A refined molecular taxonomy of breast cancer

M. Guedj[1*], L. Marisa [1*], A. de Reynies [1*], B. Orsetti [2, 3], R. Schiappa [1], F. Bibeau [4], G. MacGrogan [5], F. Lerebours [6], P. Finetti[7], M. Longy [5], P. Bertheau[8], F. Bertrand[6], F. Bonnet[5], A.L. Martin[9], J.P. Feugeas[10], I. Bièche [6], J. Lehmann-Che[10],R. Lidereau [6], D. Birnbaum [7], F. Bertucci [7,], H. de Thé [10*], C. Theillet [2, 11, 12*#]

1- Ligue Nationale Contre le Cancer, Cartes d'Identité des Tumeurs program, 14, rue Corvisart, 75013 Paris, France

2- IRCM, Institut de Recherche en Cancérologie de Montpellier, Montpellier, F-34298, France

3- CRLC Val d'Aurelle Paul Lamarque, Montpellier, F-34298, France

4- Department of Pathology, CRLC Val d'Aurelle Paul Lamarque, Montpellier, F-34298, France

5- Department of Pathology, Institut Bergonié; (EA 3669), Université Victor Segalen Bordeaux-2, Bordeaux, F-33076, France

6- Oncogenetics laboratory, INSERM U735, Institut Curie-Hôpital-Centre René Huguenin, St Cloud, F-92210, France

7- Department of Molecular Oncology, CRCM, Centre de Recherche en Cancérologie de Marseille, Institut Paoli Calmette, F-13009 Marseille, France

8- Hopital St Louis APHP, Department of Pathology, University Paris-7 Denis Diderot, Paris, F- 75010, France

9- Fédération Nationale des Centres de Lutte Contre le Cancer, 101 rue de Tolbiac, Paris, F-75013, France

10- Department of Biochemistry, Hopital Saint-Louis APHP; INSERM/CNRS UMR 944/7212 ; University Hematology Institute, University Paris-7 Denis Diderot, Paris, F- 75010, France

11- INSERM U896, Montpellier, F-34298, France

12- Université Montpellier 1, Montpellier, F-34298, France

* equal contribution

# correspondence to Charles Theillet, IRCM, INSERM U896, CRLC Val d'Aurelle-Paul Lamarque, 34298 Montpellier cedex 5, France, charles.theillet@inserm.fr

Running title: molecular classification of breast cancer
4570 words in the core text

Abstract

The current histoclinical breast cancer classification is simple but imprecise. Several molecular classifications of breast cancers based on expression profiling have been proposed as alternatives. However, their reliability and clinical utility have been repeatedly questioned, notably because most of them were derived from relatively small initial patient populations. We analyzed the transcriptomes of 537 breast tumors using 3 unsupervised classification methods. A core subset of 355 tumors were assigned to 6 clusters by all 3 methods. These 6 subgroups overlapped with previously defined molecular classes of breast cancer, but also showed important differences, notably the absence of an ERBB2 subgroup and the division of the large luminal ER+ group into 4 subgroups, 2 of them highly proliferative. Of the six subgroups, the four luminal ones were ER-/PR+/AR+, one was ER-/PR-/AR+ and one was triple negative (AR-/ER-/PR-). ERBB2-amplified tumors were split between the ER-/PR-/AR+ molecular apocrine group and the highly proliferative Lum-C group. Importantly, each of these 6 molecular subgroups showed specific copy number alterations. Gene expression changes were correlated to specific signalling pathways. Each of these 6 subgroups showed very significant differences in tumor grade, metastatic sites, relapse-free survival or response to chemotherapy. All these findings were validated on large external datasets including more than 3000 tumors. Our data thus indicate that these 6 molecular subgroups represent well-defined clinico-biological entities of breast cancer. Their identification should facilitate the detection of novel prognostic factors or therapeutical targets in breast cancer.

Introduction

Breast cancer is heterogeneous. Biological features have proven insufficient for a comprehensive description of the disease. Seminal work by Sorlie and colleagues (Sorlie et al., 2003) has delineated 5 major molecular subtypes of breast cancer associated to different outcomes. This initial classification was reproduced in independent datasets (Bertucci et al., 2006) strongly suggesting the existence of distinct molecular entities in breast cancer. The Sorlie centroid approach has subsequently been redefined and adapted to more recent technological platforms (Hu et al., 2006; Parker et al., 2009).

However, criticisms have pointed to the instability of the defined subtypes (Kapp et al., 2006; Weigelt et al., 2010) and their dependence on the original set of samples or genes. Thus, while molecular classification brings interesting insights in breast cancer taxonomy, its implementation in the clinics is put in doubt due to insufficient reliability in single sample allocation (Weigelt et al., 2010). Rather, three broad classes of breast tumors drawn along their ER, PR and ERBB2/HER2 status are commonly used in the clinic. ER-/PR-/HER2- tumors were defined as triple negative, ER+/PR+/HER2- as luminal and HER2+ tumors irrespective of their ER status form the third class (Foulkes et al., 2010). However, this simple classification is also criticized because of the biological heterogeneity within classes. In particular, the correspondence between the triple negative group and basal-like breast tumors and the heterogeneity of the large ER/PR positive group have been repeatedly questioned (Foulkes et al., 2010; Gusterson, 2009). This argues for a more elaborate stratification amenable to biological exploration and clinical choices.

