Breast Cancer Res Treat (2015) 149:425–437
DOI 10.1007/s10549-014-3262-6

PRECLINICAL STUDY

Luminal progenitor and fetal mammary stem cell expression features predict breast tumor response to neoadjuvant chemotherapy

Adam D. Pfefferle · Benjamin T. Spike · Geoff M. Wahl · Charles M. Perou

Received: 22 December 2014 / Accepted: 23 December 2014 / Published online: 10 January 2015

Abstract Mammary gland morphology and physiology are supported by an underlying cellular differentiation hierarchy. Molecular features associated with particular cell types along this hierarchy may contribute to the biological and clinical heterogeneity observed in human breast carcinomas. Investigating the normal cellular developmental phenotypes in breast tumors may provide new prognostic paradigms, identify new targetable pathways, and explain breast cancer subtype etiology. We used transcriptomic profiles coming from fluorescence-activated cell sorted (FACS) normal mammary epithelial cell types from several independent human and murine studies. Using a meta-analysis approach, we derived consensus gene signatures for both species and used these to relate tumors to normal mammary epithelial cell phenotypes. We then compiled a dataset of breast cancer patients treated with neoadjuvant anthracycline and taxane chemotherapy regimens to determine if normal cellular traits predict the likelihood of a pathological complete response (pCR) in a multivariate logistic regression analysis with clinical markers and genomic features such as cell proliferation. Most human and murine tumor subtypes shared some, but not all, features with a specific FACS-purified normal cell type; thus for most tumors a potential distinct cell type of 'origin' could be assigned. We found that both human luminal progenitor and mouse fetal mammary stem cell features predicted pCR sensitivity across all breast cancer patients even after controlling for intrinsic subtype, proliferation, and clinical variables. This work identifies new clinically relevant gene signatures and highlights the value of a developmental biology perspective for uncovering relationships between tumor subtypes and their potential normal cellular counterparts.

Keywords Breast cancer · Comparative genomics · Genetically engineered mouse models · Genomic signatures · Neoadjuvant chemotherapy · Normal mammary tissue

Electronic supplementary material The online version of this article (doi:10.1007/s10549-014-3262-6) contains supplementary material, which is available to authorized users.

A. D. Pfefferle · C. M. Perou
Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: adamp@email.unc.edu

A. D. Pfefferle · C. M. Perou (✉)
Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: cperou@med.unc.edu

B. T. Spike · G. M. Wahl
Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92130, USA
e-mail: bspike@salk.edu

G. M. Wahl
e-mail: wahl@salk.edu

C. M. Perou
Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Introduction

The mammalian breast is a dynamic organ, with major morphological changes occurring during organogenesis, puberty, pregnancy, lactation, and involution [1]. Underlying these mammary gland changes is a complex cell hierarchy that supports these processes [2–4]. The simplest model places the multipotent mammary stem cell (MaSC) at the base of this

hierarchy, having extensive, self-regenerative potential [5]. During mammary development, the MaSC has been proposed to divide asymmetrically to produce basal/myoepithelial cells as well as luminal progenitors (LumProg), which have more restricted proliferative and differentiation capabilities [5]. LumProg cells are capable of further differentiation into mature luminal (MatureLum) cells, such as estrogen receptor (ER)-positive ductal epithelium, which have an even more limited proliferative potential and some of which are terminally differentiated [5].

Breast tumors may originate from several, if not all, of the cell types within this complex mammary hierarchy. These various cellular foundations for tumor initiation may help explain the heterogeneous nature of human breast tumors [6], which consist of multiple histological and genomic subtypes; these genomic groups, which are defined by their gene expression profiles, have become known as the intrinsic subtypes of breast cancer and are referred to as basal-like, claudin-low, HER2-enriched, luminal A, and luminal B [7–10]. A simple etiological explanation for these different subtypes involves a one-to-one relationship between each intrinsic subtype and a distinct cell type of origin that largely maintains its phenotypic identity after oncogenic transformation; however, both normal and neoplastic non-stem cells can acquire stem-like properties, suggesting that the normal cell hierarchy model could also include an element of reversibility [11]. This also raises the possibility that molecular features defining tumor subtypes, may be acquired during tumorigenesis [12].

