Supplemental methods

Microarray data resources. The primary dataset generated by using Illumina HumanRef-8 v2 Expression BeadChip (http://www.illumina.com/). Total 161 breast tumors were taken between 2006 and 2008 from Princess Margaret Hospital and Mount Sinai Hospital (Toronto, ON) and finally, 149 invasive breast cancers were created as the training cohort (Table S1). The information for the validation microarray datasets [1-16] was listed in Table S2. The microarray data and their patient clinical information for the validation dataset with 295 breast cancers from Netherlands Cancer Institute [1,2] were downloaded from websites http://www.rii.com/publications/2002/nejm.html and http://microarray-pubs.stanford.edu/wound_NKI/. The other validation datasets were downloaded from NCBI Gene Expression Omnibus website http://www.ncbi.nlm.nih.gov/geo, using the accession numbers from the respective studies. All microarray data used in this study excluded replicated cases and contained clinical endpoint information. Any type of recurrence, including local recurrence and distant metastasis, was used to analyzed the relapse-free survival. All tumors must come with their clinical ER, PR and Her2 status. If the status is not available from the published materials, a request would be sent to the author, or array expression values of the three genes were used.

Agilent microarray data processing. The downloaded Agilent Hu25K data for the 295 breast cancers came with log ratios of the signals for each probe from the tumor relative to pooled sample from all patients [1,2]. The downloaded GEO series matrix files from two Agilent datasets of GSE10886 [12] and GSE6128 [13] were in log2 ratios of the tumor RNA relative to a modified Stratagene Human Universal Reference RNA, and only arrays in platform GPL1390 were used in the study. To make the two Agilent
datasets compatible with other microarray datasets, the log ratios of Agilent Hu25K dataset were converted to log₂ ratios; whereas the log₂ ratios of GSE10886 and GSE6128 datasets was first converted back to ratios and then compared that to the average ratios of all the probes in log₂ format.

**Affymetrix microarrays data processing.** The downloaded Affymetrix CEL data were processed by Expression Console version 1.1.1 of the GeneChip Operating Software (Affymetrix Inc., Santa Clara, CA). The Probe Logarithmic Intensity Error Estimation method was used to produce a summary value for each probe set by Quantile normalization and PM-MM protocols. The downloaded GEO series matrix files in normalized intensity values were directly used in next step of data processing. A value of 16 was assigned to any normalized intensity value that was less than 16, according to the recommendation from MAQC Consortium [17]. A log₂ expression ratio of an intensity value to the average signal value for each transcript in all samples was calculated.

**Integration of published gene expression signatures.** Sixteen gene expression signatures that have previously been reported to have prognostic predictability in breast cancers [1,4,5,8,12,14,19-27] are summarized in Table S3. Out of the 16 gene signatures, 14 microarray-based signatures were used to compare and evaluate the gene signature generated in this study. All array probes in the 14 signatures were re-annotated by using the tools in http://www.ncbi.nlm.nih.gov, then their official gene symbols were used to search each array data from every tumor in the training and validation cohorts. **All probes that matched to a specific gene symbol were used to classify the tumors.** The expression centroid values for each gene in the signatures were used to score the validating data series. The centroid data for PAM50 [12] was
available at https://genome.unc.edu/pubsup/breastGEO. If a centroid data was not available in their published materials, -1 was used as the good signature centroid value and +1 for poor signature. A Pearson correlation was calculated to get the quantitative scores of corresponding expression values for the genes in each tumor to the expression centroid values of the genes in each prognostic signature. The classification of Subtype [25], PAM50 [12] and CMTC, the gene signature generated in this study, were based on the nearest expression centroid method [2,25]. The adjusted threshold value of correlation coefficient -0.15 was used for WS [2,27], and 0.4 for 70GS [1,2]. The correlation coefficient value of zero was used as threshold value to classify the validation tumors for other signatures.

**Integration of published signaling pathway signatures.** Nineteen pathway signatures that enable integration of patterns in predicting activity for oncogenic signaling and other cellular pathways were collected. The training data and methods to develop gene expression signatures for pathway activity have been previously described [28,29]. To test the probabilities of the pathway activity in the 149 breast cancers in training cohort, the predicted activity patterns for the 19 pathways were represented into the three types in CMTC that was generated in this study by using a hierarchical clustering. A Pearson correlation was performed to depict the co-regulation among the pathways.

