Supplemental Methods

I. Data description

The whole set of data is composed of a total of 4407 arrays described in SupTab2, and for which available bioclinical annotations were thoroughly standardised within the CIT database (http://cit.ligue-cancer.net).

1. Patients and tumors: a total of 724 (537 discovery + 187 validation) primary breast carcinomas as well as 58 (response to CT) fine needle aspiration biopsies from locally advanced patients enrolled in a neoadjuvant trial were collected in the frame of the Cartes d'Identité des Tumeurs (CIT) program from the Ligue Nationale Contre le Cancer in 23 clinical centers in France. All tumors from this collection (n=782, 537 discovery + 187 validation + 58 response to CT) were analyzed for expression profiling on Affymetrix U133 plus 2.0 chips and 488 for copy number changes by array-CGH. This dataset was split into a CIT discovery series comprising a total of 537 tumors of which 488 samples were analyzed at both the expression and genome profiling levels, 187 which were used as part of the validation series, and 58 which were included in the response-to-chemotherapies series.

Description of this tumor collection is presented in SupTable 1. Mean follow up time was of 65 months (standard-deviation of 40)

2. Validation series (= 2291 Affymetrix microarrays + 796 non-Affymetrix microrrays): in addition to the 187 expression profiles from the initial CIT collection, we collected Affymetrix expression profiling datasets from public databases (GEO and array-express) corresponding to 2104 breast cancers, as well as Agilent, Swegene and Operon breast cancer microrray series (references are given in Sup Table 2).

3. Response-to-chemotherapy series (= 307 microarrays): in addition to the 58 expression profiles from the CIT collection, we collected 249 Affymetrix expression profiling datasets from public databases (references are given in Sup Table 2).

4. Normal-mammary-cells series (= 27): we collected 27 Affymetrix expression profiles from public databases (references are given in SupTab 2).

II. Gene expression: hybridization and pre-treatment

1. RNA extraction and Quality Control: tumor samples (10 to 50 mg) were powdered under liquid nitrogen. RNA were extracted using RNAble (Eurobio, Courtaboeuf, France), followed by a clean-up step on RNAeasy columns (Qiagen, Courtaboeuf, France). Aliquots of the RNA were analyzed by electrophoresis on a Bioanalyser 2100 (version A.02 S1292, Agilent Technologies, Waldbronn, Germany) and quantified using Nano Drop™ ND-1000 (Nyxor Biotech). Stringent criteria for RNA quality were applied to rule out degradation, especially a 28s/18s ratio above 1.8 for microarray.

2. cRNA probe production and labeling: 3 mg of total RNA were amplified and labeled according to the manufacturer's one-cyle target labeling protocol (http://www.affymetrix.com). 10 mg of cRNA were used per hybridization (GeneChip Fluidics Station 400; Affymetrix, Santa Clara, CA). The labeled cRNAs were hybridized to HG-U133 plus 2.0 Affymetrix GeneChip arrays (Affymetrix, Santa Clara, CA). Chips were scanned with a Affymetrix GeneChip Scanner 3000 and subsequent images analyzed using GCOS 1.4 (Affymetrix).

3. Affymetrix chips quality control: we used the R package affyQCReport to generate a QC report for all chips (CEL files) from the CIT discovery series. All the chips that didn't pass this QC filtering step were removed from further analysis.

4. Normalization: raw feature data from Affymetrix HG-U133A Plus 2.0 GeneChip™ microarrays are normalized using Robust Multi-array Average (RMA) method (R package affy) [1].

5. Probe sets filtering: probe sets corresponding to control genes and those whose the 90e percentile of the log intensity do not reach $\log_2(10)$ are masked yielding a total of 52,188 probe sets available for further analyses.

III. CGHarray: hybridization and pre-treatment

1. Chip description: the human genome-wide CIT-CGHarray (V6) containing 4,434 sequence-verified bacterial artificial chromosome (BAC) and P1-derived artificial chromosome (PAC) clones, was chosen to obtain a systematic coverage of the genome and detailed coverage of regions containing genes previously implicated in carcinogenesis. This array was designed by the CIT-CGH consortium (Olivier Delattre laboratory, Curie Institute, Paris; Charles Theillet laboratory, CRLC Val d'Aurelle, Montpellier; Stanislas du Manoir laboratory, IGBMC, Strasbourg) and the company IntegraGen™. The 4,434 clones, spaced at approximately 600 kb intervals, were spotted in quadruplicate on the slides.

