Oncogene (2005) 24, 4660–4671

www.nature.com/onc

Identification of molecular apocrine breast tumours by microarray analysis

Pierre Farmer[1,2], Herve Bonnefoi[3,4,5], Veronique Becette[6], Michele Tubiana-Hulin[6], Pierre Fumoleau[7], Denis Larsimont[8], Gaetan MacGrogan[9], Jonas Bergh[10], David Cameron[11], Darlene Goldstein[1,2], Stephan Duss[2], Anne-Laure Nicoulaz[2], Cathrin Brisken[2], Maryse Fiche[12], Mauro Delorenzi[1,2] and Richard Iggo*[,2]

[1]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; [2]National Centre of Competence in Research (NCCR) Molecular Oncology, Swiss Institute for Experimental Cancer Research (ISREC), Epalinges, Switzerland; [3]Hôpitaux Universitaires de Genève, Geneva, Switzerland; [4]for the Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland; [5]European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium; [6]Centre René Huguenin, St-Cloud, France; [7]Centre René Gauducheau, Nantes, France; [8]Institut Jules Bordet, Brussels, Belgium; [9]Institut Bergonié, Bordeaux, France; [10]for the Swedish Breast Cancer Group (SweBCG), Karolinska Institute, Stockholm, Sweden; [11]for the Anglo-Celtic Cooperative Oncology Group (ACCOG), Edinburgh University, Edinburgh, UK; [12]Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

Previous microarray studies on breast cancer identified multiple tumour classes, of which the most prominent, named luminal and basal, differ in expression of the oestrogen receptor α gene (ER). We report here the identification of a group of breast tumours with increased androgen signalling and a 'molecular apocrine' gene expression profile. Tumour samples from 49 patients with large operable or locally advanced breast cancers were tested on Affymetrix U133A gene expression microarrays. Principal components analysis and hierarchical clustering split the tumours into three groups: basal, luminal and a group we call molecular apocrine. All of the molecular apocrine tumours have strong apocrine features on histological examination ($P = 0.0002$). The molecular apocrine group is androgen receptor (AR) positive and contains all of the ER-negative tumours outside the basal group. Kolmogorov–Smirnov testing indicates that oestrogen signalling is most active in the luminal group, and androgen signalling is most active in the molecular apocrine group. ERBB2 amplification is commoner in the molecular apocrine than the other groups. Genes that best split the three groups were identified by Wilcoxon test. Correlation of the average expression profile of these genes in our data with the expression profile of individual tumours in four published breast cancer studies suggest that molecular apocrine tumours represent 8–14% of tumours in these studies. Our data show that it is possible with microarray data to divide mammary tumour cells into three groups based on steroid receptor activity: luminal (ER+ AR+), basal (ER− AR−) and molecular apocrine (ER− AR+).

Oncogene (2005) **24,** 4660–4671. doi:10.1038/sj.onc.1208561
Published online 9 May 2005

Keywords: breast cancer; microarrays; apocrine carcinoma

*Correspondence: R Iggo; E-mail: richard.iggo@isrec.ch
Received 29 September 2004; revised 7 January 2005; accepted 27 January 2005; published online 9 May 2005

Introduction

Despite the existence of 18 different histopathological types of breast cancer (Ellis, 2003), the major distinctions important in clinical practice are based on oestrogen receptor (ER) and ERBB2 status rather than histopathological tumour type. Unsupervised hierarchical clustering of microarray data readily identifies groups of tumours differing in ER and ERBB2 expression. The gene expression patterns corresponding broadly to ER-positive and -negative tumours have been named luminal and basal, respectively (Perou *et al*., 2000). Typically, the basal/luminal split accounts for a substantial part of the overall variation in gene expression in breast cancer microarray studies. Some of the genes showing differential expression are proven ER targets, but many others appear to be differentially expressed because of a difference in cell type. The most widely used microarray-based classification further divides breast tumours into five classes: luminal A, luminal B, basal, normal and ERBB2 (Sorlie *et al*., 2003). All of these classes except ERBB2 appear to be related to cell differentiation or cell type. Since oncogenic mutations do not generally determine cell type, it seems reasonable to attempt to derive classifications that use cell type and mutation information at separate levels. A classification based on cell type should take into account not only the cell types corresponding to the normal differentiation program in the breast but also common pathological cell lineages. Apocrine differentiation is one such change. Apocrine glands are androgen-dependent scent glands found in the axilla and perineum. Widespread apocrine metaplasia is seen in tension cysts in fibrocystic breast disease, where it arises in response to a currently ill-defined stress (Viacava *et al*., 1997). Apocrine differentiation is also seen in a group of benign and malignant breast diseases, including apocrine hyperplasia and apocrine carcinoma (reviewed by Selim and Wells, 1999; Schmitt and Reis-Filho, 2002). Whether the apocrine cells seen in fibrocystic disease and apocrine carcinoma derive from

a pathologically transformed progenitor or from expansion of a population of apocrine cells in the normal adult breast has been debated by several authors (for references, see Jones *et al*., 2001).

We report here the identification of a group of breast tumours with increased androgen signalling and some apocrine features in a microarray study on large operable or locally advanced breast cancer. These tumours are ER negative and share some features with the ERBB2 class in the Stanford classification. We use the term 'molecular apocrine' to describe these tumours, and suggest that retention of androgenic signalling may explain a substantial part of the RNA phenotype of the ER-negative tumours outside the basal class.

Results

RNA was extracted from biopsies of T2–T4 breast tumours and hybridized to Affymetrix U133A chips. Most patients had two trucut biopsies taken; the biopsy with the higher proportion of tumour cells was used for the array study. Both biopsies were analysed from two tumours to test the reproducibility of the technique. Repeat amplifications and duplicate biopsies clustered together, showing that biological variation is greater than technical variation in this data set (Supplementary Figure 1). The ER status determined by immunohistochemistry (IHC) showed an excellent correlation with the *ESR1* expression level on the arrays, which again suggests that the data are of high quality (Supplementary Figure 2).

Unsupervised clustering identifies three major groups of tumours

Principal components analysis (PCA) of all genes showed that the global expression patterns of the tumours fall naturally into three major groups (Figure 1a). Silhouette plots (Rousseeuw, 1987) confirmed that the global expression patterns are best described by division into two or three well-separated relatively homogeneous groups (Supplementary Figure 3). To understand the basis for this grouping, the expression pattern was examined using hierarchical clustering of the most variable genes (Figure 1b). To avoid redundancy, a single probeset was used per gene. Comparison with the PCA plot showed that the first component splits the tumours into basal (green in Figure 1a) and luminal (blue in Figure 1a) groups. As

