we used the euclidean distance
as the dissimilarity function and complete linkage.

All the analyses were done using statistical software R version 2.6.0. The wavelet
smoothing required package of ‘waveslim’; the weighted logistic regression required package of
‘survey’ (http://www.r-project.org/). Throughout the paper, a $p$-value < 0.05 is considered
statistically significant.

2. Loo LW, Grove DI, Williams EM, Neal CL, Cousens LA, Schubert EL, Holcomb IN, Massa
Array comparative genomic hybridization analysis of genomic alterations in breast cancer
subtypes. Cancer Research 64 (23):8541-8549
10.1093/biostatistics/kxh008