2. DNA labeling and hybridization protocols: 600ng of tumor DNA was labeled by the random priming (Bioprime DNA labelling system; Invitrogen, Cergy-Pontoise, France) with cyanine-5 (CyDye dCTP Multipack, Amersham GE Healthcare, Buckinghamshire, UK). 600 ng of reference normal DNA (a pool of 20 normal female DNAs) was labeled with cyanine-3 using the same procedure. After ethanol-coprecipitation with Human Cot-1 DNA (Roche, Basel, Switzerland), resuspension in 72.5 µl of hybridization buffer, denaturation at 100°C for 10 minutes and prehybridization at 37°C for 90 minutes, probes were hybridized on treated microarray slides in a humidity chamber at 37°C for 24 hours. After washing, slides were scanned with a GenePix 4000B scanner (Axon Instruments Inc., Union City, CA, USA) and analyzed with GenePix Pro 5.1 image analysis software, which defined the spots and determined the median intensities for the Cy3 and Cy5 signals of each BAC clone.

3. Spot filtering and normalization: raw log2-ratio feature values were filtered from further analyses (i) using a signal-to-noise threshold of 2.0 for the reference channel or (ii) when the individual single intensities for the sample or reference was less than 1.0 or at saturation (i.e. 65,000). The remaining values were normalized using the lowess within-print tip group method [2]. For BACs in which more than 1 feature value remained after filtering and that yielded an inter-feature standard deviation of less than 0.25, an average normalized log2-ratio value was calculated.

4. Smoothing: the iterative, data-adaptive smoothing technique Adaptive Weights Smoothing (AWS, http://www.wias-berlin.de/project-areas/stat/publications/paper.html; Polzehl and Spokoiny) was then applied to the normalized log2-ratio values (as adapted in the R GLAD package v1.8) [3]. This yielded smoothed log2-ratios values in homogeneous segments along the chromosome.

5. Determination of DNA copy number: for each sample, the level (L_N) corresponding to a normal (i.e. diploid) copy number is determined as the first mode of the distribution of the smoothed log2-ratio values across all autosomes. The standard deviation (SD) of the difference between normalized and smoothed log2-ratio values is calculated. Then for all clones in a segment, the 'GNL' copy number status (G: gain - N: normal - L: loss) is determined as follows: based on the segment smoothed log-ratio value (X): if $X > L_N + SD$ then status=gain (G), if $X < L_N - SD$ then status=loss (L), else status=normal. In a given segment, outlier clones that yielded normalized log2-ratio values (Y) such that $Y > L_N + 3 \times SD$ (respectively $Y < L_N - 3 \times SD$) are classified as gains (respectively losses).

IV. TP53 typing

TP53 status was determined by the yeast functional assay, in which mutant TP53 transcripts yield red yeast colonies and wild-type transcripts yield white ones [5]. Tumors were

considered TP53 mutant when: (i) more than 15% of the yeast colonies were red, (ii) analysis using the split versions of the test could identify the defect in the 59 or 39 part of the gene, confirming the initial determination [6], and (iii) sequence analysis from mutant yeast colonies could identify an unambiguous genetic defect (mutation, deletion, or splicing defects). All tumors with more than 15% red colonies fulfilled these three criteria. Note that the four tumors with low percentage of mutant colonies (15%–25%) all exhibited stop or frame-shift mutations, defects known to be associated with nonsense mediated RNA decay, resulting in low mRNA abundance. Prediction of dominant negative activity was performed using IARC software (http://www-p53.iarc.fr/index.html).

V. Subgroups discovery by applying a semi-unsupervised approach

Introductory note: Except when indicated, statistical analyses were carried out using either an assortment of R system software (http://www.R-project.org, V2.10.1) packages including those of Bioconductor [7] or original R code. R packages and versions are indicated when appropriate.

Our rational was to produce a robust classification scheme independent of previously proposed approaches and ensure the greatest possible homogeneity to identified subgroups. To this aim, subgroup determination was based on the CIT discovery series including 537 Affymetrix U133Plus2 microarrays. We applied an approach of clustering that iterates unsupervised and supervised steps, which was, therefore, designated as “semi-unsupervised” clustering approach.

The overall approach applied in our study is summarized in SupFig 1.

Step 1: Unsupervised probe sets selection

Probe set unsupervised selection was based on two criteria:

(i) p-value of a variance test (see below) < 0.01

(ii) a coefficient of variation < 10 and a rCV percentile > 99% (see below). After filtering we were left with 244 probe sets corresponding to 188 known genes.

Variance test: For each probe set (P) we tested whether its variance across samples was different from the median of the variances of all the probe sets. The statistic used was ((n-1)×Var(P) / Var_{med}), where n refers to the number of samples. This statistic was compared to a percentile of the Chi-square distribution with (n-1) degrees of freedom and yielded a p-value for each probe set. This criterion is the same used in the filtering tool of BRB ArrayTools software [8].

Robust coefficient of variation (rCV): For each probe set, the rCV is calculated as follows: having ordered the intensity values of the n samples from min to max, we eliminate the minimum value and the maximum value and calculate the coefficient of variation (CV) for the rest of the values.