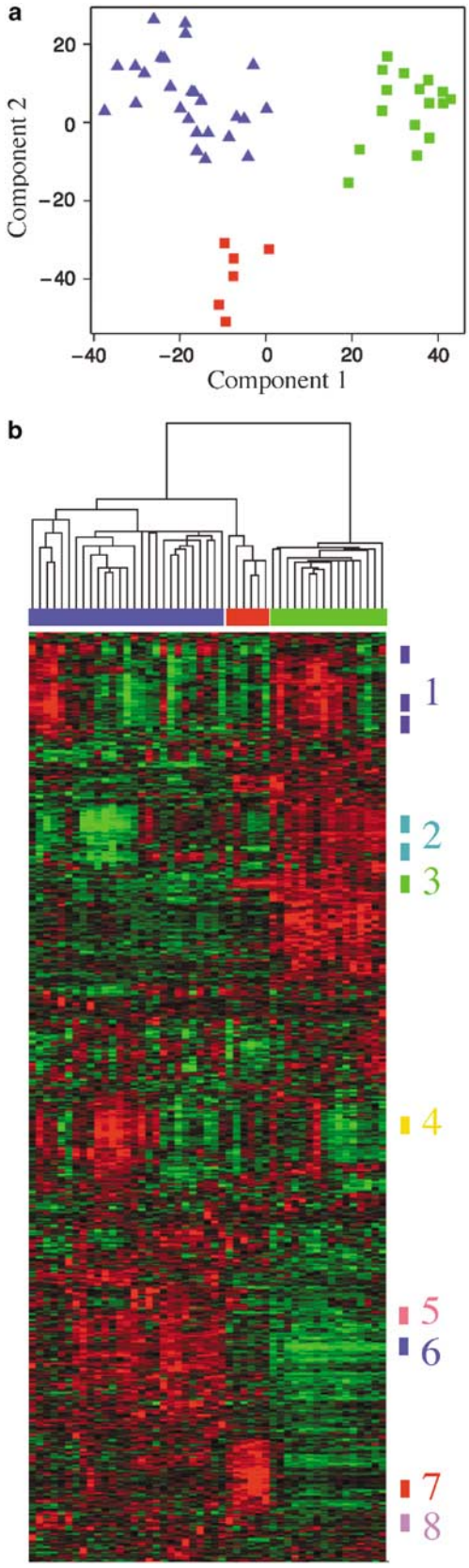

Figure 1 Unsupervised analysis. (**a**) Principal components analysis using all probesets. The first two components are plotted. Points are labelled according to ER status determined by immunohistochemistry: triangle, ER positive; square, ER negative. The three major groups in the data are labelled in blue, green and red. (**b**) Hierarchical clustering using 3198 nonredundant probesets. The horizontal bar under the dendrogram uses the same colour scheme as in (**a**). Small parts of the image have been expanded in (**c**) and Supplementary Figure 5 to show interesting gene clusters: 1, interferon, T cell and B cell genes; 2, proliferation and 8q amplicon genes; 3, apocrine/basal and hypoxia genes; 4, stromal genes; 5, 17q21–23 amplicon genes; 6, luminal genes; 7, apocrine/luminal genes; 8, ERBB2 amplicon genes. The Treeview files are available as Supplementary data 1

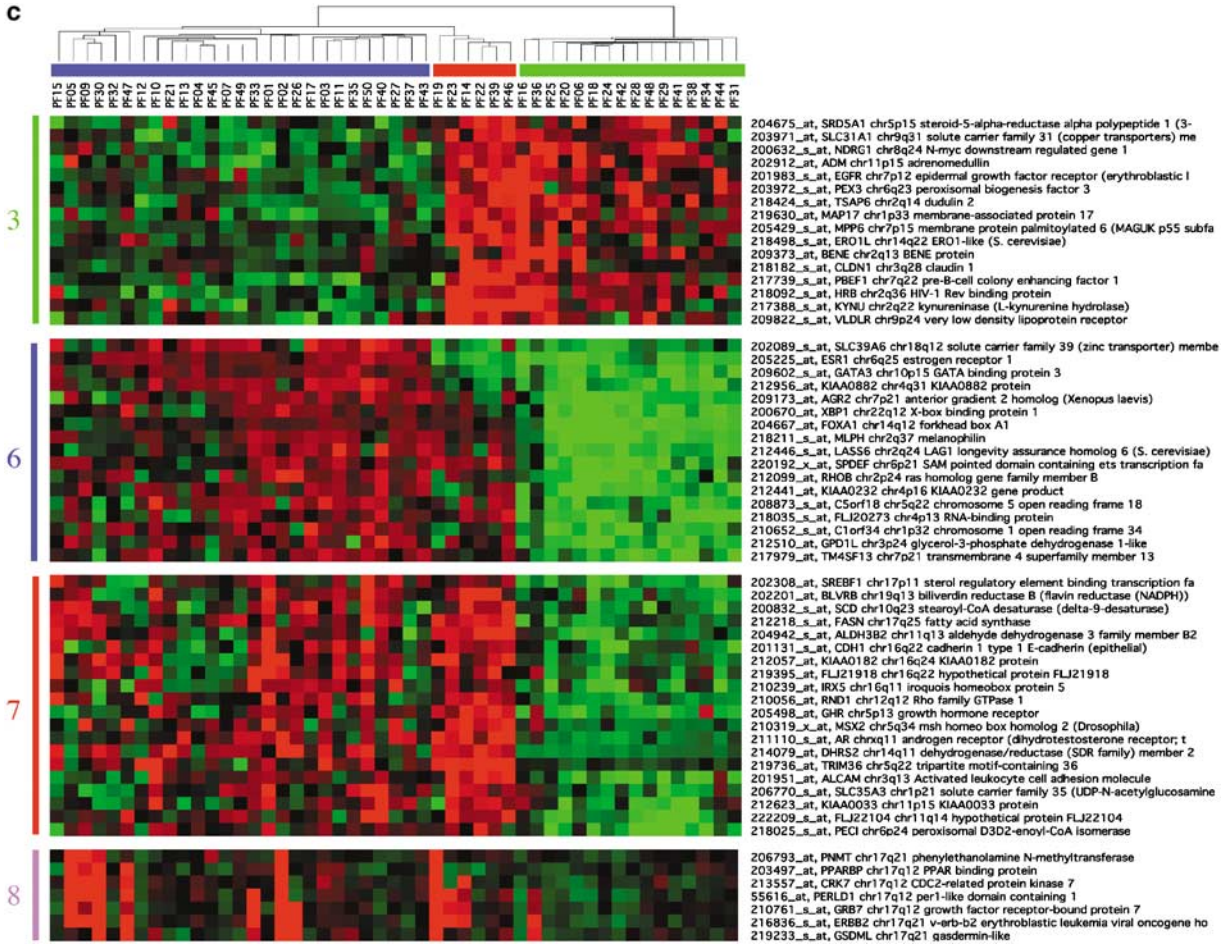

Figure 1 *Continued*

noted previously (Perou *et al*., 2000), the basal groups express higher levels of basal keratins like 5 and17 and lower levels of keratins 8, 18 and 19 (Supplementary Figure 4).

Consistent with previous reports, large gene clusters corresponding to tumour infiltration with stromal cells, B cells and T cells, as well as large proliferation and interferon clusters, are easily identifiable in the clustered data (Supplementary Figure 5). Regions of presumptive copy number change are also visible, for example amplification of 8q and 17q (Supplementary Figure 5). These gene clusters dictate the subdivisions within the basal and luminal groups in the tumour dendrogram, but not the three-way split noted in the PCA plot (Figure 1a). The third PCA group corresponds instead to the third major group in the tumour dendrogram and includes 6/49 tumours (12%). It is ER negative (ESR1, Figure 1c panel 6) and shares some gene clusters with luminal tumours (Figure 1c panels 6 and 7) and others with the basal tumours (Figure 1c panel 3). The pattern of keratin expression resembles that in luminal tumours, with the addition of keratin 7 (Supplementary Figure 4).