This prompted us to construct a robust molecular classification on a large number of samples to reach high statistical power. To this aim, we produced transcriptomes of a series of 537 primary breast cancers and, using a semi-supervised analysis, revealed 6 stable molecular subgroups. A related classification rule was defined. Each of the 6 molecular subgroups showed distinct genomic changes, correlated with a specific set of signaling pathways and was associated with significant differences in tumor grade, metastatic sites and metastasis-free survival. We propose that this classification scheme could lay the bases of an operative tool to reliably classify breast cancers in more homogeneous molecular subgroups. This classification could be highly beneficial in future investigations aiming at identifying novel prognostic factors or therapeutical targets in breast cancer.

Results:

Semi-supervised gene expression analysis identifies six prototypic molecular subtypes

Our aim was to identify molecular subgroups representing homogeneous subsets of breast cancer. Our methodology is detailed in SupFig1 and the SupMethods section. Briefly, we produced a large dataset comprising 537 primary breast cancer transcriptomes on Affymetrix U133-Plus2.0 arrays to ensure proper statistical power. First, this tumor-set was classified with 3 unsupervised methods (hierarchical clustering, Gaussian mixture models and k-means) in parallel. Of the 537 tumors, 355 yielded a consensus subgroup assignment (i.e. were assigned to the same subclass) between all 3 methods. This subset was named core-set and used for further analysis. Second, a minimal list of 256 discriminative genes with maximal intragroup homogeneity and intergroup heterogeneity was generated by ANOVA (SupTab3). Hierarchical clustering based on this list delineated six homogeneous tumor subgroups, homogeneity being confirmed by the principal component analysis (PCA) (Figure 1B). To allow the classification of independent sample profiles to one of the 6 subgroups we built a single sample predictor (SSP) based on a distance-to-centroids approach (using the previously mentioned 256 genes) (SupMethods). The 182 tumors of the discovery set lying outside of the core-set were classified using this SSP.

The overall distribution of the six subgroups was determined by 3 large gene clusters shared by at least two subgroups. The first one (cluster-VI, Figure 1A, 1C, SupTab3), containing *ESR1* and correlated genes, defined two ER-negative (ER-) and four ER-positive (ER+) subgroups (Figure 1A, 1C). The second gene cluster (cluster-IV) included the androgen receptor (*AR*) gene and encompassed five subgroups. Of the six subgroups, four were ER+/PR+/AR+, one was ER-/PR-/AR+ and one was triple negative (AR-/ER-/PR-) (Figure 1A, 1C). Interestingly, cluster-IV included transcription factors *FOXA1, SPDEF* and *XBP1*, which are usually associated to the ER-cluster (Bertucci et al., 2006). The third cluster (cluster-II), was predominantly composed of genes regulating DNA replication and cell cycle progression, thus defining elevated cell proliferation. This cluster encompassed both ER- and 2 ER+ subgroups (Figure 1A , 1C).

Each subgroup was defined by a specific gene cluster (SupFig2) in which we found genes previously part of the Sorlie centroids. Hence, for simplicity we named our subgroups according to the Sorlie subtype (Sorlie et al., 2003). ER+ subgroups were split according to expression levels of the cell cycle cluster. Low proliferation ER+ subgroups were differentiated by clusters-III and IX (Figure 1A, SupFig2), comprising respectively genes from the Sorlie luminal-A and normal-like centroids (SupTab3) and were, thus, designated LumA and NormL. The two high proliferation ER+ subgroups differed sharply in ER-cluster expression levels. The subgroup expressing highest levels of ER was named LumB. The other subgroup, positioned at the boundary between ER+ and ER- tumors, was designated LumC (Figure 1A, 1B, 1C). Noteworthy, 40% of LumC tumors overexpressed the *ERBB2/HER2* gene.

Next was the AR+/ER-/PR- subgroup (Figure 1B), defined by cluster-VIII. The AR+/ER- status of this subgroup was reminiscent of the previously described “molecular-apocrine” subtype (Farmer et al., 2005) and we designated it mApo. Although *ERBB2/HER2* was overexpressed by 72 % of the tumors in this subgroup, cluster-VIII did not comprise genes co-amplified with *ERBB2/HER2*. In fact, ERBB2/HER2+ tumors distributed in mApo and

LumC subgroups (Table 1). Finally, the AR-/ER-/PR- subgroup, defined by cluster-I, presented the greatest distance to all others (Figure 1). Since it shared genes with the “basal-like” subtype, it was designated BasL (SupTab3).

Molecular subgroups show distinct clinical correlations, metastatic sites and outcomes

BasL and mApo at one end of the spectrum and LumA and NormL at the other end showed an inverse balance between high-grade and ER/PR positivity (Table 1). TP53 mutation incidence reached 83% in the BasL subgroup and gradually went down to 4% in NormL and LumA tumors (Table 1). This distribution of high grade/ER- vs. low grade/ER+ cancers was also coherent with the median age of onset: 50 and 62 for BasL and LumA patients, respectively. Correlation with histological type was observed as well. While the BasL subgroup was composed of 98% ductal carcinomas, NormL presented 19% of invasive lobular tumors, representing 53% of all lobular cancers in the dataset, in coherence with previous findings (Bertucci et al., 2008).