Step 2: Preliminary clustering and samples coreset

A preliminary set of five molecular subgroups was determined by applying three parametric and non-parametric statistical methods of clustering on the 537 microarrays and the 244 probe sets: (i) Agglomerative Hierarchical Clustering with Pearson correlation as a similarity measure and the Ward's linkage method to minimize sum of variances (as in Step 1); the number of subgroups (=5) was assessed qualitatively by considering the shape of the clustering; (ii) Mixed-Gaussian-Models (R package mclust); the number of subgroups (=5) is assessed with the Bayesian-Information-Criterion (BIC); (iii) K-Means-Clustering (R package stat); the number of subgroups is set to the same value (=5) than the one determined with the two other methods.

The five classes defined according to the three unsupervised methods were matched and the 394 samples for which the three methods showed convergence were selected. Conversely, 143 samples associated to discordance were taken out.

Step 3: Identification of a molecular signature

A supervised analysis was performed on the 394 samples and all the probe sets to determine probe sets best discriminating the molecular subclasses. To this aim, 21,000 probe sets were selected according to a classical Analysis-of-Variance (FDR < 1e-7) (R

package kerfdr) and then ranked by random-forest (R package randomForest). This produced a minimal list corresponding to 375 probe sets (256 known genes) leading to the best re-classification of the samples (SupTab3).

Step 4: Final clustering and sample

Using the 375 probe sets selected in Step 3, we re-applied Step 2 on our discovery set including 537 microarrays. This led to the identification of six main molecular subgroups by the convergence of the three clustering methods. They represented a total of 355 samples that constituted the coreset that was used for further investigations.

VI. Predictors

1. CIT predictors: There are two CIT predictors, one for Affymetrix (RMA normalized) data and one for non-Affymetrix (pre-treated) data. Both predictors (as well as the related confidence score –see below–) are implemented in the citbcmst R package (CRAN repository http://cran.r-project.org/web/packages/citbcmst/index.html) coming with a Sweave user documentation.

Predictor for Affymetrix (RMA normalized) profiles : given a sample profile S to be assigned to one of the 6 CIT subgroups, and the set X of probe sets that were measured for S, the following steps are processed : 1) identify the set Y of probe sets common to X and to the set of 375 probe sets given in supplemental table 3. 2) compute centroids of the 6 subtypes on these reduced dataset of Y probe sets, using the 355 samples from the CIT coreset 3) compute the distance of the new input sample(s) to those 6 centroids 4) assign sample(s) to the subgroup corresponding to the closest centroid. Here the (DLDA) distance between S and the centroid of a subgroup K is defined as:

$$\sum_{i=1..N} \frac{(measure(S, gene_i) - m(subgroup_K, gene_i))^2}{s(gene_i)}$$, where $\mu(subgroup_K, gene_i)$ designates the mean expression of the gene (probe set) i across the samples from the CIT coreset being in

subgroup K (K in {lumA, lumB, lumC, normL, mApo, basL}) and σ ($gene_i$) the standard-deviation of the gene (probe set) i across the samples from the CIT coreset.

NB: $\mu_{K,i}$ and σ_i values are given in SupTab3.

Predictor for non-Affymetrix (pre-treated) profiles : given a sample profile S to be assigned to one of the 6 CIT subgroups, and the set X of genes (HUGO gene symbols) that were measured for S, the following steps are processed :1) identify the set Y of genes common to X and to the set of 256 genes given in SupTab3. 2) Aggregate data by gene (HUGO gene symbols). 3) compute centroids of the 6 subtypes on these reduced dataset of Y genes, using the 355 samples from the CIT coreset. 4) compute the distance of the new input sample(s) to those 6 centroids 5) assign sample(s) to the subgroup corresponding to the closest centroid. Here the distance used is (1-Pearson coefficient of correlation).

Confidence score for the CIT predictors: In order to have a confidence evaluation of the subtype assignation, we have defined a score to identify outliers and characterizes a sample assignment to a subgroup as certain or uncertain. If a sample is close to several centroids, i.e. if the difference of distance to centroid is inferior to the 1st decile of the difference between centroids on data used to compute centroids, the score is set to uncertain. If the distance to the assigned centroid is n times superior to the mad (median absolute deviation) of distances to the centroid within the related subgroup in the training set, the sample is set to outlier; n is defined on data used to compute centroid as the maximum between the 6 subtypes of $(\max_{\text{distances to centroid c}} - \text{med}_{\text{distances to centroid c}})/\text{mad}_{\text{distances to centroid c}}$.

