SUPPLEMENTARY SECTION

Identification of a gene expression signature (GES) for medullary breast cancer

Results: identification of GES

A first analysis, the comparison between basal MBCs and basal DBCs is reported in the core text of manuscript. A second supervised analysis, reported here, searched for a GES that would discriminate between all 22 MBCs and all 44 DBCs (ER-negative and ER-positive).

We identified 527 genes (527-GES) as discriminator between MBCs and DBCs (theoretical number of produced false positives inferior to 1), with 311 genes overexpressed and 216 underexpressed in MBCs. They represented 413 different sequences, corresponding to 378 characterized genes and 35 ESTs (Supplementary Table 4). Onto-Express biological processes represented by genes overexpressed in MBCs were “immune response” (GO :0006955 ; 20 genes, $p<0.001$), “cell surface receptor-linked signal transduction” (GO:0007166 ; 8 genes, $p<0.001$), “proteolysis and peptidolysis” (GO:0006508 ; 12 genes, $p<0.01$), “cell proliferation” (GO:0008283; 8 genes, $p<0.01$). Conversely, the biological process more active in DBCs versus MBCs was “regulation of transcription, DNA-dependent” (GO:0006355; 22 genes, $p=0.01$). Concerning chromosomal location, there was a prominence of the 12p13 region for genes upregulated in MBCs ($p<10^{-8}$, Fisher's exact test). The classification power of this GES is illustrated in Supplementary Figure 2A. A threshold of 0 (orange solid line in Supplementary Figure 2A) sorted the samples into two classes ("predicted MBC class", positive scores; "predicted DBC class", negative scores) that

correlated with the pathological type: all MBCs classified in the "predicted MBC class", and 32 of 44 DBCs in the "predicted DBC class" ($p<0.001$, Fisher exact test). By LOO cross-validation, 79% of samples were correctly assigned by the predictor, suggesting the validity of the procedure.

Results: validation of GES

Validation of this GES was obtained by RNA profiling of another series of samples on another microarray platform. Two discriminator genes (*GATA3* and *MSN*) were also validated at the protein level on a larger series of samples screened on TMA.

As a first validation study, we analyzed personal data from 73 SBR grade III breast cancers (6 MBCs and 67 DBCs) treated at IPC, and profiled with cDNA-spotted microarrays (IPC/Ipsogen data set). Thirty-two samples (3 MBCs and 29 DBCs) were common to the present series, and 41 were independent. After filtering, 3.891 genes were available for analysis. Sixty-two genes of the 527-gene signature were included in the 3.891 genes. Based on these 62 genes, hierarchical clustering of our 66 present samples identified two major groups closely associated with the pathological type ($p<0.001$, Fisher's exact test – Supplementary Figure 2B, left). We used the 62 genes to cluster the 73 samples of the IPC/Ipsogen data set (Supplementary Figure 2B, right). We identified a small group of 13 samples that included all the MBCs ($p<0.001$, Fisher's exact test). Importantly, similar significant discrimination persisted when applied to the 41 independent samples ($p=0.003$, Fisher's exact test). Because the number of genes common to the intrinsic gene set and the 3.891 genes was too small, we could not assign a molecular subtype to the 205 samples and evaluate the robustness of the 534-GES in basal samples.

As a second validation study, we studied proteins corresponding to discriminator genes by IHC on a larger series of samples. We selected *GATA3* and *MSN* (moesin) because of a putative role in mammary oncogenesis (1, 2), the availability of a corresponding monoclonal antibody that well performed on paraffin-embedded tissues and their opposite expression patterns; RNAs from *GATA3* and *MSN* were respectively underexpressed and overexpressed in MBCs as compared to DBCs. A total of 547 breast cancers, including 385 DBCs (97 ER-negative and 288 ER-positive), were available in TMA1. A second TMA (TMA2) contained 40 MBCs. Examples of IHC results are shown in Supplementary Figure 2C. First, we confirmed the correlation between RNA and protein expression in the 55 tumors common to gene profiling and TMAs. The GATA3 protein (cut-off = 1) was significantly overexpressed in DBCs (37.5% of informative DBCs versus 0% of MBCs; p=0.01, Fisher's exact test), and moesin (cut-off = 20) in MBCs (85% of MBCs versus 29% of DBCs, p=0.001, Fisher's exact test). We then validated the differential protein expression between MBCs and DBCs in the 370 samples specific to TMAs: GATA3 was overexpressed in DBCs (60% of DBCs versus 0% of MBCs, p<0.001, Fisher's exact test) and moesin in MBCs (25% of MBCs versus 13%, p=0.15, Fisher's exact test).