Molecular subgroups showed differences in sites of metastatic relapse. In line with previous studies (Smid et al., 2008), LumA and NormL predominantly metastasized to the bone and rarely or never to the brain, while BasL and mApo tumors metastasized to the brain and less to the bones (Table 1). *ST6GALNAC5*, *COX2/PTGS2* and *HBEGF,* whose expression has recently been associated to brain metastasis (Bos et al., 2009), were increased in BasL (SupFig3). Clear differences were also found in metastasis-free survival (MFS) (Figure 2). BasL and mApo subgroups showed earliest recurrence (18 to 60 months). LumA and NormL had the slowest course. Whereas, metastatic recurrence plateaued between 60 and 180 months in BasL and mApo, it progressively increased after 60 months in ER+ subgroups. LumA and NormL tumors presented recurrences after 120 months post-surgery. Interestingly, patterns of recurrence (early vs. late) matched cell cycle cluster expression levels in the different subgroups.

Performance on external datasets

We applied our classification scheme to a large Affymetrix dataset comprising 2291 breast cancer transcriptomes we have collected from the litterature (SupMethods). The 6 molecular subgroups were perfectly reproduced, both in terms of distribution and clinical correlations and outcomes (SupTab4a, Figure 2B). To further ascertain its robustness, we tested our classification on 3 expression datasets from different technological platforms (Swegene, Qiagen/Operon and Agilent). Our prediction rule being designed for Affymetrix datasets we had to adapt it to different technological contexts (SupMethods). Overall molecular subgroups were reproduced on different platforms (SupTab4b, SupFig4). Differences were noted according to the dataset, which may possibly be due to different tumor recruitment in each series. To test inter-platform reproducibility we classified the GSE3155 dataset which was analyzed in parallel on 2 dual-color (Agilent and Stanford) and 1 uni-color (Applied Biosystems) platforms (SupTab4c). Classification on both dual-color datasets showed a 90% overlap, suggesting a good inter-platform reproducibility. However, overlap dropped dramatically when dual and uni-color platforms were compared (48 and 52%). This indicates that classification rules need adaptation to technological specificities of each platform to perform optimally.

Comparison with other molecular classifiers

We next compared our classification with the Sorlie, Hu and Parker centroids (Sorlie et al., 2003; Hu et al., 2006; Parker et al., 2009). Variable overlaps were found for BasL, LumB, LumA and NormL subgroups (Figure 1D). However, significant differences were noted for the mApo and LumC subgroups which overlapped at variable levels with the ERBB2 subtype, but also with basal-like, luminal A and B and normal-like groups, depending on the classifier (SupTab5). Classification differences affected the distribution of bioclinical markers among molecular subgroups. Main differences were in the fraction of ER+/PR+ and AR+ tumors in basal-like subtypes and the distribution of ERBB2 positive tumors (SupTab6). MFS curves showed better separation of good and bad outcome subgroups with the CIT classification (SupFig5, SupFig6).

Molecular subgroups show differential activation of signaling pathways

We selected 40 cancer relevant pathways from public databases and tested for specific enrichment in our molecular subgroups (SupMethods). Genes specific for each subgroup were identified using 4 algorithms. Pathways were ranked for each subgroup on the mean of p-values across the four methods. As shown in Figure 3 each subgroup was associated to different up or downregulated signaling pathways. The upregulation of DNA replication and repair in BasL and LumB contrasted with its downregulation in NormL. The upregulation 4/5 immune system pathways in LumC was of further note. These data indicate that molecular subgroups relate to different signaling pathways and biological processes.

Molecular subgroups show specific genomic anomalies

Of the 537 tumors profiled for RNA expression, 488 tumors were analyzed by array-CGH. Twenty-one regions of gain and 33 regions of loss were found in more than 30% of the tumors (Figure 4A, top panel). BasL and LumB showed extensive copy number alterations (CNA), whereas NormL and LumA were the least rearranged. Qualitative differences were also apparent (Figure 4A) and we searched for CNAs specifically associated to each subgroup. BasL and LumB tumors presented the greatest number of CNAs with respectively 39 and 46 specific CNAs (Figure 4A, SupTab7). The number of specific events was lower in the other subgroups ranging from 2 to 8. Expectedly, amplifications at 17q12 were found in 70% of mApo tumors. LumA showed gains at 4q35 and 16p11-p13, whereas NormL tumors could be differentiated from LumA by gains at 9q33, 8p23, 16p13 and loss at 16q12.

CNAs were associated to large-scale gene expression modifications. A total of 786 genes comprised in intervals of gains or losses showed significantly modified expression levels. A number of regions of gains overexpressed genes encoding cell cycle and proliferation activators and, conversely, known tumor suppressors, pro-apoptotic or DNA damage checkpoint genes were found downregulated in regions of loss (SupTab7). These findings suggest that CNAs are part of a selective process associated with tumor progression, with differences from one subgroup to another. In that respect, 28 CNAs presented inverse patterns in different subgroups. These inverted patterns involved mainly BasL and LumB, but were also found in mApo and LumB or LumB and NormL (SupFig7). Strikingly, they were associated to inverse expression of key cancer genes. These data support the notion that breast cancer subgroups arise along distinct genetic pathways.