2. Sorlie, Hu and Parker classifiers: Sorlie [17], Hu [21] and Parker [22] centroids were respectively retrieved from (1) (2) and (3) (see below). To build the corresponding predictors, the procedure used for the CIT non-Affymetrix predictor (see above) was repeated here. For Sorlie centroids the 552 clone ids from the intrinsic gene set corresponded to 334 unique HUGO gene symbols, which were then mapped to Affymetrix (U133A or U133Plus2) probe-

sets. For Hu centroids of the 306 original UniGene ids, 232 corresponded to a unique HUGO gene symbol, which were then mapped to Affymetrix (U133A or U133Plus2) probe-sets. For Parker centroids the 50 HUGO gene symbols were directly mapped to Affymetrix (U133A or U133Plus2) probe-sets.

(1) http://genome-www.stanford.edu/breast_cancer/robustness/data/IntrinsicGeneList.txt

(2) https://genome.unc.edu/pubsup/breastTumor/data/306genes-X-249samples-X-5subtypes+5centroids.xls

(3) https://genome.unc.edu/pubsup/breastGEO/pam50_centroids.txt

3. Van't veer and GGI predictors: These predictors were built as described in [18] and [19] using the CIT coreset as training set. To train the GGI predictor we used the grade information of the CIT coreset.

4. Jönsson classification system (arrayCGH-based): This predictor was built as described in [20]. The 6 Jönsson centroids are relative to genomic regions determined with the GISTIC algorithm [23]. To apply this predictor to the CIT arrayCGH coreset (n=320), we averaged the smoothed log2 ratios obtained as described above (see III.4) by Jönsson GISTIC region and used (1-Pearson coefficient of correlation) as distance between these profiles and Jönsson centroids.

5. Performance on external dataset evaluation: To evaluate the performance of our classification system on non-Affymetrix external dataset, we analyzed the GSE3155 dataset (Sorlie et al 2006 [24]), where the same 20 samples were analyzed on 3 different platforms (Applied Biosystem (AB), Agilent (AG) and Stanford (ST) microarrays) and where the analysis on AB platform was done in duplicate. We assigned each microarray profile to a CIT subtype as mentioned above (Predictor for non-Affymetrix profiles in CIT predictor subsection), using the same genes for all platforms. We assessed the intra-platform robustness of our classifier as the concordance between duplicated samples (AB platform). We assessed the inter-platform robustness as the concordance between the 3 platforms 2 by 2 (CIT vs AB, CIT vs AG, CIT vs ST).

VII. Statistical tests

1. Differential expression: to identify genes differentially expressed between the sample subgroups, based on the RMA $\log_2$ single-intensity expression data, we used Welch's T-tests (t.test function, R package stats) as well as the ANOVA (aov function R package stats).

2. Differential genomic status: to identify clones / regions with differential genomic status, based on the GNL (Gain/Normal/Loss) copy number status, we used the chi-square test (or the equivalent Fisher-exact test when appropriate) (chisq.test and fisher.test functions, R package stats).

3. Clinical factors: association of the sample subgroups to bio-clinical factors was tested by applying the chi-square test (or the equivalent Fisher-exact test when appropriate) for qualitative factors (gene expression, mutation, histological type, SBR grade and metastatic sites) and the ANOVA test for quantitative variables (age of diagnostic).

4. Survival: disease outcome was investigated by applying a Cox model on Kaplan-Meier curves stratified for each subgroup (function Surv, R package survival). P-values at 60, 120 and 180 months resulted from a log-rank test on Cox estimates (function survdiff, R package survival). The proportional-hazards assumption was tested to examine the model's appropriateness.

5. Response to chemotherapy: Association of subgroups to response to chemotherapy was assessed using the chi-squared test. For adjustment relatively to other factors (ER and grade) we used the Cochran-Mantel-Haenszel chi-squared test.

6. P-values adjustment: p-values adjustment for multiple-testing was performed using the p.adjust function from stats R package which estimates the FDR using the Benjamini and Hochberg (BH) method [9].

7. Π1 proportion: the proportion of tests under H1 was calculated using the Storey method.

VIII. Principal Component Analysis for dimensional reduction and visualization

Principal Component Anallysis (PCA) identifies new variables, the principal components, which are linear combinations of the original variables [10]. The first principal component is the direction along which the samples show the largest variation. The second principal component is the direction uncorrelated to the first component along which the samples show the largest variation. Each component can then be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the samples when projected onto the component.

IX. Gene cluster expression

Gene cluster expression values are based on the mean expression of all the genes of a given cluster. Their distribution for each molecular subgroup is represented by using boxplots. Except for the lum-C cluster, all the gene clusters represented in Figure 1 results from the molecular signature of 256 genes.

X. Validation on a large set of data and comparison between CIT the other classification systems

Validation series were treated independently in order to avoid confusion due to technical bias and different molecular composition between series. Pre-treatment was processed following the same workflow described previously and applied to the main CIT discovery series for Affymetrix series (starting with CEL files); for non-Affymetrix series, as pre-treated data were available they were directly used.