**Statistics and data analysis.** All microarray data were represented as \( \log_2 \) ratios for the expression analysis of gene transcription and entered into the Acuity software version 4 (Molecular Devices, Sunnyvale, CA) with their annotation files and clinical information for data analysis. Variant significance \( t \) test and ANOVA test were used to evaluate the differential expression between cancer groups. A Benjamini-Hochberg
method was used to control false discovery rate, and the most conservative correction method Bonferroni was applied to the \( P \) values of corresponding \( t \) tests between different microarray expression patterns. Chi-square test and Fisher's exact test were used to test the significance of the clinical and pathological variables between different cancer types. The hierarchical analysis was used to generate and present the expression patterns. Kaplan-Meier analysis was used to compare patient's survivals in differential gene expression groups, and their differences were determined by the Log-rank Test. Univariate and multivariate analyses of prognostic factors were performed by using Cox proportional hazard method. Receiver Operating Characteristic analysis was used to score the Area Under the Curve. All reported \( P \) values were two-sided, and a \( P \) value of less than 0.05 was considered statistically significant.

**Illumina array quality measures and data processing.** To measure the quality of the Illumina microarray, we incorporated a control RNA sample using Universal Human Reference RNA (Stratagene; La Jolla, CA) into each of the 30 Illumina BeadChips. The Reference microarray dataset is available at GEO website with the accession number GSE16984. For each of the 22,184 unique probes in the dataset, there was an average of 42.3±8.1 replicated beads. The correlation analysis of the expression intensity values revealed a very high average correlation coefficient of 0.9908±0077 among the 30 controls. In the sample specimens, the average correlation coefficient was 0.9918±0108 for the 10 pairs of duplicated fine needle aspiration biopsies taken from the same tumors and 0.8491±0407 among different tumors. All duplicates of the cancer samples were combined for each tumor, and a total 149 microarray data of breast cancers was used for next analysis for the selected 149 invasive breast cancers. By adjusting the lowest intensity value, 713 probes with a log\(_2\) ratio value of "0" across all samples were considered as under detectable signals and were eliminated from the next step of the
analysis. Respectively, within the 149 breast cancers, the expression levels of ESR1 and ERBB2 from microarray were consistent very well with clinical ER and Her2 status measured by immunohistochemistry or fluorescent in situ hybridization ($P < 0.0001$).

**Generation of gene expression profile for Her2+/TN phenotype.** Of the 149 breast cancers in the training cohort, 44 were Her2-positive (Her2+) or triple negative (TN, ER-/PR-/Her2-). The 44 Her2+/TN tumors were used as a group to distinguish the gene expression pattern compared to the other 105 tumors. A t test was performed to screen the most differentially expressed genes between the two groups. A total of 1428 probes (representing 1376 genes, some genes were represented by multiple oligonucleotide probes in the microarray) were selected at a level of Bonferroni corrected $P$ value less than 0.01. The hierarchical clustering analysis using the 1428-probe set resulted in division of a group of 39 tumors with 36 Her2+/TN status from the other group of 110 tumors with 8 Her2+/TN status. As shown in Figure S1A, the group with less Her2+/TN tumors can visibly be separated into two subgroups which we labeled as group 1 and 2 according to the gene expression profile, and the group enriched with Her2+/TN tumors was shown as group 3. Because we wanted to look for the molecular basis of dividing breast cancers into 3 groups similar to oncologists in the clinical settings, we went on to perform a second screen using all the differentially expressed genes that were best in separating the 149 breast cancers of the training cohort into three clusters with most Her2+/TN in one group. A total of 1349 probes (1304 genes) were selected at a level of the $P$ value less than 0.001 by an ANOVA test among the three groups. As a result, a more apparent three-cluster pattern was seen using the 1349-probe set (Figure S1B). Out of the 42 tumors in group 1, only one was Her2+/TN; there are 7 Her2+/TN in the 68 tumors of group 2, and 36 Her2+/TN in the 39 tumors of group 3.
References


