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 $\log_2$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 eucledian 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.

1. Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, Porter P (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6 (2):211-226. doi:6/2/211 [pii] 10.1093/biostatistics/kxi004

2. 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 Research 64 (23):8541-8549

3. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5 (4):557-572. doi:5/4/557 [pii] 10.1093/biostatistics/kxh008

4. Tibshirani R, Wang P (2008) Spatial smoothing and hot spot detection for cgh data using the fused lasso. Biostatistics 9 (1):18-29. doi:kxm013 [pii] 10.1093/biostatistics/kxm013

5. Huang J, Gusnanto A, O'Sullivan K, Staaf J, Borg A, Pawitan Y (2007) Robust smooth segmentation approach for array cgh data analysis. Bioinformatics 23 (18):2463-2469. doi:btm359 [pii] 10.1093/bioinformatics/btm359

6. Engler DA, Mohapatra G, Louis DN, Betensky RA (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. Biostatistics 7 (3):399-421. doi:kxj015 [pii] 10.1093/biostatistics/kxj015