CIT subgroups prediction: samples for each series were independently classified into the six CIT main subgroups by applying the relevant predictor (see chapter VI above) depending on the platform (Affymetrix versus non-Affymetrix).

Sorlie / Hu / Parker classification prediction: samples of each series were independently classified using the predictors obtained as described in chapter VI.

XI. Relation to normal mammary epithelial cell hierarchy

We collected expression data of normal mammary gland sorted cells from 3 public datasets [GSE11395 { Affymetrix X3P}, GSE18931 { Affymetrix HG U133 Plus 2.0 }; GSE16997 {Illumina HumanWG-6_V3 }]. For each of these 3 series we calculated an independent gene signature by comparing the differentiated cells and stem cells (-like) populations, using a 0.05 p-value threshold. More precisely (i) in Raouf dataset we compared the differentiated luminal samples (n=3) to the bipotent and committed samples (n=6); (ii) in Pece dataset we compared the PKH negative samples (n=3) to the PKH positive samples (n=3); (iii) in Lim dataset we compared the mammary stem cells (MaSC) samples (n=3) to the mature luminal samples (n=3).

For Pece and Lim datasets the compared conditions were respectively derived from the same 3 pools and same 3 patients, so we used paired T-tests. For Raouf dataset we used the Bayes moderated T-test [11] implemented in the *limma* R package.

We then built a meta-signature (163 genes, SupTab9) by selecting genes present in all 3 signatures (that is genes for which the p-value was less than 0.05 in all 3 series).

The expression profiles from the CIT coreset restricted to the genes from this meta-signature were then used in a Principal Component Analysis. The CIT coreset samples profiles were then projected on the first 2 Principal Components (upper panel in Figure 5). The same space was then used to project normal mammary gland samples profiles (lower panel in Figure 5) from Lim dataset.

XII. Cancer pathways analysis

Cancer pathways: we selected a set of KEGG (ftp://ftp.genome.ad.jp/pub/kegg/pathways/hsa), Biocarta (http://www.biocarta.com) and

MSigDB (http://www.broadinstitute.org/gsea/msigdb) biological pathways known to be associated to cancer, and mapped the related genes to non-redundant HUGO Gene symbols.

Subgroup pathways: given a cancer pathway and the molecular subgroup to be compared to the others, four methods are used: GSA (R package GSA, Efron and Tibshirani [12]), Globaltest (R package globaltest, [13]), SAM-GS (original R code implementing the algorithm bu Dinu et al [14]), Tuckey approach (original R code implementing an algorithm described by Goeman et al [15]). Each method will yield a p-value based on Monte-Carlo simulations: the lower the p-value, the more the genes from the gene set are differentially expressed between the sample's subgroups. In order to rank the gene sets by order of interest, we used the mean rank of the p-value across the four methods.

XIII. DNA copy number aberrations

1. Cumulated profiles of alteration: cumulated profiles of alteration, in the whole cohort of 488 CGH arrays and by subgroup, were obtained by computing the proportion of samples harboring a gain or a loss of copy, at each clone of the array.

2. Identification of frequently-altered genomic regions in the whole cohort: frequently-altered genomic region in the whole cohort were determined by identifying regions for which the proportion of alteration (in gain or loss) exceed 20%, 30%, 40% and 50%.

3. Identification of subgroups specific regions: for each subgroup, specific regions were determined by (i) applying at each clone a test of proportion comparing the proportion of alteration (gain and loss) in the samples of a given subgroup versus the others; (ii) resulting p-values were corrected for multiple-testing by FDR; (iii) then, subgroup-specific genomic regions defined as significantly more altered in the subgroup of interest, were delimited by applying a threshold of 1% on p-values; (iv) the resulting regions are refined by applying the

segmentation-clustering approach described in Picard et al (2005) [16] for aCGH, to the signal of p-values within each region.

4. Integration of genomic and expression data: in order to identify putative candidate genes impacted by the alterations identified in the whole cohort and specific to each subgroup, we integrated expression data to the genomic profiles. We mapped BAC clones and probe sets based on their genomic position. Then given a set of samples (whole cohort or molecular subgroup) and a genomic region of alteration, we compared the expression of each gene mapped within this region by applying a t-test between samples harboring the alteration (gain for regions of gain and loss for regions of loss) and samples that do not present the alteration.

5. Identification of focal alterations: for a given set of samples, we identified focal amplifications based on the intra-sample rank of the normalized log2-ratio values, by selecting clones for which a significant number of samples showed values superior to the 99^{th} percentile.

6. Functional resolution of the CGH Array platform used: we used the method by Coe et al [25] to assess the functional resolution of our BAC-array platform. The theoretical sensitivity, single-copy sensitivity and the breakpoint precision for a 95% cut-off were estimated using the associated software ResCalc.