Genetically engineered mouse models (GEMMs) of breast carcinoma develop heterogeneous tumors [13, 14], but the extent to which they represent human disease is an area of active investigation. We previously showed that murine mammary tumors comprise at least 17 distinct intrinsic subtypes/classes, with eight classes being identified as strong human subtype counterparts by gene expression similarity [14]. As with human breast cancer, the degree to which murine models reflect normal mammary epithelial subpopulations requires further analysis. Characterization of the cellular features of these murine classes is also needed to better determine their preclinical utility, to shed light on trans-species associations [14], and to help interpret preclinical study observations [15–18].

Several studies have independently profiled fluorescence-activated cell sorted (FACS) purified normal mammary cell types from both human [19–21] and murine [22, 23] mammary tissues. Here, we use a meta-analysis approach to compare the transcriptomic profiles from FACS-enriched mammary cell populations with each other and with primary tumors. These data not only identify a number of clinically relevant biomarkers that may be useful for predicting chemotherapy benefit, but also suggest a cell type of origin for many tumor subtypes.

Methods

Detailed methods can be found in Supplemental File 1.

Mammary cell subpopulation gene signatures

Gene expression measurements from FACS-enriched mammary subpopulations were obtained from three human and two murine published studies: GSE16997 [19], GSE19446 [22], GSE27027 [23], GSE35399 [20], and GSE50470 [21]. Using a meta-analysis approach, a consensus 'enriched' gene signature was produced for each mammary subpopulation. 'Enriched' signatures comprised genes that were identified as being uniquely and highly expressed (false discovery rate (FDR) < 5 %) within a given subpopulation as determined using a two-class (subpopulation X versus all others) significance analysis of microarrays (SAM) analysis [9, 24]. Each 'enriched' signature was further refined by supervised clustering using the human UNC308 breast tumor dataset [9] to identify subpopulation 'features', which were defined as having at least ten genes with a Pearson correlation greater than 0.5 across all tumors [15, 25]. Expression scores for gene signatures were determined by calculating the mean expression of the signature within each tumor; all gene signature lists are provided in Supplemental Table 1.

Mammary cell subpopulation centroids

Mammary cell subpopulation centroids were created using the union of the 'enriched' epithelial gene signatures. Distance weighted discrimination (DWD) single sample predictor [26] was used to calculate the shortest Euclidean distance between each tumor and each epithelial cell-enriched centroid. Samples with a positive silhouette width were considered to have a strong association with a given subpopulation [27].

Chemotherapy response

A combined breast cancer gene expression dataset of patients treated with neoadjuvant anthracycline and taxane chemotherapy regimens was created from three public datasets: GSE25066 [28], GSE32646 [29], and GSE41998 [30]. Univariate (UVA) and multivariate (MVA) logistic regression analyses were used to determine if gene signatures derived from normal cell populations were capable of predicting pathological complete response (pCR).

Results

Comparison of human mammary subpopulation transcriptomic datasets

Several groups have independently obtained transcriptomic profiles of normal human breast cells and compared the genomic biology of these different cell types with human tumors [19–21]. In these studies, normal mammary tissues obtained from female donors were FAC sorted using cell surface markers to enrich for specific mammary subpopulations before microarray analysis (Table 1; Fig. 1). While these initial studies were important, the datasets themselves were relatively small ($n = 12$ for Lim et al. [19], $n = 72$ for Shehata et al. [20], $n = 18$ for Prat et al. [21]), and few if any comparisons across studies were performed. Importantly, FACS-based cell fractionation can only enrich for specific subpopulations. Therefore, transcriptomic profiles reflect features of other contaminating cell types to varying degrees. As such, study-specific biases may be present in any single dataset; therefore, we used consensus information from all three FACS-enriched human transcriptomic datasets to reduce technical and study-specific biases.