To determine whether the third PCA group is caused by contamination of tumour samples with normal tissue, three reduction mammoplasty samples were tested and the clustering was repeated using the Stanford 'intrinsic' gene set (Perou *et al*., 2000). The normal samples do not cluster with the tumours in the third group (Figure 2a). To determine whether the third group corresponds to one of the previously defined Stanford classes, the expression profile of each tumour was compared to reference profiles based on the mean expression level of the intrinsic genes in each Stanford class (Sorlie *et al*., 2003). The normal samples and putative basal tumours show high correlations with the corresponding Stanford profiles (Figure 2b). The remaining ER-negative Stanford classes are ERBB2 and, to some extent, luminal B. By exclusion, tumours in our third group are likely to fall into one of these categories. Neither profile explains a substantial part of the expression pattern in our tumours (the coefficient of determination, r^2, is plotted in Figure 2b, since this indicates the proportion of variation explained under a linear model). The biological basis for the luminal B class is unclear, which makes it difficult to infer by other means whether our third group is related to Stanford luminal B tumours. There are 11 tumours with ERBB2 amplification in our set, as defined by the ERBB2 mRNA level (Figure 1c panel 8

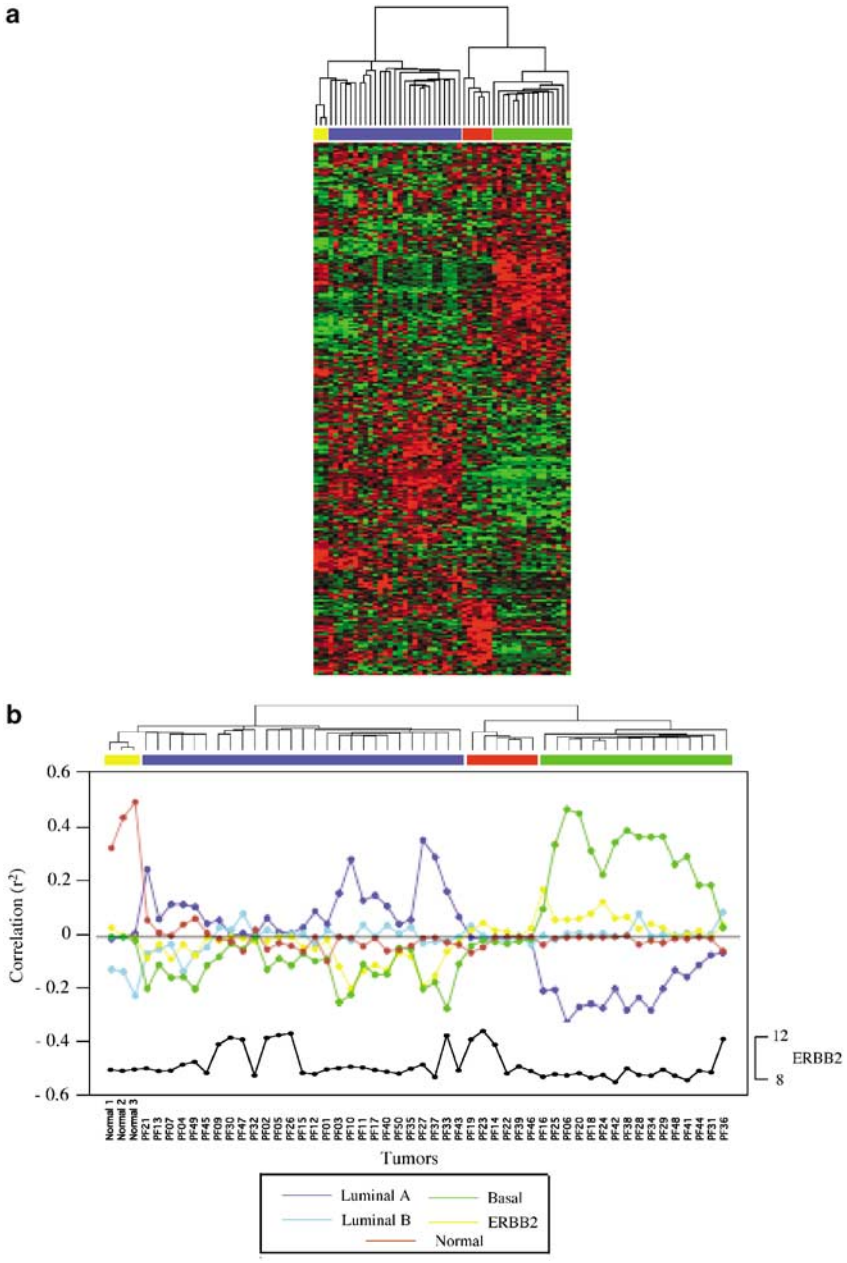

Figure 2 Analysis using the intrinsic gene set. (**a**) Hierarchical clustering of 269 nonredundant probesets mapped to the Stanford intrinsic gene set. The horizontal bar under the dendrogram uses the same colour scheme as in Figure 1, with normal tissue shown in brown. The Treeview files are available as Supplementary Data 2. (**b**) Correlation with the mean expression profile of intrinsic genes in the Stanford tumour classes. The coefficient of determination is shown (r^2), with preservation of the sign. The black curve shows the measured ERBB2 values in arbitrary $\log_2$ units

and Figure 2b) and increased expression of neighbouring genes on 17q (Supplementary Figure 6). Three tumours in our third group are ERBB2 positive (3/6). All but one of the remaining ERBB2-positive tumours are in the ER-positive group (7/27). ERBB2 amplification alone is thus not a satisfactory explanation for the formation of the third group. Nevertheless, for reasons given below, we suggest that the Stanford ERBB2 class most nearly describes the phenotype of the tumours in our third group.

The third group contains tumours with increased androgen signalling

The large number of differences between the expression pattern of the third group and the luminal or basal tumours suggests that the basis for the three-way split may be a difference in cell differentiation or cell type. The third group contains all the ER-negative tumours outside the basal group, but, in contrast with the basal tumours, retains expression of the androgen

receptor (AR, Figure 1c panel 7, and Figure 3a). Genes that are highly expressed in the third group were identified by Wilcoxon test. The gene ontology (GO) terms for these genes were examined to gain clues to the function of the cells. The best-represented GO terms at a GO depth of 4 were for 'metabolism', which includes many genes for fatty acid and lipid synthesis (Table 1). A trivial explanation might be contamination of the samples with adipocytes, but adipocyte markers like FABP4, CD36, GPX3 and LPL (Perou *et al*., 1999) are not highly expressed in the third group, the samples do not cluster with normal tissue, and histological examination ruled out extensive adipocyte contamination. Androgen signalling leads to increased metabolic gene expression and provides an alternative explanation for the GO results, particularly given the retention of AR expression in this group.

Two approaches were used to test whether the androgen receptor is actively signalling in the third group. In the first approach, metagenes were constructed based on the mean expression level of AR and ER target genes. The target gene lists were based on published array studies listing genes induced by androgen in LNCaP prostate cancer cells (Nelson *et al*., 2002) and oestrogen in MCF7 cells (Frasor *et al*., 2003). To avoid circularity, the AR and ER metagenes do not include the AR and ESR1 probesets. Despite the fact that the target lists were established on different platforms, the metagenes separate the tumours