In order to estimate the percent of focal regions that can be detected using CGH array, we evaluated the concordance between focal regions (amplification <0.5Mb and homozygous deletion <1Mb) found for 72 samples from the CIT coreset hybridized on Illumina 610K SNParray with the GNL status of the CGH clones mapping those regions.

XIV. Significance of prognostic parameters and molecular signature

A Cox proportional-hazards model was fit to assess differences in 5-year survival to compare the CIT classification to prognostic parameters and molecular signatures. The proportional-hazards assumption was tested for each model to examine the model's appropriateness.

Prognostic factors: CIT classification {normL; lumA; other subgroups}, ESR1 (EXP), ERBB2 (EXP), N {0; 1+}, T {0-1; 2+}, ± SBR grading, ± adjuvant chemotherapy, ± adjuvant hormonotherapy.

Prognostic molecular signatures: Sorlie [17], van't Veer [18], GGI [19], Hu [21], Parker [22].

First a univariate analysis is performed to assess the marginal prognostic value of each variable independently from the others. In addition, a multivariate analysis is performed using variables having available values for a sufficient number of samples. To be comparable, each variable have to be assessed on the same samples; consequently we analyzed prognostic factors separately from molecular signatures.

XV. ER, PR, AR and ERBB2/HER2 scoring

This was done both on the basis of IHC staining and on that of Affymetrix expression measures. We used both in parallel, because of the limited availability of IHC data in the validation series.

IHC scoring was as follows: ER and PR status were defined as positive when 10% of carcinoma cells (or more) were stained, whereas HER2 was done according to the Herceptest® system **[**0 = no or less than 10% membrane staining positive cells; 1+ $\geq$ 10% stained cells with weak staining; 2+ $\geq$ 10% stained cells with weak or moderate complete staining; and 3+ $\geq$ 10% cells with strong and complete staining**]**. Tumors scored 3+ were considered HER2-positive, whereas 2+ tumors were verified by CISH or FISH for gene amplification. The current ASCO/ CAP guidelines (Hammond et al. JCO 2010 and Wolff et al. JCO 2007) were not used in this work, since the selection of patients and tumor samples began in 2005 while the ASCO/ CAP guidelines were published in 2007.Pathologists involved in this work regularly participate to national insurance quality tests on ER, PR and HER2 techniques, proposed either by the "AFAQAP" (Association Française pour l'Assurance Qualité en Anatomie Pathologique) or by UKNEQAS. They also have regular

consensus meeting and participate to multicenter studies, initiated by the “GEFPICS” (Groupe d’Etude des Facteurs Pronostiques en Immunohistochimie dans les Cancers du Sein), aiming at improving inter-laboratory and inter-observer agreement.

Affymetrix scoring was done as follows: we used the following probe sets AR: 211621_at (HG-U133A) and 226197_at (HG-U133plus2.0); ESR1: 205525_at ; PGR: 208305_at (HG-U133A) and 228554_at (HG-U133plus2.0); ERBB2 : 216836_s_at. To define positivity/negativity thresholds were adapted to each series using the density R function: distributions were bi-modal; the thresholds were put at the point of smallest density between the 2 modes.

Comparison of IHC and Affymetrix measures: We also compared IHC and Affymetrix expression measures for ER, PGR and ERBB2. We hypothesized that if we used the same set of samples with complete information in both approaches, the “best” definition should be the one yielding the highest level of intra-condition homogeneity and inter-conditions heterogeneity assessed as the differential expression level (i.e. H1 proportion) in a T-test comparison of the two conditions (positivity/negativity) for all of the 56K probe sets of the Affymetrix HGU133 plus2.0 chip. In all 3 cases (ER, PGR and ERBB2) the Affymetrix expression-derived definition clearly yielded the highest H1 proportion compared to the IHC-derived definition. We thus used both the IHC and the Affymetrix derived definitions of ER, PGR and ERBB2 in the principal and supplemental tables. Of note, the Affymetrix-derived definitions have the great advantage to ensure a standardized definition to be used over the public Affymetrix datasets.

XVI. Pathological review and SBR Grading

This was performed in each contributing center which are all academic cancer hospitals, where breast cancer is a major pathology. Histological grade was defined according the

modified SBR (Sarff, Bloom and Richardson) grade according to Elston and Ellis (Ellis et al. Histopatholgy 1992;20:479-498).

XVII. Percent of non-diploid cells

To get an objective estimate of the rate of non-diploid cells in the analyzed tumors and determine its distribution in the molecular subgroups, we analyzed 72 samples from the CIT coreset on Illumina 610K SNParray and used the formula recently published by Van Loo et al, (PNAS, 2010) to compute the rate of non-diploid cells.