Focal DNA amplification (defined as high-level gains occurring in regions not larger than 3 Mb) occurred significantly more frequently in LumB, mApo and LumC than in the other subgroups (SupTab8a). We further investigated the occurrence of focal CNAs and analyzed a subset of 72 tumors from the CIT discovery set with high resolution Illumina 610K-SNP-arrays (SupTab8b). We detected 246 gains and 337 losses (mean size 132 and 161 kb respectively). We noted that 53% of the gains were also detected in our BAC-array data, while the overlap was lower for losses (19%). However, gains showed modest copy number increase and were infrequently recurrent. Losses showed greater recurrence but this corresponded mainly to probable CNVs (identical starts and ends).

We verified the overlap of our subgroups with the recently proposed CNA-based classification (Jönsson et al., 2010) and observed an overall coherence with our findings. Their CNA-based Basal-complex class overlapped with our BasL, 17q12 with part of our mApo and LumC, Luminal complex and amplifier with LumB and LumC, while the Luminal-simple corresponded globally to LumA and NormL (Figure 1D).

Fraction of non-tumor cells and distribution in molecular subgroups

The fraction of non-tumor cells is frequently discussed as a confounding factor in molecular analyses of breast cancer fostering the proposition that the normal-like group was a possible artefact (Prat et al., 2010). **T**o get an objective estimate of the rate of non-diploid cells in our dataset and determine its distribution within molecular subgroups, we computed the Illumina

610K-SNP data using a recent formula (Van Loo et al., 2010). Significant differences were seen among molecular subgroups (SupFig8a), with, surprisingly, mApo showing the lowest rate of non-diploid cells. NormL ranked third and LumA and LumB presented the highest fraction of non-diploid cells. Our results agreed with recent data (Van Loo et al., 2010). However, a variable fraction of tumor cells may also be diploid, leading to an overestimation of normal cells. To assess this, a histological estimate of the non-tumor cell fraction was performed on the tumors analyzed with the Illumina 610K-SNP-arrays. This showed that SNP-based estimates of non-diploid cells were lower than pathological tumor cell content (SupFig8b). Overall these data are coherent with the idea of NormL representing a bona-fide breast cancer subgroup.

Breast cancer subgroups and mammary epithelial cell hierarchy

To test whether our subgroups relate to distinct cells of origins in the mammary gland, we took advantage of 3 published expression profiling datasets of sorted normal mammary epithelial cell subpopulations (Lim et al., 2009; Pece et al., 2010; Raouf et al., 2008). We inferred a signature that discriminated the Mammary Stem Cell (MaSC) enriched, luminal progenitor (LPC), mature luminal (MLC) and stromal cell populations and used this signature to classify our breast tumor expression data (SupMethods). As shown in Figure 5, the PCA ordered normal mammary epithelial cell fractions according to a differentiation gradient and breast tumors from BasL, mApo, LumC, LumB /NormL to LumA, suggesting a proximity of BasL and mApo with either MaSC or LPC, whereas ER-positive subgroups showed a gradient between LPCs and MLCs. The correlation of BasL and mApo with least differentiated cells (MaSC or LPC) in the normal mammary gland was confirmed in a second analysis (SupTab9).

Prognostic significance of molecular subgroups,

We next compared the prognostic significance in terms of metastatic relapse of our molecular subgroups to classical prognostic factors (ER, ERBB2/HER2, SBR grading and nodal involvement). As shown in Table 2, our classification signature performed better in both univariate and multivariate analyses than the classical prognostic factors, in both the discovery and validation sets. However, the absence of central pathology review in both datasets prevents us to draw firm conclusions on the independent prognostic power of our signature. In a comparative analysis with 5 expression signatures (Hu et al., 2006; Parker et al., 2009; Sorlie et al., 2003; Sotiriou et al., 2006; van 't Veer et al., 2003), our signature came second after the van't Veer signature in the discovery set and performed best in the validation set (SupTab10), demonstrating the important difference in terms of prognosis among molecular subgroups.

Molecular subgroups show differential response to chemotherapy

To test whether our classification could predict chemotherapy response, we analyzed 3 datasets of locally advanced breast cancers treated by neoadjuvant therapy followed by surgery and assessment of the pathological response. ER-negative breast cancers were overrepresented in the 3 cohorts, but our signature allowed the assignment of tumors to 4 subgroups after pooling LumB and LumC, as well as LumA and NormL to reach sufficient sample size by subgroup. Despite different chemotherapy protocols in individual cohorts, obvious differences in response were observed. BasL and mApo showed the best response rates with respectively 44 and 65% and 37 and 42% of complete response. ER-positive subgroups showed 19% of complete response in LumB/LumC tumors and 0% in LumA/NormL(Table 3A). Prediction of complete pathological response (pCR) of the CIT classification was then compared to that ER status and SBR Grade in the three pooled datasets. Both in the univariate and multivariate analysis the CIT classification showed the strongest score (Table 3B).

Discussion

Breast cancer heterogeneity, reflected in molecular subgroups, can be attributed to differences in molecular alterations, cellular origin or both. We present a classification of breast cancer into 6 molecular subgroups, which differed upon gene expression, genomic profiles, differentiation level and clinical features.

First, gene expression differences strongly suggested that they outlined distinct biological entities, reflecting founder mutations and/or cell-of-origin. Specific sets of signalling pathways were associated to each subgroup.