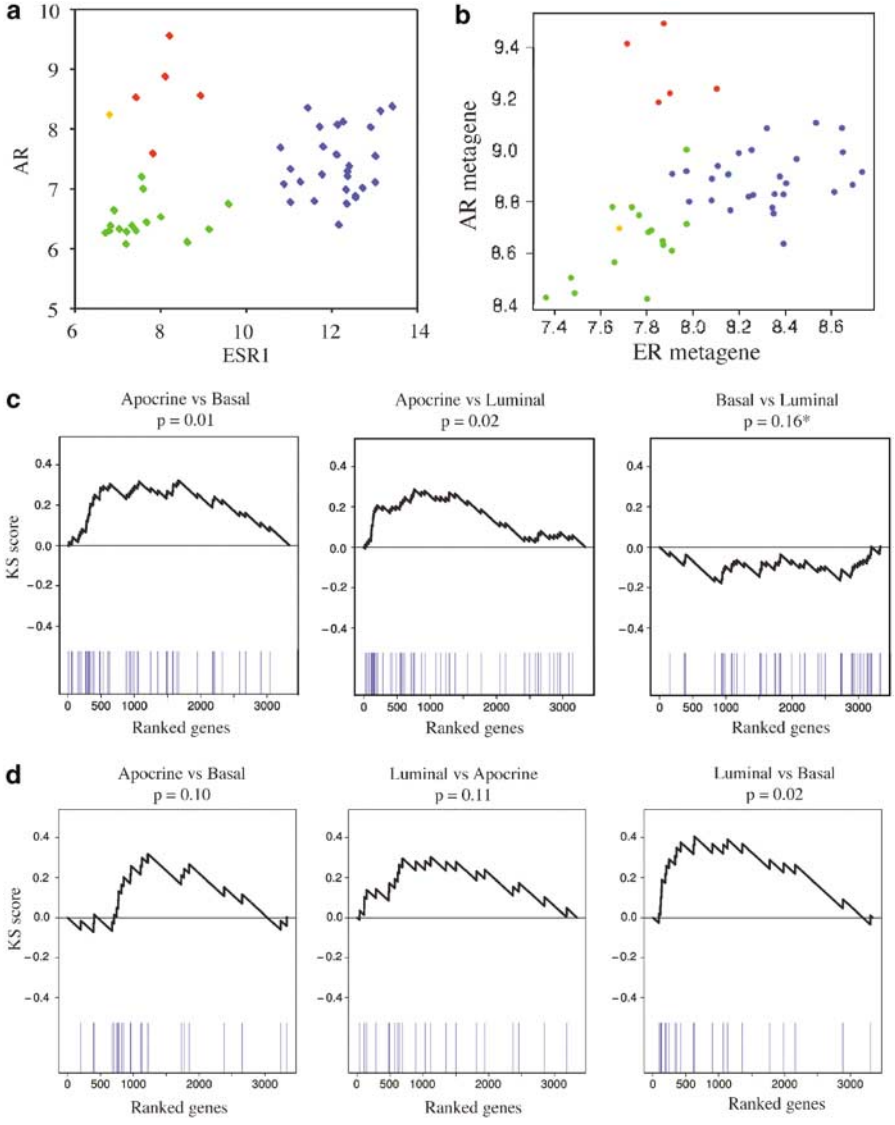

Figure 3 AR and ER expression split the tumours into three groups. (**a**) Expression levels for AR and ER. (**b**) Expression levels for AR and ER metagenes, representing the mean level of expression of a pool of androgen- and oestrogen-inducible genes. The tumours are coloured in (**a** and **b**) according to the same scheme as in Figure 1, with PF19 in brown. (**c**, **d**) KS tests for androgen- and oestrogen-dependent gene expression (**c**, androgen; **d**, oestrogen). The *P*-values were calculated by randomization of the assignment of tumours to groups. The asterisk in (**c**) indicates that this *P*-value (0.16) was obtained by ranking the genes from right to left

Table 1 GO terms over-represented in the apocrine group

GO depth	*Actual hits*	*Possible hits*	P-*value*	*Raw* P-*value*	*GO description*	*GO ID*
0	168	14475			All	GO:0003673
1	156	12779			Biological process	GO:0008150
4	24	436	3E-07	6E-10	Organic acid metabolism	GO:0006082
4	10	52	3E-07	6E-10	Sulphur metabolism	GO:0006790
4	12	118	9E-06	2E-08	Aromatic compound metabolism	GO:0006725
4	22	614	2E-03	5E-06	Lipid metabolism	GO:0006629
4	13	252	6E-03	1E-05	Alcohol metabolism	GO:0006066
4	13	280	2E-02	4E-05	Amino acid and derivative metabolism	GO:0006519
4	14	336	3E-02	6E-05	Amine metabolism	GO:0009308

The 222 genes (1% of probe sets) that best separate the third group from the other tumours were mapped to GO terms. Out of the 222 genes tested, 156 were present in the biological process division of the GO classification. Terms giving a $P<0.05$ at a GO depth of 4 are shown

into the same three groups as were seen in the global expression pattern, indicating that different levels of oestrogen and androgen signalling correlate with and constitute a possible explanation for the splitting of the tumours into three major groups (Figure 3b). PF19 has lower AR activity than the other tumours in the third group (coloured brown in Figure 3b), consistent with its behaviour as an outlier in the clustering analysis (Supplementary Figure 3). To test whether the differences in target gene expression are significant, a 'Kolmogorov–Smirnov' (KS) test was performed (Lamb *et al*., 2003). This test can be used to determine whether a particular group of genes, for example androgen-inducible genes, is more highly expressed in a particular group of tumours. To perform a KS test, genes are ranked according to a criterion, for example similarity to AR expression. The position of target genes (eg, androgen-inducible genes) within this ranked list is then determined. A running score that starts and ends at zero is kept, increased at the positions of the target genes and decreased otherwise. The test statistic is the maximum of the running score and is high when the genes in the target list are enriched early in the ranking (Lamb *et al*., 2003). To determine the significance of the score, the test is repeated after randomization of the ranking or the target list.

To confirm that this approach can identify tumours with active androgen or oestrogen signalling, KS tests were performed with the androgen and oestrogen target genes using all samples and ranking genes according to similarity to *AR* or *ESR1* expression, respectively (Supplementary Figure 7). To determine whether the KS scores were significant, empirical *P*-values were obtained by repeating the test with randomly selected target genes. In both cases, the KS test found a significant association between receptor expression and target gene expression ($P<0.001$). Genes that define membership of the three sample groups were then ranked by Wilcoxon test statistic as a criterion for higher expression in one group, and these ranked lists were used to test for significant differences in androgen and oestrogen signalling between pairs of groups (Figure 3c and d; the AR+/ER− group is labelled apocrine for reasons given below). To determine whether the KS scores were significant, empirical permutation *P*-values were estimated by repeating the test with randomly permuted group labels. The results showed that AR activity is significantly higher in the AR+/ER− compared to the luminal and basal groups, while ER activity is significantly higher in the luminal than in the basal group and intermediate in the AR+/ER− tumours. This provides strong support for the interpretation that the AR+/ER− group identified by PCA and hierarchical clustering contains tumours with active AR but inactive or weak ER signalling.

The third group contains tumours with apocrine features

Exposure of mammary epithelium to supraphysiological doses of androgen, for example in female to male transsexuals, leads to a histological change referred to as apocrine metaplasia (reviewed by Liao and Dickson, 2002). The malignant counterpart of apocrine metaplasia is apocrine carcinoma. To test whether the tumours with increased androgen signalling had apocrine features, such as abundant eosinophilic cytoplasm and prominent nucleoli, tissue sections were examined histologically by a pathologist without prior knowledge of the microarray assignments. Apocrine grade was scored on a three-point scale as described by Miller *et al*. (1985), although it must be pointed out that formal diagnosis of apocrine carcinoma is generally very difficult using frozen sections because of insufficient preservation of cell morphology. None of the biopsies contained extensive regions of benign apocrine metaplasia that could account for the gene expression data. Except for the tumour flagged as an outlier in the silhouette plot (PF19, Supplementary Figure 3), the tumours in the third group all had marked apocrine features (Figure 4a and b). The association between apocrine histology and clustering in the AR+ ER− group was highly significant (Table 2a and Supplementary Table sheet 1). Since tumours in the third group have an expression profile compatible with increased androgen signalling and some morphological hallmarks of apocrine tumours, but do not meet the strict histopathological criteria for diagnosis as classical apocrine carcinomas, we suggest calling them 'molecular apocrine' tumours.