Discussion

A total of 527 genes discriminated MBCs from DBCs. Onto-Express showed that "immune response" was the most represented biological processes in MBCs versus DBCs. *ETV6*/*TEL*, located at 12p13, was the most upregulated gene in MBCs. It encodes a transcription factor of the ETS family, frequently rearranged in several types of cancer. The *ETV6-NTKR3* gene fusion is the primary event in secretory

breast carcinoma (3). Overexpression of *ETV6* in MBCs might be related to 12p13 amplification and/or stem cell biology (see below). *NFKB2* and several genes involved in the positive regulation of the NF-$_K$B cascade (*RELB*, *TNFRSF10B*, *ATP2C1*, *IKBKE*, *CFLAR*, *IL15*, *IL15RA*, *VCAM1*, *TNFRSF6*, *IL2RA*, *CSF1, IRAK1*) were upregulated in MBCs. Activation of the NF-$_K$B pathway is prominent in ER-negative breast cancers (4) and is consistent with the presence of an inflammatory stroma and the high proliferation rate of MBCs. Examples of basal genes overexpressed in MBCs include *MSN* and *ICAM1*. *MSN* codes for moesin, a protein important for cell-cell recognition, signaling and cell movement. Its expression is associated with ER-negative breast cancer (5). We recently found *MSN* gene and moesin protein overexpression in basal and medullary breast cell lines as compared to luminal ER-positive cell lines (2). The same trend was observed, although at the limit of significance, in our independent validation panel of tumors. *ICAM1* codes for intercellular adhesion molecule 1 (CD54). It is overexpressed in MBCs as compared to DBCs (6), and likely contributes to the formation of a leukocyte infiltrate in MBCs.

The most underexpressed gene in MBCs was *FOXA1*, which codes for transcription factor HNF3A (hepatocyte nuclear factor 3-alpha). HNF3A expression correlates with that of ER in breast tumors (7). Several other genes downregulated also code for transcription factors (*ESR1*, *GATA3*, *XBP1*, *RARA*, *RXRA, SPDEF*), as confirmed with Onto-Express, which identified “regulation of transcription, DNA-dependent” as the most active process in DBCs versus MBCs. These results were consistent with the phenotype of MBCs (ER-negativity, dense lymphocyte infiltrate, high mitotic index, and basal subtype).

References

1. Mehra R, Varambally S, Ding L*, et al.* Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. Cancer Res. 2005;65*:*11259-64.
2. Charafe-Jauffret E, Ginestier C, Monville F*, et al.* Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene. 2006;in press.
3. Tognon C, Knezevich SR, Huntsman D*, et al.* Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell. 2002;2*:*367-76.
4. Biswas DK, Shi Q, Baily S*, et al.* NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. Proc Natl Acad Sci U S A. 2004;101*:*10137-42.
5. Carmeci C, Thompson DA, Kuang WW, Lightdale N, Furthmayr H, Weigel RJ. Moesin expression is associated with the estrogen receptor-negative breast cancer phenotype. Surgery. 1998;124*:*211-7.
6. Bacus SS, Zelnick CR, Chin DM*, et al.* Medullary carcinoma is associated with expression of intercellular adhesion molecule-1. Implication to its morphology and its clinical behavior. Am J Pathol. 1994;145*:*1337-48.
7. Lacroix M, Leclercq G. About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. Mol Cell Endocrinol. 2004;219*:*1-7.
8. Sorlie T, Tibshirani R, Parker J*, et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003;100*:*8418-23.

Legend of Supplementary Figures

Supplementary Figure 1: Hierarchical clustering of expression data from Sorlie *et al*. (8) based on 476 common genes and 122 samples

Hierarchical clustering of the 122 Stanford/Norway samples based on RNA expression levels of 476 genes common to our 27.243 genes/ESTs and the intrinsic 500-gene set used by Sorlie *et al.* (8). Under the dendrogram of samples, the horizontal colored boxes delimit the five tumor subgroups: luminal A (dark blue box), luminal B (light blue box), ERBB2-overexpressing (pink box), basal (red box) and normal breast-like (green box). Branches of the core samples used for computing each of the five centroids are similarly color-coded in the dendrogram. Black branches represent samples with low correlation to any subgroup. Colored bars to the right indicate the locations of the ERBB2 (pink bar), basal (red bar) and luminal (dark blue bar) gene clusters.

Supplementary Figure 2: Supervised classification of 66 breast cancer samples based on the MBC/DBC molecular signature

A/ Similar to Figure 3A, but applied to the 527 genes identified as discriminator between the 22 MBCs and the 44 DBCs. The position of *MSN* and *GATA3* are indicated. *B/* Validation of the GES by RNA profiling of another series of samples on another microarray platform. The 62 genes common to the MBC/DBC gene expression signature (527 genes) and the IPC/Ipsogen dataset (3.891genes) are submitted to hierarchical clustering with respect to samples from present study (*left*, n=66) and from IPC/Ipsogen study (*right*, n=73). In each analysis, clustering evidenced two major groups of samples. The pathological type of samples is

represented as in Figure 1. Genes correlated with each pathological type (MBC, black and DBC, white) are similarly color-coded in the two datasets. *C/* Analysis of protein expression using IHC on tissue microarray sections. Examples of IHC staining for GATA3 and moesin (epithelial cells: single arrow) are shown in DBC and MBC samples (magnification is x100). *Top,* in the ER-positive DBC sample, staining is strong for GATA3 and low for moesin. Conversely, in the MBC sample (*Bottom*), staining is negative for GATA3 and positive for moesin. Note the diffuse lymphoplasmocyte stromal infiltrate (dashed double arrow) in the MBC sample.