The distribution of the 6 subgroups was determined by the combination of the expression of three large gene clusters organized around (i) the estrogen receptor, (ii) androgen receptor, (iii) cell cycle regulator genes. The ER cluster is well known as defining luminal breast tumors (Bertucci et al., 2006) and the expression of AR in breast cancer is long-known (Isola, 1993), but has been confounded with that of the ER cluster (Doane et al., 2006). Its combined expression with the ER cluster yields 3 broad classes determined by nuclear receptor expression; AR-/ER-/PR- (triple negative) corresponding to the BasL subgroup, AR+/ER-/PR- (mApo), AR+/ER+/PR+ (triple positive) including the 4 ER+ subgroups. The AR cluster comprises key genes previously associated to the ER cluster, such as the pioneer factor FOXA1, which recruits ER, AR and RAR/RXR (Carroll et al., 2006; Lupien et al., 2008).

The existence of an ER-/AR+ breast tumor subset (our mApo subgroup) has been proposed (Doane et al., 2006; Farmer et al., 2005) and its important overlap with *ERBB2/HER2* amplification is intriguing, possibly reflecting cross-talks between the AR and ERBB2/HER2 pathways (Naderi & Hughes-Davies, 2008). However, it is notable that our classification did not define an ERBB2 subgroup. Instead, ERBB2-amplified cancers distributed in mApo (ER-) and LumC (ER+) subgroups. We found less expression differences between mApo/ERBB2+ and mApo/ERBB2- than between mApo and LumC tumors (SupFig9). Interestingly, Staaf and coworkers (Staaf et al., 2010) showed that ER- and ER+ ERBB2-amplified tumors presented different 17q CNA patterns. These observations could have implications in the clinic as they indicate that ERBB2+ breast cancer correspond to a biologically heterogeneous group. Moreover, it seems important to distinguish ERBB2+ and mApo tumors, because the so-called triple-negative group comprises both BasL and ERBB2-/mApo tumors despite clear molecular and clinical differences.

Second, subgroups were also characterized by different patterns of genomic anomalies. These data were concordant with previous results (Chin et al., 2006; Natrajan et al., 2009) and the CGH classification recently proposed by Jönsson and coworkers (Jönsson et al., 2010). Moreover, the existence of chromosomal regions showing inverse patterns (gain in one subgroup/loss in another) further supported the notion that these subgroups progress along distinct genetic routes, which possibly involve different mechanisms of genetic instability.

Third, our data indicated that subgroups differed in their differentiation level, pointing to possible differences in cell-of-origin. This was suggested by similarities between the transcriptome of distinct cellular contingents in the normal mammary gland and those of molecular subgroups. While BasL and mApo showed proximity to MaSC or Luminal Progenitors, ER+ subgroups formed a gradient between LPCs (LumC) and mature luminal cells (LumA). Our findings are consistent with recent work suggesting that LPCs were the cells of origin of basal cancer and Brca1 mammary tumors (Lim et al., 2009; Molyneux et al., 2010). These findings bring insight on the prevalence of grade 3 tumors in BasL and mApo contrasting sharply with that of low grade cancers in NormL and LumA. Our data thus suggest that breast cancer may arise from at least two distinct cell types and that the final phenotype will result from genetic and epigenetic changes occurring during cancer progression. This may also have some link with the striking gradient of *TP53* mutations observed between BasL and NormL subgroups. The correlation with elevated expression of the cell-cycle cluster and increased genomic instability was also notable. Moreover, there is a striking parallel between the incidence of TP53 inactivation and the response rates of neo-adjuvant chemotherapies. These data are in line with our previous observation proposing that TP53 is not the mediator of chemotherapy-induced cell death (Bertheau et al., 2007).

Fourth, molecular subgroups show striking differences with respect to metastatic relapse both in terms of kinetics and site of recurrence. While BasL and mApo tumors preferentially metastasized to the brain and rarely to the bone, ER+ subgroups exhibited an inverse pattern, strengthening previous studies (Smid et al., 2008). Our data suggest that these differences could be due to differential expression of key metastasis genes (Bos et al., 2009). Hence, metastasis to a specific organ can also be the result of a subgroup-specific gene program and coexist with the *de novo* acquisition of stochastic mutations, as recently shown by massively parallel sequencing work (Ding et al., ; Yachida et al.). Outcomes of the different subgroups were very different as well. BasL and mApo showed earlier relapse, but a remarkably stable metastasis-free survival for the next 100 months. In contrast, while all ER+ subgroups did better during the first years, a continuous incidence of late relapse was observed. LumB and LumC outcome progressively became worse than that of BasL or mApo. However, a number of recurrences occurring after 5 years in ER+ subgroups are probably linked to interruption of anti-estrogen treatments.

The status of the NormL subgroup is of particular interest because its existence has been put in doubt and attributed to an elevated content of normal cells (Prat et al., 2010). In line with recently published data (Van Loo et al., 2010), we showed that NormL tumors did not present a lower fraction of non diploid cells than mApo or LumC. Furthermore, our data showed that 70% of NormL tumors showed loss at 16q, further supporting that this subgroup does not result from a co-cluterization of breast tumors presenting smaller fractions of tumor cells.