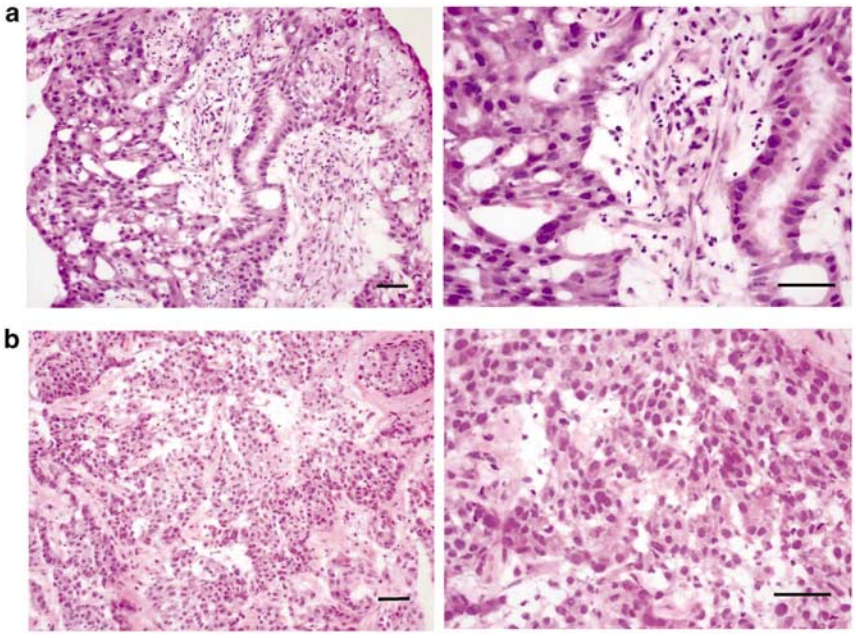

Figure 4 Histological appearances of tumours PF39 (**a**) and PF23 (**b**). Haematoxylin and eosin stain, scale bar 100 μm

Table 2 Apocrine scores

	Luminal	*Basal*	*Molecular apocrine*
(a) Farmer *et al*.[a]			
Apocrine score 1	14	8	0
Apocrine score 2	10	5	1[b]
Apocrine score 3	2[c]	1[c]	5
(b) Perou *et al*.[d]			
Apocrine score 1	5	4	0
Apocrine score 2	4	1	2
Apocrine score 3	0	0	2

Frozen sections were scored for apocrine features by a pathologist without prior knowledge of the microarray assignments. Apocrine grade 3 has the most marked apocrine features. (a) Our tumours. (b) Perou *et al*.'s (2000) tumours scored using images on the Stanford web site (http://genome-www.stanford.edu/breast_cancer/molecularportraits/). The individual tumor assignments are given in the supplementary table (sheets 1 and 4) [a]*P*-value = 0.0022 by Fisher's exact test, two sided. [b]Sample PF19. [c]Samples PF09, 16 and 47; samples PF02, 42 and 48 were excluded because of inadequate tissue preservation. [d]*P*-value = 0.044 by Fisher's exact test, two sided

The length of the AR polyglutamine repeat is normal in the molecular apocrine group

The underlying mechanism responsible for apocrine differentiation is not known, but increased sensitivity to androgen signalling is one possibility. The *AR* gene contains a polymorphic CAG_n polyglutamine repeat whose expansion impairs receptor function (La Spada *et al*., 1991; Chamberlain *et al*., 1994). Higher grade breast tumours express shorter repeats (Yu *et al*., 2000). The KS data showing increased androgen signalling in the apocrine group would be compatible with selection for expression of shorter repeats in these tumours. To test for this, repeat length was measured in the RNA used for the array study. Genomic DNA tested from 12 tumours confirmed that the repeat is highly polymorphic, with two different alleles present in most of the samples (data not shown). Consistent with the location of the *AR* gene on the X chromosome, only a single allele was detected at the RNA level. The mean number of CAG repeats was 18 in the apocrine group (range 17–19), compared with 18 and 20 in the basal and luminal groups (Supplementary Table sheet 1). All of the tumours in the apocrine group thus express *AR* alleles potentially encoding fully active receptors.

Identification of molecular apocrine tumours in other microarray data sets

To determine whether molecular apocrine tumours are present but under-reported in other breast cancer microarray studies, we identified genes that discriminate between the three groups in our own data and looked at the pattern of expression of these genes in four published data sets (West *et al*., 2001; van't Veer *et al*., 2002; Huang *et al*., 2003; Sorlie *et al*., 2003). The luminal/apocrine/basal or LAB gene set (Supplementary Table sheet 2) contains the most discriminant genes by Wilcoxon test for each pairwise comparison in our data (luminal vs basal, basal vs apocrine, apocrine vs luminal). The LAB genes were mapped to the other data sets and the top 90 genes for each pairwise comparison in each data set were retained. The group-specific mean expression levels of the LAB genes in the luminal, molecular apocrine and basal groups in our tumours were used to generate reference profiles for each tumour group. Correlation with these reference profiles was used to assign tumours to the luminal,

Table 3 Classification of tumours in this study and the Huang *et al*. (2003), West *et al*. (2001), Sorlie *et al*. (2003) and van't Veer *et al*. (2002) studies using the LAB gene set

Data set	*Basal*	*Luminal*	*Molecular apocrine*	*Undefined*	*Total*
Our data	16 (33)	27 (55)	6 (12)	0 (0)	49 (100)
Perou *et al*. (2000)	7 (19)	14 (38)	5 (14)	11 (30)	37 (100)
West *et al*.(2001)	19 (39)	21 (43)	4 (8)	5 (10)	49 (100)
Huang *et al*. (2003)	15 (17)	42 (47)	9 (10)	23 (26)	89 (100)
Van't veer Veer *et al*. (2002)	20 (21)	50 (52)	10 (10)	16 (17)	96 (100)

To assign class membership, the decision rule was that the Spearman correlation coefficient should exceed the threshold at which a class would be assigned in $<1\%$ of tests with scrambled data. If more than one class exceeded this threshold, a class was assigned only if the correlation coefficient exceeded the next best value by the same threshold amount

molecular apocrine and basal classes (Table 3 and Supplementary Table sheet 3). The number of tumours with the molecular apocrine profile in all five data sets (8–14%) is consistent with historical estimates of the number of breast tumours with marked apocrine histological features (2–15%)(Miller *et al*., 1985). To visualize the result, our tumours were used to train two partial least squares (PLS) models, one that separates the basal from luminal tumours, and another that separates the molecular apocrine from all other tumours. Figure 5a shows a projection of the tumours in each data set onto the first components of these PLS models. This shows that the LAB assignments successfully divide the tumours in each study into separate clusters containing the putative luminal (blue), molecular apocrine (red) and basal (green) tumours. The tumours from the Sorlie and van't Veer studies are labelled in Figure 5a according to the Stanford classes assigned by Sorlie *et al*. (2003): square for luminal, triangle for basal, diamond for ERBB2, circle for normal, cross for unclassifiable (the individual assignments are given in Supplementary Table sheet 3).

It is not possible independently to verify these predictions in most cases because histological apocrine scores are not available for the tumours in the other studies, but images of some of the Sorlie tumours are given on the Stanford University web site. The histological apocrine grade of these tumours was scored by a pathologist without prior knowledge of the microarray assignments. There was a significant association between apocrine histology and molecular apocrine class, although the numbers are small and the resolution of the images limits the quality of the analysis (Table 2b and Supplementary Table sheet 4). The clinical outcome of the patients with molecular apocrine tumours in our study is not yet known, but survival data are available for the Sorlie and van't Veer patients. In both studies, the molecular apocrine profile is associated with poor long-term survival (Figure 5b).

