Supplementary Methods

DNA Isolation, hybridization and DNA copy number analysis. Genomic DNA was isolated from 5 to 10 30 μm tumor cryostat sections (10–25 mg) with QIAamp DNA mini kit (Qiagen, Venlo, Netherlands) according to the protocol provided by the manufacturer. Genomic DNA from each patient sample was allelo-typed using the Affymetrix GeneChip® Mapping 100K Array Set (Affymetrix, Santa Clara, CA) in accordance with the standard protocol. Briefly, 250 ng of genomic DNA was digested with either Hind III or XbaI, and then ligated to adapters that recognize the cohesive four base pair (bp) overhangs. A generic primer that recognizes the adapter sequence was used to amplify adapter-ligated DNA fragments with polymerase chain reaction (PCR) conditions optimized to preferentially amplify fragments ranging from 250 to 2000 bp size using DNA Engine (MJ Research, Watertown, MA). After purification with the Qiagen MinElute 96 UF PCR purification system, a total of 40 μg of PCR product was fragmented and about 2.9 μg was visualized on a 4% Tris borate ethylenediaminetetraacetic acid (TBE) agarose gel to confirm that the average size of DNA fragments was smaller than 180 bp. The fragmented DNA was then labeled with biotin and hybridized to the Affymetrix GeneChip® Human Mapping 100K Array Set for 17 hours at 48^0C in a hybridization oven. The arrays were washed and stained using Affymetrix Fluidics Station, and scanned with GeneChip Scanner 3000 G7 and GeneChip® Operating software (Affymetrix). The Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) (Affymetrix) software was used to generate a SNP call for each probe set on the array. SNP call was determined for 96.6% of the probe sets across the study, with a standard deviation of 2.6%. The Affymetrix Chromosome Copy Number Tool (CCNT) 3.0 (1) software was then used to generate a value representing the copy number of each probe set. This was done by comparing the hybridized intensities of each chip to a

manufacturer provided reference set of intensity measurements for over 100 normal individuals of various ethnicities. The copy number measurements were then smoothed using the genomic smoothing function of the software with a window size of 0.5 Mb. The Affymetrix GeneChip@ Human Mapping 100K Array Set contains 115,353 probe sets for which the exact mapping positions were defined. The median length of the interval between the probe sets was 8.6 kb, 75% of the intervals were less than 28 kb and 95% were less than 94.5 kb.

Identification of chromosome regions with prognostic CNAs. The first step in our analysis was to identify chromosome regions whose CNAs were correlated with distant metastasis. Briefly, in the training set the univariate Cox proportional-hazards regression was used to evaluate the statistical significance of the correlation between the copy number of each individual SNP and the time to distant metastasis. Then, to define prognostic chromosomal regions, chromosomes were scanned in steps of 1 Mb using a sliding window of 5 Mb which contained an average of 250 SNPs to compile the Cox regression *P*-values of all SNPs within the window and to determine a smoothed *P*-value of all these SNPs as a whole relative to permutated data sets. Briefly, for a given window of size 5 Mb containing n SNPs, let β_i and P_i denote the Cox regression coefficient and the *P*-value from the Cox regression for the i^{th} SNP, respectively. A score S for this window was defined by summarizing the statistical significance of all SNPs within this window as a whole as follows:

$$S = \sum_{i=1}^{n} -\log(P_i) \cdot I_i$$

where

$$I_i = \begin{cases} 1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \end{cases}$$

The indicator variable I_i was used to account for and to distinguish the positively correlated copy number changes from the negatively correlated ones, indicated by the

signs of the Cox regression coefficients β_i. The positive coefficients reflect that relapsing patients had higher copy numbers than disease-free patients and the negative coefficients suggested the opposite. To compute the smoothed *P*-values from the scores, we used permutations to derive the null distribution of the scores. Four hundred permutations were performed by shuffling the survival time and event indicator together with respect to patients identifiers and re-computing a permuted score S for each of the permuted datasets. The smoothed *P*-values were calculated as the fraction of the permuted scores that were more extreme than the original score. From the smoothed *P*-values that were spaced at 1 Mb apart, the prognostic chromosomal regions were defined as the chromosomal segments within which the consecutive smoothed *P*-values were all less than 0.05.

Construction of CNS and predictive model. Once the prognostic chromosome regions were identified, we mapped the well defined genes with an Entrez Gene ID within those regions using the University of California Santa Cruz Genome Browser (http://genome.ucsc.edu) Human March 2006 (hg18) assembly. Next, two filtering steps were used to select those genes with greater confidence of having prognostic values to build a CNS. First, we filtered down for those genes that have at least one corresponding Affymetrix U133A probe set. Only those genes that had statistically significant Cox regression *P*-values ($P < 0.05$) from the gene expression data were followed through. Second, the correlation between the gene expression levels and copy numbers must be greater than 0.5. If the gene contained multiple SNPs inside, then the SNP with the best Cox regression *P*-value was selected; if contained no SNP, then the nearest SNP was chosen. For U133A probe set, the one with the best Cox *P*-value was used.

To build a model using the genes in the CNS to predict distant metastasis, we transformed the genes numeric copy number estimates into discrete values: amplification, no change, or deletion. In order to do the transformation, we first estimated

the diploid copy numbers for each gene by performing a normal mixture modeling on the representative SNP's copy number data and using the main peak of the modeled distribution as the estimate of the diploid copy number. Then for amplification, it was defined as 1.5 units above the diploid copy number estimate to ensure low false positives due to the intrinsic data variability; whereas deletion was defined as 0.5 units below the diploid copy number estimate because of the nature of the alteration and the narrow distribution of the copy number data for copy number loss. Once the copy number data were transformed, we used the following simple and intuitive algorithm to build a predictive model. The algorithm classified a patient as a relapser if at least *n* genes had copy numbers altered in that patient, and as a non-relapser otherwise. We examined all possible scenarios for *n* ranging from 1 to all genes in the CNS and determined the value of *n* by examining the performance of the signature in the training set as measured by a significant log-rank test *P*-value and setting a lower limit for the percentage of positives (predicted relapsers) to avoid the situation of very small number of positives as *n* increases. The minimum number of altered genes in order for a patient to be considered a relapser was 7 for ER-positive tumors and 3 for ER-negative tumors.

Validation of CNS. The performance of the CNS was assessed both in the copy number data set of the remaining validation patients and in the external array comparative genomic hybridization data set (2) using the same algorithm described above. For the external data set, because it was derived from totally different array comparative genomic hybridization technology and the data format was log2 ratios, the cutoff for amplification was set at 0.45 while the cutoff for deletion was -0.35 to ensure comparable percentage of positives generated as the SNP array technology. As with the construction of the CNS, the validation was done in the ER-positive and ER-negative tumors separately using the corresponding subsets of genes in the CNS. The final

performance shown, however, represented the combined performance for both ER-positive and ER-negative patients in the validation set.

References

1. Huang J, Wei W, Zhang J*, et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 2004;1:287-99.
2. Chin SF, Teschendorff AE, Marioni JC*, et al.* High-resolution array-CGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol 2007;8:R215.