Our results are in favor of the existence of different breast cancer subtypes bearing distinct biologies and clinical courses. We propose that stratifying breast cancers according to such a classification could be highly beneficial when searching for new prognostic or response to treatment indicators. These would be subgroup specific instead of expressing the differences between highly and poorly proliferating tumors. Furthermore, such a classification, once adapted in a format compatible with clinical setting, could efficiently contribute to disease management. Indeed, the different subgroups outlined here occur in different age groups, metastasize to different organs and exhibit distinct survival kinetics. Similarly, the association with immune system activation pathways in LumC may be indicative for an anti-tumor immunity in this specific subgroup. All of these are clear indications that they represent distinct clinical and biological entities.

Materials and methods

Patients and tumors

A total of 724 primary breast carcinomas were collected and analyzed for expression profiling on Affymetrix U133-Plus2.0 chips and a subset of 488 samples were analyzed by array-CGH. In addition, 58 fine needle aspiration biopsies from patients undergoing neoadjuvant chemotherapy were analyzed by transcriptome and included in the response-to-chemotherapies set. Full description can be found in SupTab1 and SupTab2. Mean follow up time was of 65 months. Four RNA from normal human breast tissue were used as reference. Histological grade as well as ER, PR and HER2 levels determination are detailed in the SupMethods.

Discovery and validation sets

Our 724 breast tumor transcriptome dataset was split in a CIT-discovery-set comprising 537 (75%) tumors of which 488 were analyzed by array-CGH and 187 (25%) cases that were set apart for the validation-set. The Affymetrix validation-set comprised the 187 samples from CIT and 2291 transcriptomes collected from GEO and array-express (SupTab2).

Expression profiling and data analysis

RNA profiling

Methods used for RNA purification, quality control, fluorescent probe production, hybridization and data processing were essentially as previously described (de Reynies et al., 2009).

Transcriptome analysis and molecular subgroup determination

Our rational was to ensure the greatest possible homogeneity to identified subgroups. Subgroup determination was based on the CIT discovery set including 537 transcriptomes and a clustering approach iterating unsupervised and supervised steps (SupFig1, SupMethods). Microarray data were first classified with a set of 244 most variant probesets using in parallel Hierarchical clustering, k-means and Gaussian mixture model. Tumors that were assigned to the same group by the 3 methods were kept defining a core-set of 355 tumors. Based on this core-set most discriminative genes were selected by ANOVA and ranked by random-forest, producing a 256 gene signature, leading to the identification of 6 homogeneous molecular subgroups. Validation datasets were independently classified in the CIT molecular subgroups by applying a classical distance-to-centroid approach, implemented in the citbcmst R package available at the following URL http://cran.r-project.org/web/packages/citbcmst/index.html. and coming with a (Sweave) user documentation. The complete classification procedure is detailed in the SupMethods.

Comparison with the Sorlie, Hu and Parker classifiers

Sorlie (Sorlie et al., 2003), Hu (Hu et al., 2006) and Parker (Parker et al., 2009) centroids were respectively retrieved from

http://genome-www.stanford.edu/breast_cancer/robustness/data/IntrinsicGeneList.txt,

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

https://genome.unc.edu/pubsup/breastGEO/pam50_centroids.txt.. To build the classifiers corresponding clone UniGene_IDs were mapped to Affymetrix (U133A or U133Plus2) probe-sets. For Sorlie this was possible for 334 UniGene_IDs gene symbols, for Hu 232 UniGene_IDs and Parker all genes could be directly mapped.

Comparison with the Jönsson array-CGH-based classification

The 6 Jönsson centroids are relative to genomic regions determined with the GISTIC algorithm (Jönsson et al., 2010). Details are provided in the SupMethods.

Cancer pathways analysis

Cancer relevant pathways were retrieved from KEGG (ftp://ftp.genome.ad.jp/pub/kegg/pathways/hsa), Biocarta (http://www.biocarta.com) and GO (http://www.geneontology.org/) and related genes mapped to non-redundant HUGO Gene symbols. Four gene set analysis methods were used (SupMethods) yielding p-values based which were transformed in ranks. Gene sets were ranked by order of interest according to the mean of the ranks across the four methods.

Array-CGH

Array-CGH was performed on a 4434 BAC-array with a median inter-BAC gap of 0.6 Mb. DNA labeling, hybridization and data processingare as described in the SupMethods.

Statistical tests

Clinical correlations were determined by chi-square for qualitative factors and ANOVA for quantitative variables. Disease outcome was investigated with Kaplan-Meier curves using metastatic recurrence as an endpoint and subgroup for stratification. Metastasis-free survival (MFS) was calculated from the date of diagnosis until first metastatic relapse. P-values at 60 and 180 months resulted from a log-rank test on Cox estimates. Benjamini and Hochberg (BH) method was applied for multiple-testing adjustment.

Acknowledgements This work was supported by the Carte d'Identité des Tumeurs Program from the *Ligue Nationale Contre le Cancer.* The authors wish to thank Pr Jacqueline Godet and Dr Jacqueline Métral for their constant support during the course of this work. The expertise of the personnel serving the RNA extraction and qualification (Saint-Louis Hospital, Paris), Affymetrix expression array (IGBMC, Strasbourg), CIT database and array-CGH (IRCM, Montpellier) CIT platforms is gratefully acknowledged. Further thanks are granted to Dr Stan du Manoir for critical reading of the manuscript, Vessela N. Kristensen and Peter Van Loo for making their data available to us.