Discussion

All published microarray studies on breast cancer have reported striking differences in the gene expression profile of ER+ and ER− tumours. This reflects the importance of steroid receptor signalling in mammary epithelial cells. Based on our data, we suggest that there is a group of tumours containing ER− AR+ cells, in which androgen signalling replaces oestrogen signalling as a major determinant of the steroid-related expression profile of the cells. This is consistent with previous immunohistochemical reports that divided breast tumours into ER+ AR+, ER− AR+ and ER− AR− cells (Moinfar *et al*., 2003). The question that arises is how to integrate this information with previous microarray schemes for classification of breast cancer. Cell type influences the expression of so many genes that it dictates the highest levels of split in most tumour dendrograms. In breast cancer, cell type and steroid receptor signalling are intimately linked, which explains why the luminal/basal split closely follows ER status. The next lower levels of the dendrogram in our data are dictated by clusters of lymphocyte, stromal and proliferation genes. Although there is broad agreement, it is difficult precisely to map the divisions in our luminal group to those reported by others, because there is no agreed definition of these groups. In our view, the best way to make sense of subdivisions within a particular cell type will be to identify signature oncogenic defects. For example, MYC amplification at 8q24 is associated with poor survival (Jain *et al*., 2001). The low proliferation group in the luminal cluster in our data overlaps with a large cluster of coordinately regulated genes on chromosome 8q (Supplementary Figure 5 panel 2). MYC is a candidate gene to drive the amplification, but the region amplified encompasses much of the long arm of 8q in many studies, and MYC does not cluster within the main 8q amplicon in our data, suggesting that MYC is not the only relevant oncogene on 8q (Isola *et al*., 1995; Pollack *et al*., 2002; Molist *et al*., 2004).

Among ER-negative tumours, there is one group with a relatively homogeneous expression profile, named basal by Perou *et al*. (2000) because of higher expression of basal keratins. At issue is the phenotype of the ER-negative tumours outside the basal group. Several other microarray studies on breast cancer identified an ERBB2 class of tumours as one of the primary subdivisions in the tumour dendrogram (Perou *et al*., 2000; Sorlie *et al*., 2003; Sotiriou *et al*., 2003). The ERBB2 class is a subdivision of the ER-negative group, but many ERBB2-positive tumours are found outside the ERBB2 class in all studies, and some tumours within the ERBB2 class lack ERBB2 overexpression or amplification. Thus, ERBB2 expression does not satisfactorily

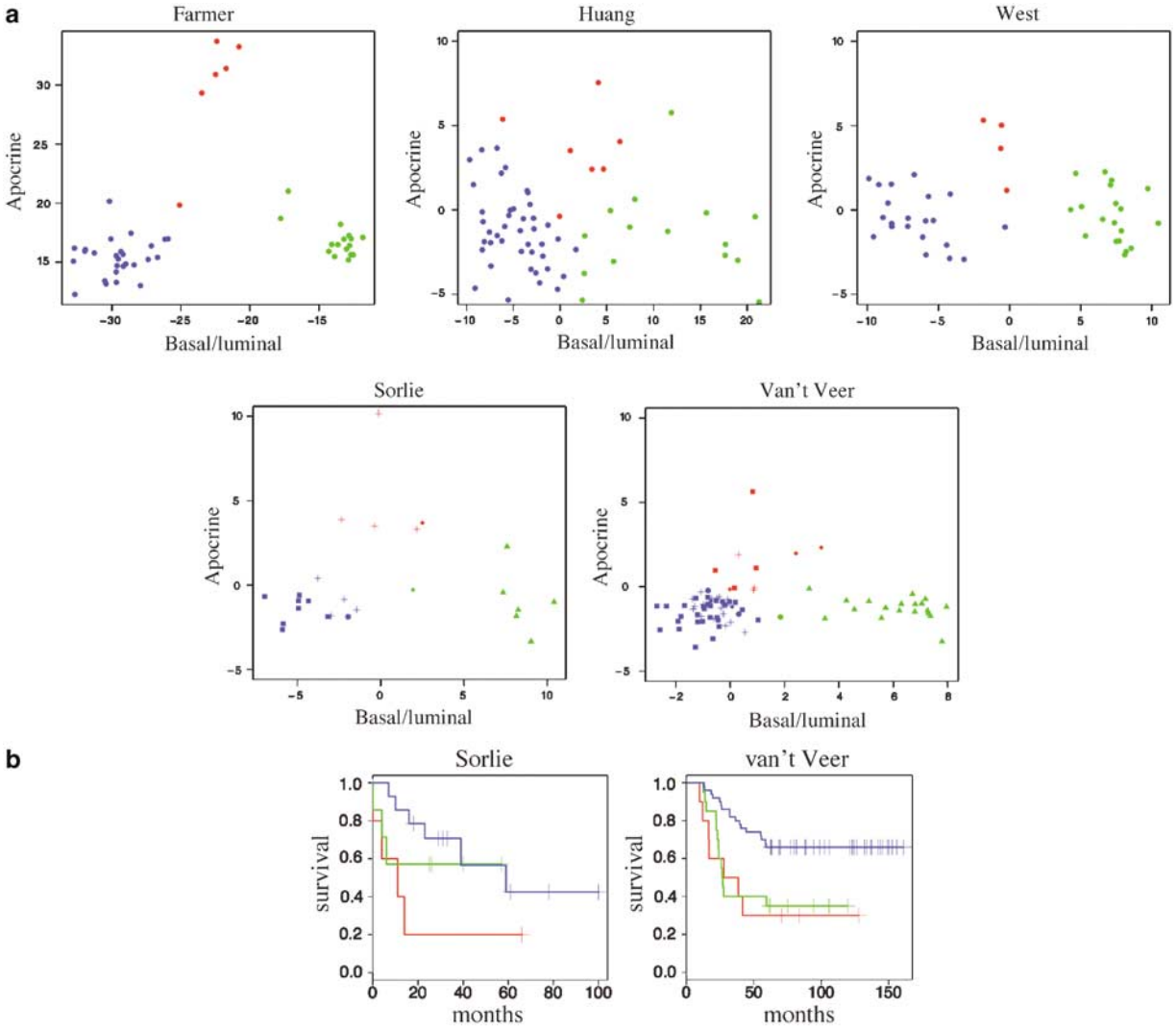

Figure 5 Identification of molecular apocrine tumours in published data sets. (**a**) Tumour classification using PLS components. The LAB gene set was used to make PLS models that maximize covariance with the basal vs luminal and the apocrine vs other groups in our data. The first PLS components were then used to plot tumours in the West *et al*. (2001); Huang *et al*. (2003); Sorlie *et al*. (2003) and van't Veer *et al*. (2002) data sets. The original Sorlie *et al*. (2003) classification of the van't Veer and Sorlie data is given by the shape of the symbols: square for luminal, triangle for basal, diamond for ERBB2, circle for normal, cross for unknown. (**b**) Kaplan–Meier survival curves for the van't Veer and Sorlie patients according to the LAB classification in Table 3. Red, molecular apocrine; green, basal; blue, luminal; tumours labelled undefined in Table 3 are omitted

define membership of the ERBB2 class. The ER-negative non-basal group probably contains several different types of tumour, but in our data we can see mainly one phenotype, which we have called molecular apocrine. Taking a very broad view, we suggest that the Stanford ERBB2 class is the same as the molecular apocrine class in our data. The Stanford classification has difficulty defining these tumours, as shown by the fact that of the five tumours we identified as molecular apocrine, three, five and one were placed in the ERBB2 class in successive analyses of the same data by Perou *et al*. (2000), Sorlie *et al*. (2001) and Sorlie *et al*. (2003), respectively.