Following DWD normalization [26], an unsupervised cluster of the most variably expressed genes was performed using Gene Cluster v3.0 by selecting all genes with an absolute $\log_2$ expression value greater than three in at least four samples (212 genes) (Fig. 2a). In general, the four major array dendrogram nodes correspond to the four FACS-enriched mammary subpopulations, indicating that the most highly and variably expressed genes are similarly expressed across the different studies. Even when using all genes in the dataset, there is a high Pearson correlation within a given subpopulation across studies and low correlations to other subpopulations (Fig. 2b).

On a per-sample basis, the first principle component separated the stroma and adult mammary stem cell (aMaSC) samples from the LumProg and MatureLum samples (Fig. 2c). The second principle component separated the stroma and aMaSC samples into distinct groups, while the third principle component separated the LumProg and MatureLum samples into distinct groups. The aMaSC subpopulation displayed the highest level of variation, which is likely attributable to varying degrees of contamination by other cell types.

Human mammary cell subpopulation enriched gene signatures

As shown in Fig. 2, there is a natural degree of variation between samples of a given subpopulation. We therefore developed gene signatures for each human mammary subpopulation by integrating consensus information across all three datasets (Table 1) to identify the highest confidence subpopulation-specific genes. First, genes highly expressed (FDR < 5 %) within each mammary subpopulation were found using a two-class (subpopulation X versus all others) SAM analysis [24] within each dataset [19–21]. Second, the overlap of genes highly expressed within a particular subpopulation across studies was determined. Lastly, as it is possible in the above analysis to have the same gene in the signature of more than one subpopulation, genes that were identified to be significantly associated with more than one subpopulation were also removed. This resulted in a single, consensus *Homo sapiens*-enriched (HsEnriched) signature per subpopulation (Fig. 3a). The average Euclidean distance was

Table 1 Human FACS-enriched normal mammary cell subpopulation studies

Enriched population	FACS markers	Species	Source	Abbreviation	Reference
Stroma	CD49fneg, EpCAMneg	Human	Adult	aStr-Lim09	Lim et al. [19]
	CD49fneg, EpCAMneg	Human	Adult	aStr-Shehata	Shehata et al. [20]
	CD49fneg, EpCAMneg	Human	Adult	aStr-Prat	Prat et al. [21]
Stem cell	CD49fpos, EpCAMneg	Human	Adult	aMaSC-Lim09	Lim et al. [19]
	CD49fpos, EpCAMneg	Human	Adult	aMaSC-Shehata	Shehata et al. [20]
	CD49fpos, EpCAMneg	Human	Adult	aMaSC-Prat	Prat et al. [21]
Luminal progenitor	CD49fpos, EpCAMpos	Human	Adult	LumProg-Lim09	Lim et al. [19]
	CD49fpos, EpCAMpos	Human	Adult	LumProg-Shehata	Shehata et al. [20]
	CD49fpos, EpCAMpos	Human	Adult	LumProg-Prat	Prat et al. [21]
Mature luminal	CD49fneg, EpCAMpos	Human	Adult	MatureLum-Lim09	Lim et al. [19]
	CD49fneg, EpCAMpos	Human	Adult	MatureLum-Shehata	Shehata et al. [20]
	CD49fneg, EpCAMpos	Human	Adult	MatureLum-Prat	Prat et al. [21]