References

Bertheau P, Turpin E, Rickman DS, Espie M, de Reynies A, Feugeas JP*, et al.* (2007). Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen, *PLoS Med*, **4**: e90.

Bertucci F, Finetti P, Cervera N, Maraninchi D, Viens P ,Birnbaum D. (2006). Gene expression profiling and clinical outcome in breast cancer, *Omics*, **10**: 429-443.

Bertucci F, Orsetti B, Negre V, Finetti P, Rouge C, Ahomadegbe JC*, et al.* (2008). Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles, *Oncogene*, **27**: 5359-5372.

Bos PD, Zhang XH, Nadal C, Shu W, Gomis RR, Nguyen DX*, et al.* (2009). Genes that mediate breast cancer metastasis to the brain, *Nature*, **459**: 1005-1009.

Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J*, et al.* (2006). Genome-wide analysis of estrogen receptor binding sites, *Nat Genet*, **38**: 1289-1297.

Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo W-L*, et al.* (2006). Cancer Cell, Vol. 10, pp 529-541.

de Reynies A, Assie G, Rickman DS, Tissier F, Groussin L, Rene-Corail F*, et al.* (2009). Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival, *J Clin Oncol*, **27**: 1108-1115.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW*, et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft, *Nature*, **464**: 999-1005.

Doane AS, Danso M, Lal P, Donaton M, Zhang L, Hudis C ,Gerald WL. (2006). An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen, *Oncogene*, **25**: 3994-4008.

Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D*, et al.* (2005). Identification of molecular apocrine breast tumours by microarray analysis, *Oncogene*, **24**: 4660-4671.

Foulkes WD, Smith IE ,Reis-Filho JS. (2010). N Engl J Med, Vol. 363, pp 1938-1948.

Gusterson B. (2009). Do 'basal-like' breast cancers really exist?, *Nat Rev Cancer*, **9**: 128-134.

Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF*, et al.* (2006). BMC Genomics, Vol. 7, pp 96.

Hua S, Kittler R ,White KP. (2009). Genomic antagonism between retinoic acid and estrogen signaling in breast cancer, *Cell*, **137**: 1259-1271.

Isola JJ. (1993). Immunohistochemical demonstration of androgen receptor in breast cancer and its relationship to other prognostic factors, *J Pathol*, **170**: 31-35.

Jönsson G, Staaf J, Vallon-Christersson J, Ringnér M, Holm K, Hegardt C*, et al.* (2010). Breast Cancer Res, Vol. 12, pp R42.

Kapp AV, Jeffrey SS, Langerod A, Borresen-Dale AL, Han W, Noh DY*, et al.* (2006). Discovery and validation of breast cancer subtypes, *BMC Genomics*, **7**: 231.

Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH*, et al.* (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers, *Nat Med*, **15**: 907-913.

Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W*, et al.* (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription, *Cell*, **132**: 958-970.

Molyneux G, Geyer FC, Magnay F-A, McCarthy A, Kendrick H, Natrajan R*, et al.* (2010). Cell Stem Cell, Vol. 7, pp 403-417.

Naderi A ,Hughes-Davies L. (2008). A functionally significant cross-talk between androgen receptor and ErbB2 pathways in estrogen receptor negative breast cancer, *Neoplasia*, **10**: 542-548.

Natrajan R, Lambros MB, Rodríguez-Pinilla SM, Moreno-Bueno G, Tan DSP, Marchió C*, et al.* (2009). Clin Cancer Res, Vol. 15, pp 2711-2722.

Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T*, et al.* (2009). J Clin Oncol, Vol. 27, pp 1160-1167.

Pece S, Tosoni D, Confalonieri S, Mazzarol G, Vecchi M, Ronzoni S*, et al.* (2010). Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content, *Cell*, **140**: 62-73.

Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI*, et al.* (2010). Breast Cancer Res, Vol. 12, pp R68.

Raouf A, Zhao Y, To K, Stingl J, Delaney A, Barbara M*, et al.* (2008). Transcriptome analysis of the normal human mammary cell commitment and differentiation process, *Cell Stem Cell*, **3**: 109-118.

Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG*, et al.* (2008). Subtypes of breast cancer show preferential site of relapse, *Cancer Res*, **68**: 3108-3114.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A*, et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets, *Proc Natl Acad Sci U S A*, **100**: 8418-8423.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J*, et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis, *J Natl Cancer Inst*, **98**: 262-272.

Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J, Grabau D, Arason A*, et al.* (2010). Breast Cancer Res, Vol. 12, pp R25.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Bernards R ,Friend SH. (2003). Expression profiling predicts outcome in breast cancer, *Breast Cancer Res*, **5**: 57-58.

Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W*, et al.* (2010). Proc Natl Acad Sci USA, Vol. 107, pp 16910-16915.

Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DS, Dowsett M*, et al.* (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis, *Lancet Oncol*, **11**: 339-349.

Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B*, et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer, *Nature*, **467**: 1114-1117.