In our hands, supervised analysis comparing the 11 ERBB2-positive tumours with the rest suggests that the expression profile of ERBB2-positive tumours is mainly defined by genes in the ERBB2 amplicon itself. Using the first PLS component from pairwise comparisons (apocrine *vs* others, ERBB2 *vs* others), the apocrine component explains 7.7% and the ERBB2 component 5.9% of the total variance in our data, despite the smaller number of apocrine tumours. A leave-one-out cross-validation test of pairwise discrimination with nearest centroid classification on the first PLS component shows that the luminal, apocrine and basal groups are perfectly predicted, while the ERBB2-positive group is not, suggesting that the ERBB2-positive tumours do not have a single homogeneous phenotype. The molecular apocrine group is certainly enriched in ERBB2-positive tumours, which suggests that there may be a link between ERBB2 signalling and the molecular apocrine phenotype. The possibility that the two are

related is further suggested by the observation that ERBB2 is overexpressed but not amplified in apocrine adenosis (Selim *et al*., 2000) and the finding that high grade *in situ* and invasive tumours are frequently AR+/ER−/ERBB2+ (Moinfar *et al*., 2003). Bertucci *et al*. (2004) suggested that coamplification of PPARBP with ERBB2 may result in modifications of fatty acid metabolism in ERBB2-positive tumours. This seems unlikely to explain the molecular apocrine phenotype in our data, because PPARBP is highly expressed in only one of the three ERBB2-positive tumours in the molecular apocrine group (PF19, Supplementary Figure 6). Another possible explanation is that ERBB2 reinforces androgenic signalling by stabilizing AR at the protein level (Mellinghoff *et al*., 2004). This might lead to an apocrine phenotype in ER-negative tumours that retain AR expression, but not lead to apocrine changes in cells expressing both AR and ER because of dominant oestrogenic signalling in these cells.

However, ERBB2 amplification could only explain increased androgenic signalling in half of the molecular apocrine tumours in our data. Another mechanism for activation of AR is required in the remainder. Comparative genomic hybridization has identified clonal abnormalities in apocrine hyperplasia and carcinoma, with the number of changes increasing as the cells progress towards malignancy, but no amplified or deleted locus clearly linked to androgen signalling has been identified (Jones *et al*., 2001). An imbalance in AR and ER signalling could be caused by mutation of the androgen receptor itself. AR mutations have been extensively investigated in prostate cancer, principally to explain resistance to antiandrogen therapy (Lopez-Otin and Diamandis, 1998). The receptor contains a highly polymorphic CAG repeat which is expanded in spinal and bulbar muscular atrophy (SBMA, also known as Kennedy's disease), an androgen-insensitivity syndrome (La Spada *et al*., 1991). Conversely, deletion of the repeat increases receptor activity, although the effects are modest (Chamberlain *et al*., 1994). The number of repeats in the tumours from the molecular apocrine group was 17–19, which is at the shorter end of the normal range. Increased AR activity in this group can thus not be explained by a pathological contraction of the repeat. It remains possible that full sequencing of the gene would reveal the presence of activating mutations. An alternative explanation for increased AR target gene expression is that androgen synthetic pathways are more active in these cells. For example, SRD5A1, the 5α-reductase which converts testosterone to dihydrotestosterone, is overexpressed in the apocrine and basal groups (Figure 1c panel 3). This is consistent with previous data showing increased 5α-reductase biochemical activity in apocrine tumours (Miller *et al*., 1985). The GO analysis revealed the presence of many genes involved in steroid metabolism in the apocrine group. Some, like HSD17B2 and UGT2B28, are involved in the catabolism of oestrogens and androgens (Wu *et al*., 1993; Levesque *et al*., 2001). Increased expression of catabolic as well as synthetic enzymes is a feature of steroid responsive tissues (Labrie *et al*., 2003). A further possibility is that androgen-independent expression of AR target genes is caused by expression of AR coactivators or factors that cooperate with AR, such as SPDEF, an ets-related transcription factor active in prostatic epithelium (Oettgen *et al*., 2000). SPDEF is expressed in the apocrine and luminal clusters (Figure 1c panel 6). It is itself an androgen-regulated gene, and could thus participate in a positive feedback loop (Nelson *et al*., 2002). Beyond steroid receptor activity, the mechanism of basal, luminal or apocrine fate selection is unknown, in part because the factors mediating lineage choice in the normal breast are poorly understood.

The clinical features of apocrine carcinoma are generally considered not to differ appreciably from conventional forms of ductal carcinoma (Frable and Kay, 1968; Rosen, 1996). Although apocrine carcinoma is rare, tumour cells with apocrine features have been reported in up to 63% of cases (Miller *et al*., 1985). The poor survival of the patients with the molecular apocrine profile in the Sorlie and van't Veer studies (Figure 5b) is consistent with the poor prognosis expected for ER-negative, ERBB2-positive tumours. The gene cluster shared by the basal and molecular apocrine tumours (Figure 1c panel 3) includes several genes induced by hypoxia, including ERO1L, ADM and NDRG1 (Cormier-Regard *et al*., 1998; Park *et al*., 2000; Gess *et al*., 2003), which would again be consistent with a poor prognosis. Taken together, these data suggest that the molecular apocrine profile does not identify only patients with classical apocrine carcinomas.

The basal, luminal and apocrine groups differ in the expression of many genes which are potential or actual drug targets. Androgens have been tested on many occasions for the treatment of breast cancer (reviewed by Santen *et al*., 1990; Birrell *et al*., 1998). The assumption behind old studies was that androgen would inhibit the proliferation of breast cancer cells (reviewed by Labrie *et al*., 2003), but it is now known that androgen treatment stimulates the proliferation of some breast cancer cell lines and inhibits the proliferation of others (Birrell *et al*., 1995). The question raised by our data is whether apocrine tumours, which in some respects resemble prostate cancer cells, would benefit from androgen blockade. Previous studies failed to show any significant activity of antiandrogens in breast cancer (Millward *et al*., 1991). Future studies of endocrine therapy should take into account the division of breast cancer into luminal (ER+ AR+), apocrine (ER− AR+) and basal (ER− AR−) groups, which should, in the simplest case, receive oestrogen blockade, androgen blockade and no steroid-based therapy, respectively. Failure to stratify treatment in this way would make it very difficult to detect a true benefit in the molecular apocrine group. Besides AR, it is noteworthy that the apocrine group differs in the expression of genes involved in androgen and lipid metabolism, many of which could, in principle, be targeted by drugs. For example, the apocrine tumours express HMGCR, the target of statins. Inhibition of HMGCR has been reported to inhibit growth of breast cancer cell lines

through a reduction in geranylgeranylation of rhoA (Denoyelle *et al*., 2003). The apocrine tumours also express GHR, PRLR and EGFR, which can be inhibited using a variety of direct and indirect approaches (Santen *et al*., 1990; Fuh *et al*., 1992; Chen *et al*., 1999; Wennbo and Tornell, 2000; Paez *et al*., 2004). The potential therapeutic implications of identifying molecular apocrine tumours suggest that further studies to produce a more robust definition of the molecular apocrine profile and the mechanisms leading to its expression are justified.

Materials and methods

Biopsies of large operable or locally advanced/inflammatory breast tumours were taken before treatment from patients enrolled in a neoadjuvant clinical trial (EORTC 10994). Total RNA was extracted from $4 \times 25\,\mu$m sections of biopsies and amplified by an Eberwine T7 procedure according to the Affymetrix small sample protocol (Affymetrix, Santa Clara, USA). There was no obvious bias related to RNA quality, as assessed by Agilent Bioanalyzer curves, despite the inclusion of biopsies from patients treated in 14 centres throughout Europe. Details of microarray procedures and data analysis techniques are given in the supplementary methods file. The raw data have been deposited in the NCBI GEO database with series accession number GSE1561. The decision rule to assign tumours to classes was that the Spearman correlation coefficient should exceed the threshold at which a class would be assigned in 1% of tests with scrambled data. If more than one class exceeded this threshold, a class was assigned only if the correlation coefficient exceeded the next best value by the same threshold amount. Membership of the groups used for the different tests is given in the supplementary methods file.