REFERENCES

1. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4:249-264.

2. Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30:e15.

3. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics. 2004; 20:3413-3422.

4. Rouveirol C, Stransky N, Hupe P, et al. Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics. 2006; 22:849-856.

5. Flaman JM, Frebourg T, Moreau V et al. A simple p53 functional assay for screening cell lines, blood and tumors. PNAS. 1995; 92:3963-3967.

6. Waridel F, Estreicher A, Bron L et al. Filed cancerisation and polyclonal p53 mutation in the upper aero-digestive tract. Oncogene. 1997; 14:163-169.

7. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5:R80.

8. Simon R, and Peng Lam A. BRB-ArrayTools software v3.1 User's Manual linus.nci.nih.gov/BRB-ArrayTools.html, 2003.

9. Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B. 1995; 57 289-300.

10. Ringnér M. What is principal component analysis. Nature Biotechnology. 2008; 26:303-304.

11. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004)

12. Efron, B. and Tibshirani, R. On testing the significance of sets of genes. Stanford tech report rep 2006. http://www-stat.stanford.edu/~tibs/ftp/GSA.pdf.

13. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004; 20:93-99.

14. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics. 2007; 8:242.

15. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23:980-987.

16. Picard F, Robin S, Lavielle M et al. A statistical approach for array CGH data analysis. BMC Bioinformatics. 2005; 11:6:27

17. Sorlie TR, Tibshirani J, Parker T et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. PNAS. 2003; 100:8418-8423

18. van't Veer LJH, Dai MJ, van de Vijver YD et al. Expression profiling predicts outcome in breast cancer. Breast Cancer Res. 2003; 5:57-58

19. Sotiriou CP, Wirapati S, Loi A. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. JNCI. 2006; 98:262-272

20. Jönsson G. et al – Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. Breast Cancer Res, 2010.

21. Hu Z et al. - The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics, 2010.

22. Parker JS et al. - Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. JCO, 2009.

23. Beroukhim et al., Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. PNAS, 2007

24. Sorlie T, Wang Y et al., Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomic, 2006.

25. Coe et al., Resolving the resolution of array CGH. Genomics, 2007.

Legends to Supplemental Figures

SupFigures

SupFig1. Overall discovery process: steps of the unsupervised analysis. Workflow scheme of the unsupervised approach used to define and characterize molecular classification of the CIT discovery set.

SupFig2. Subtypes gene cluster characterization
Expression levels boxplot of the genes in clusters specific to each subtype.

SupFig3. Metastasis related genes expression levels in CIT subtypes. Expression level boxplots stratified in the 6 CIT subgroups of ST6GALNAC5, HBEGF, PTGS2 whose increased expression has recently been associated to brain metastasis (Bos et al., 2009; Minn et al., 2005).

SupFig4. Kaplan-Meier survival curves in CIT subtypes defined in non-Affymetrix datasets. Metastasis-Free survival (left) and overall survival (right) was determined in 3 whole genome expression datasets produced on different technological platforms**:** (A,B) Operon/Qiagen dataset (Chanrion et al. 2008), (C) Swegene dataset (Jönsson et al. 2010), (D,E) Agilent dataset (van de Vijver et al. 2002). Only overall survival data were provided for the Swegene dataset. The log rank test *p*-values are given for two delay cutoffs 60 months (5 years) and 120 months (10 years).

SupFig5. Metastasis-Free Survival according to molecular subgroups defined by the CIT, Sorlie, Hu and Parker and Jönsson classifications in the CIT discovery set. (A) CIT (B) Sorlie (C) Parker (D) Hu (E) Jönsson. The log rank test *p*-values are given at 60 months (5 years), 120 months (10 years) and 180 months (15 years).

SupFig6. Metastasis-Free Survival according to molecular subgroups defined by the CIT, Sorlie, Hu and Parker and Jönsson classifications in the Affymetrix validation set: (A) CIT, (B) Sorlie (C) Parker (D) Hu (E) Jönsson. The log rank test *p*-values are given at 60 months (5 years), 120 months (10 years) and 180 months (15 years).

SupFig7. Copy Number Alterations (CNAs) showing inverse patterns according molecular subgroups. CNAs are depicted as color bars arranged according to their position on the genome: BasL (red), mApo (orange), lumC (pink), lumB (light blue), NormL (green). Bars represent $-\log_{10}$ *p*-values of the increase in proportion in a given subgroup tested against all others. Gains are depicted as bars going up, losses going down.

SupFig8a. Fraction of non diploïd cells in breast tumors stratified by CIT subgroups. Boxplot represent the fraction of non diploid cells according to CIT subgroups estimated from 72 SNP data using the Van Loo et al. approach. Numbers between brackets indicate the number of samples per box.

SupFig8b: correlation between the snp-based estimation of the fraction of non diploid cells and the pathological estimation of non tumor cells in the same samples.