Legends to the Figures

Figure 1: **Breast tumor classification according to the CIT classification into 6 subgroups of tumors**. **(A)** Heatmap representing the expression of the 256 genes (8 clusters of genes represented by vertical color bars on the left of the heatmap) through the 6 groups. **(B)** Principal-Component-Analysis (PCA) of the samples of the core-set according to the 256 gene signature. The first principal component (PC1) represents the combined expression of the 3 transversal clusters (ER, AR and cell cycle), the second component (PC2) differentiates LumB and NormL**. (C)** Distribution of mean expression levels of the three transversal gene clusters (ER, AR and Cell Cycle) over the six main molecular subgroups. **(D)** Comparison of the CIT classification with those obtained using the Sorlie, Hu, Parker and Jönsson systems.

Figure 2: **Breast cancer molecular subgroups show distinctly different disease outcome**. Kaplan-Meier curves shown in this Figure represent disease free survival with metastatic relapse as an end point. Panels A and B show survival curves in the CIT and validation set respectively. Abrupt breaks in some curves of panel A are related to small numbers of patients with long term follow up in these subgroups. These appear smoothed out in panel B because of greater numbers in the validation set.

Figure 3: **Molecular subgroups show differential activation of major signaling pathways:** Correlations between a given pathway and a subgroup are indicated by color boxes. Red boxes show up-regulation of the pathway, green down-regulation. Up or down-regulation was deduced using KEGGanim tool where relative expression measures are projected in the related KEGG pathway interaction graph. Pathways showing no clear direction of regulation were excluded.

Figure 4: **Breast cancer molecular subgroups present different copy number change (CNC) profiles.** CNC profiles were established using genome wide array-CGH on the a 488 breast tumor dataset and subsequently stratified according to the CIT classification. The left panel shows frequency of gains (vertical bars going up) or losses (bars going down) at a given location on the genome. Graphs from top to bottom correspond to profiles of the whole CIT breast cancer set and each of the six molecular subgroups. The right panel represents regions of CNC correlating to a specific subgroup. Specific genomic regions for the whole CIT set are the ones for which the proportion of alterations (in gain or loss) exceeded 20%. Subgroup specific regions are those that present significant increase in proportion (at a 0.1 FDR level) in a given subgroup tested against all others. Bars represent p-values after a standard logarithmic transformation.

Figure 5: **Principal Component Analysis of the CIT core-set expression profiles based on a meta-signature comparing normal mammary epithelial cell subpopulations.** A 163 gene signature was produced by comparing different normal mammary cell contingents from 3 independent studies (GSE16997, GSE18931, GSE11395) and used in a principal component analysis. Samples from the CIT core-set and normal mammary gland samples from GSE16997 were projected in the 2 first principal components in the upper and lower panel respectively.

Legends to the Tables

Table 1: molecular subgroups show differential correlation to breast cancer clinico-biological parameters and different sites of metastatic relapse. Expression of ER, PR and ERBB2/HER2 were determined by immunohistochemistry as well as by RNA expression (for greater details see SupMethods). TP53 mutation status was determined by the yeast functional assay (SupMethods). P-values for qualitative variables (ER, PR, ERBB2/HER2, TP53 mutation, histological type, SBR grading) result from a Fisher exact test. P-values for quantitative variables (median age) result from an ANOVA. Metastasis relapse (MR) was

determined 5 and 15 years after surgery. Frequency of MR in a subgroup was calculated as the ratio of MR with the total number of MR. For each subgroup, percentages of M.R. in a given site are determined by the number of M.R. in this site over the whole number of MR in the subgroup. MR may occur at more than one site, hence the sum of percentages may not equate 100.

Table 2: **Prognostic significance of the CIT classification.** Relative risk was calculated taking metastatic relapse as an endpoint and compared to that of (A) clinical parameters and (B) of 3 molecular classifiers (Sorlie, Hu, Parker) and 2 prognostic signature (GGI, Van't Veer). The dataset comprised 426 patients from the CIT discovery set for which MFS information was available. Complete clinical information was available in 371 cases explaining the smaller numbers in the multivariate analysis on prognostic factors. Prognostic significance was assessed by applying a Cox model. Columns refer to the Hazard Ratio (HR), the 95% confidence-interval (CI) and the p-values for both univariate and multivariate models.

Table 3: **Differential response to chemotherapy according to molecular subgroups of the CIT classification.** Table 3A shows the correlation between pathological complete response (pCR) and CIT molecular subgroups. Pathologic complete response (pCR) and absence of response (no pCR) to chemotherapy was analyzed in three clinical trials (Hess et al 2006, Bonnefoi et al 2007, CIT set). Due to the small number of data, four main subgroups and two intermediate subgroups were combined into two groups: [LumB; LumC; LumB/C] and [NormL; LumA; NormL/LumA]. Treatment description**: (EC)** six cycles of a dose-dense regimen of 75 mg/m^2 epirubicin and 1,200 mg/m^2 cyclophosphamide, given every 14 days; **(T/FAC)** 24 weeks of sequential paclitaxel and fluorouracil-doxorubicin-cyclophosphamide; **(FEC)** fluorouracil, epirubicin, and cyclophosphamide for six cycles; **(TET)** docetaxel for three cycles followed by epirubicin plus docetaxel for three cycles. Correlations were calculated using Fisher exact test. Table 3B shows uni and multivariate analyses of factors predictive of

pCR in the three pooled datasets. Univariate analysis was done using the Fisher exact test and Multivariate analysis by logistic regression.