Acknowledgements
We thank the women participating in the EORTC 10994/BIG 00-01 study for generously donating tumour samples. We thank the doctors, nurses and data managers from the European Organization for Research and Treatment of Cancer (EORTC), the Anglo-Celtic Cooperative Oncology Group (ACCOG), the Swiss Group for Clinical Cancer Research (SAKK) and the Swedish Breast Cancer Group (SweBCG) for their active participation. We thank Annick Ducraux for technical assistance and Monica de Vos for data management. We thank Dr Wassim Raffoul for providing reduction mammoplasty tissue. We thank Dr Patrick Descombes for advice on chip hybridization and the Geneva NCCR 'Frontiers in Genomics' for use of their Affymetrix workstation. We thank Dr Michael Morris for measuring AR repeat length. We thank Drs Pascale Anderle, Thierry Sengstag and Viviane Praz for advice on Io and Cleanex. We thank Dr Felix Naef for helpful discussions on data analysis. We thank the Swiss National Science Foundation NCCR Molecular Oncology program, MEDIC Foundation, EORTC Translational Research Fund and Oncosuisse for financial support.

References

Bertucci F, Borie N, Ginestier C, Groulet A, Charafe-Jauffret E, Adelaide J, Geneix J, Bachelart L, Finetti P, Koki A, Hermitte F, Hassoun J, Debono S, Viens P, Fert V, Jacquemier J and Birnbaum D. (2004). *Oncogene*, **23,** 2564–2575.

Birrell SN, Bentel JM, Hickey TE, Ricciardelli C, Weger MA, Horsfall DJ and Tilley WD. (1995). *J. Steroid Biochem. Mol. Biol.*, **52,** 459–467.

Birrell SN, Hall RE and Tilley WD. (1998). *J. Mammary Gland Biol. Neoplasia*, **3,** 95–103.

Chamberlain NL, Driver ED and Miesfeld RL. (1994). *Nucleic Acids Res.*, **22,** 3181–3186.

Chen WY, Ramamoorthy P, Chen N, Sticca R and Wagner TE. (1999). *Clin. Cancer Res.*, **5,** 3583–3593.

Cormier-Regard S, Nguyen SV and Claycomb WC. (1998). *J. Biol. Chem.*, **273,** 17787–17792.

Denoyelle C, Albanese P, Uzan G, Hong L, Vannier JP, Soria J and Soria C. (2003). *Cell Signal.*, **15,** 327–338.

Ellis I. (2003). *Pathology and Genetics of Tumours of the Breast and Female Genital Organs, Vol. 5: World Health Organization classification of tumours* Tavassoli F, Devilee P (eds) IARC Press: Lyon, pp 13–59.

Frable WJ and Kay S. (1968). *Cancer*, **21,** 756–763.

Frasor J, Danes JM, Komm B, Chang KC, Lyttle CR and Katzenellenbogen BS. (2003). *Endocrinology*, **144,** 4562–4574.

Fuh G, Cunningham BC, Fukunaga R, Nagata S, Goeddel DV and Wells JA. (1992). *Science*, **256,** 1677–1680.

Gess B, Hofbauer KH, Wenger RH, Lohaus C, Meyer HE and Kurtz A. (2003). *Eur. J. Biochem.*, **270,** 2228–2235.

Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR and Huang AT. (2003). *Lancet*, **361,** 1590–1596.

Isola JJ, Kallioniemi OP, Chu LW, Fuqua SA, Hilsenbeck SG, Osborne CK and Waldman FM. (1995). *Am. J. Pathol.*, **147,** 905–911.

Jain AN, Chin K, Borresen-Dale AL, Erikstein BK, Eynstein Lonning P, Kaaresen R and Gray JW. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 7952–7957.

Jones C, Damiani S, Wells D, Chaggar R, Lakhani SR and Eusebi V. (2001). *Am. J. Pathol.*, **158,** 207–214.

Labrie F, Luu-The V, Labrie C, Belanger A, Simard J, Lin SX and Pelletier G. (2003). *Endocr. Rev.*, **24,** 152–182.

Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, Zahnow CA, Patterson N, Golub TR and Ewen ME. (2003). *Cell*, **114,** 323–334.

La Spada AR, Wilson EM, Lubahn DB, Harding AE and Fischbeck KH. (1991). *Nature*, **352,** 77–79.

Levesque E, Turgeon D, Carrier JS, Montminy V, Beaulieu M and Belanger A. (2001). *Biochemistry*, **40,** 3869–3881.

Liao DJ and Dickson RB. (2002). *J. Steroid. Biochem. Mol. Biol.*, **80,** 175–189.

Lopez-Otin C and Diamandis EP. (1998). *Endocr. Rev.*, **19,** 365–396.

Mellinghoff IK, Vivanco I, Kwon A, Tran C, Wongvipat J and Sawyers CL. (2004). *Cancer Cell*, **6,** 517–527.

Miller WR, Telford J, Dixon JM and Shivas AA. (1985). *Breast Cancer Res. Treat.*, **5,** 67–73.

Millward MJ, Cantwell BM, Dowsett M, Carmichael J and Harris AL. (1991). *Br. J. Cancer*, **63,** 763–764.

Moinfar F, Okcu M, Tsybrovskyy O, Regitnig P, Lax SF, Weybora W, Ratschek M, Tavassoli FA and Denk H. (2003). *Cancer*, **98,** 703–711.

Molist R, Remvikos Y, Dutrillaux B and Muleris M. (2004). *Oncogene*, **23,** 5986–5993.

Nelson PS, Clegg N, Arnold H, Ferguson C, Bonham M, White J, Hood L and Lin B. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 11890–11895.

Oettgen P, Finger E, Sun Z, Akbarali Y, Thamrongsak U, Boltax J, Grall F, Dube A, Weiss A, Brown L, Quinn G, Kas K, Endress G, Kunsch C and Libermann TA. (2000). *J. Biol. Chem.*, **275,** 1216–1225.

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE and Meyerson M. (2004). *Science*, **304,** 1497–1500.

Park H, Adams MA, Lachat P, Bosman F, Pang SC and Graham CH. (2000). *Biochem. Biophys. Res. Commun.*, **276,** 321–328.

Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO and Botstein D. (1999). *Proc. Natl. Acad. Sci. USA*, **96,** 9212–9217.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO and Botstein D. (2000). *Nature*, **406,** 747–752.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL and Brown PO. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 12963–12968.

Rosen PP. (1996). *Diseases of the Breast* Harris JR (ed) Lippincott: Philadelphia, pp 413–414.

Rousseeuw PJ. (1987). *J. Comput. Appl. Math.*, **20,** 53–65.

Santen RJ, Manni A, Harvey H and Redmond C. (1990). *Endocr. Rev.*, **11,** 221–265.

Schmitt FC and Reis-Filho JS. (2002). *Breast*, **11,** 463–465.

Selim AG, El-Ayat G and Wells CA. (2000). *J. Pathol.*, **191,** 138–142.

Selim AG and Wells CA. (1999). *J. Clin. Pathol.*, **52,** 838–841.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P and Borresen-Dale AL. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 10869–10874.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL and Botstein D. (2003). *Proc. Natl. Acad. Sci. USA*, **100,** 8418–8423.

Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL and Liu ET. (2003). *Proc. Natl. Acad. Sci. USA*, **100,** 10393–10398.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH. (2002). *Nature*, **415,** 530–536.

Viacava P, Naccarato AG and Bevilacqua G. (1997). *Virchows Arch.*, **431,** 205–209.

Wennbo H and Tornell J. (2000). *Oncogene*, **19,** 1072–1076.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson Jr JA, Marks JR and Nevins JR. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 11462–11467.

Wu L, Einstein M, Geissler WM, Chan HK, Elliston KO and Andersson S. (1993). *J. Biol. Chem.*, **268,** 12964–12969.

Yu H, Bharaj B, Vassilikos EJ, Giai M and Diamandis EP. (2000). *Breast Cancer Res. Treat.*, **59,** 153–161.

Supplementary Information accompanies the paper on Oncogene website (http://www.nature.com/onc)