SupFigure9. heterogeneity of ERBB2-positive tumors. Histogram of moderated *t*-test *p*-values between *(top)* mApo samples and lumC samples having an ERBB2 amplification (ERBB2+), *(middle)* between ERBB2+ and ERBB2- samples in the mApo subtype and *(down)* between ERBB2+ and ERBB2- samples in the lumC subtype. The H1 proportion is the estimate of the proportion of *p*-values generated under H1 hypothesis, i.e. that a gene is differentially expressed between the two groups.

Legends to Supplemental Tables

SupTab1. Clinical and Molecular Description of the CIT discovery cohort. Descriptive values concerning SBR grade, Tumor Size from surgery, Age at diagnosis, Histological type, nodal involvement and ESR1 (ER), PGR (PR) and ERBB2 (HER) protein expression (IHC, 0:absent, 1:present).

SupTab2. datasets used in this work. List of all the dataset, CIT and publicly available, used, per analysis type (training, validation, response-to-treatment, normal mammary cell) with the references to the related article, the database accession number, the platform type and the number of samples/arrays used.

SupTab3. CIT Subtypes Centroids. List of the 375 pbs (256 genes) with the mean values per subtype constituting the centroids used to classify new datasets. For each probesets, are also given the Affymetrix Gene Symbol annotation (version na29, Jun 30, 2009), chromosomal location and the gene cluster.

SupTable4a. Bio-Clinical correlations of CIT molecular subgroups defined in the Affymetrix validation set. Association (Chi-squared or Fisher test) between clinico-molecular annotations and CIT subtypes.
SupTable4b. Bio-Clinical correlations of CIT molecular subgroups defined in the non-Affymetrix validation set. Association (Chi-squared or Fisher test) between clinico-molecular annotations and CIT subtypes.

SupTab4c inter platform classification agreement. The GSE3155 dataset generated in parallel on 3 technological platforms using the same samples; Agilent (AG, dual-color), Stanford (ST, dual-color), Applied Biosystems (AB, uni-color). Each dataset was classified using our classification rule and results were compared. Samples were run twice on the AB platform allowing to test for intra platform reproducibility as well.

SupTab5. Comparison of sample attribution between CIT classification and other classifications on the coreset (*left*) and on the Affymetrix validation set (*right*). (A,E) CIT versus Sorlie, (B,F) CIT versus Parker, (C,G) CIT versus Hu, (H) CIT vs Jönsson

SupTab6. Bio-Clinical correlations with the classifications of Sorlie, Hu, Parker and Jönsson. This SupTable includes 8 spreadsheets corresponding to sub-tables. Distributions are presented as in ST4, each Table is presented for comparison with the Table presenting the distribution using the CIT classification.

SupTab7. Copy Number Alterations (CNAs) regions specific to each subgroup. CNAs significantly associated to a molecular subgroup are listed according to subgroup and ordered according to chromosomal location. Each region is defined by its start and end positions (5' end of the BAC clone on the left and the 3'end of the BAC on the right extremity). Size represents the distance between these two values. Mean p value reprensents the level of association of the region to the molecular subgroup. Genes listed indicate genes with significant expression changes.

SupTab8a. Molecular subgroups show preferential amplification patterns.

SupTab8b: **Focal CNA were detected on a subset of 72 breast tumors from the CIT discovery set.** All tumors were previously analyzed by BAC-arrays. Total numbers of CNAs were determined and distributed as gains, losses, recurring and probable CNVs (events showing identical starts and/or ends). Mean size of each category of event was calculated and the overlap between the number of regions determined on the Illumina-arrays and those from the BAC-array.

Sup Tab9: Enrichment of mammary epithelial cell subpopulations signatures in CIT subgroups. Results of gene set differential analysis (GSA) between CIT basL and lumA (A) and between CIT basL and lumB (B) for each gene set of genes deregulated between the 3 normal mammary epithelial cell subpopulation in Lim et al., 2009; Pece et al.2010 and Raouf et al. 2008.

Sup Tab10: Prognostic significance of the CIT classification relative to that of clinical parameters (A) and of 3 molecular classifiers (Sorlie, Hu, Parker) and 2 prognostic signatures (GGI, Van't Veer). Relative risk was calculated taking metastatic relapse as an endpoint. The dataset comprised 1186 patients from the Affymetrix validation set for which MFS information was available. Complete clinical information was available 995 cases of the Affymetrix validation set explaining the smaller numbers in the multivariate analysis on prognostic factors. SBR grade was available in about 25% of cases was thus excluded of the multivariate analysis because it reduced statistical power significantly.

Sup Tab11. Summary of all tables/figures in the article with a brief description, sample list used, gene list used if relevant and reference to supplememtal method chapter.