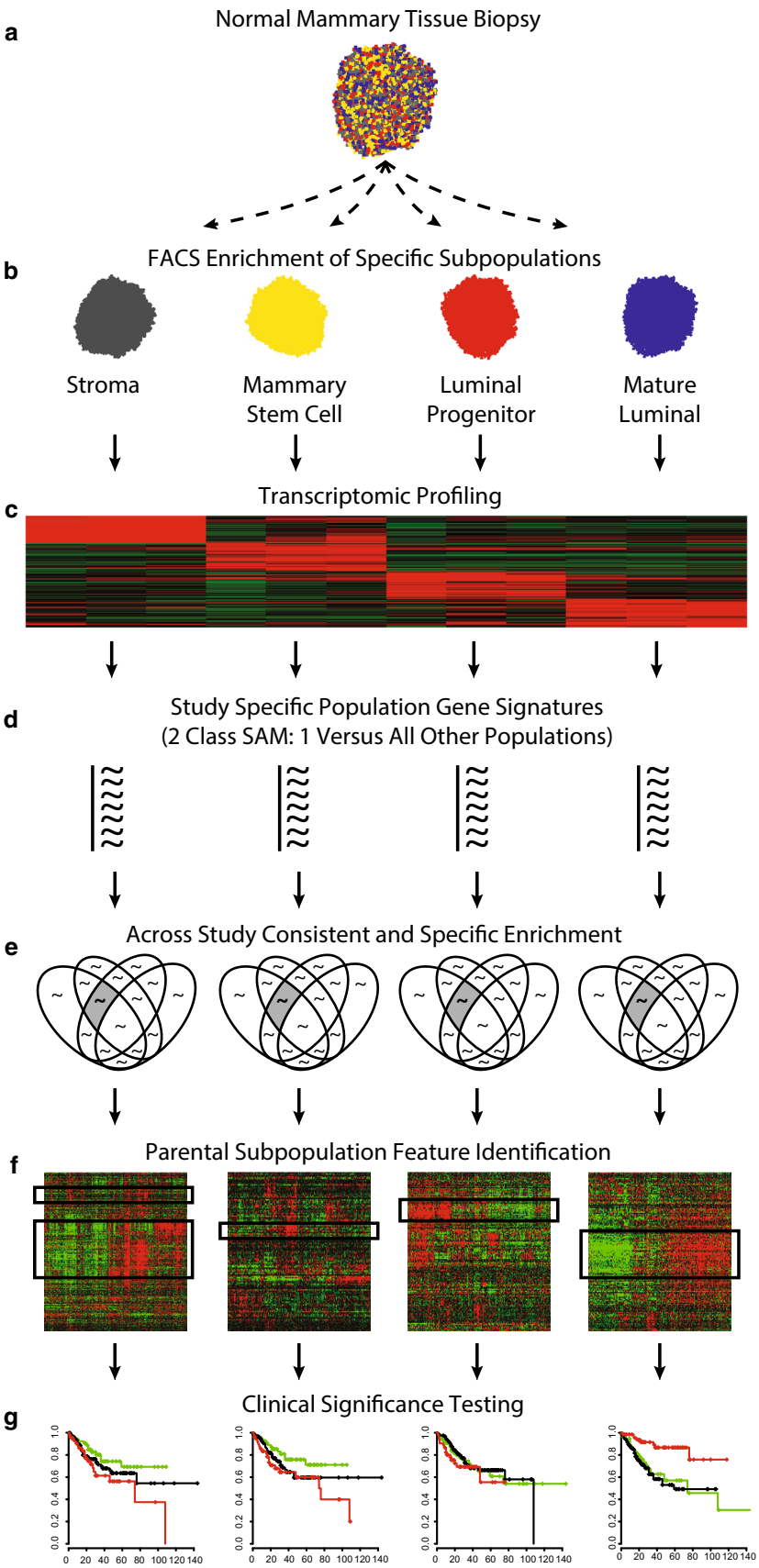


Fig. 1 Flowchart of analysis. Normal mammary tissue biopsies were taken from female patients (**a**) and FACS-enriched into distinct mammary cell subpopulations (**b**). Transcriptome profiling was performed on each subpopulation using gene expression microarrays by three different studies (**c**). Within each study, genes highly expressed within each subpopulation were determined using a two-class SAM (**d**). Genes commonly and specifically enriched within each subpopulation across studies were determined to identify 'enriched' gene signatures (**e**). Each 'enriched' signature was refined by supervised hierarchical clustering to identify gene 'features' highly correlated across a diverse set of human breast tumors (**f**). These gene signatures were then used for clinical testing (**g**)

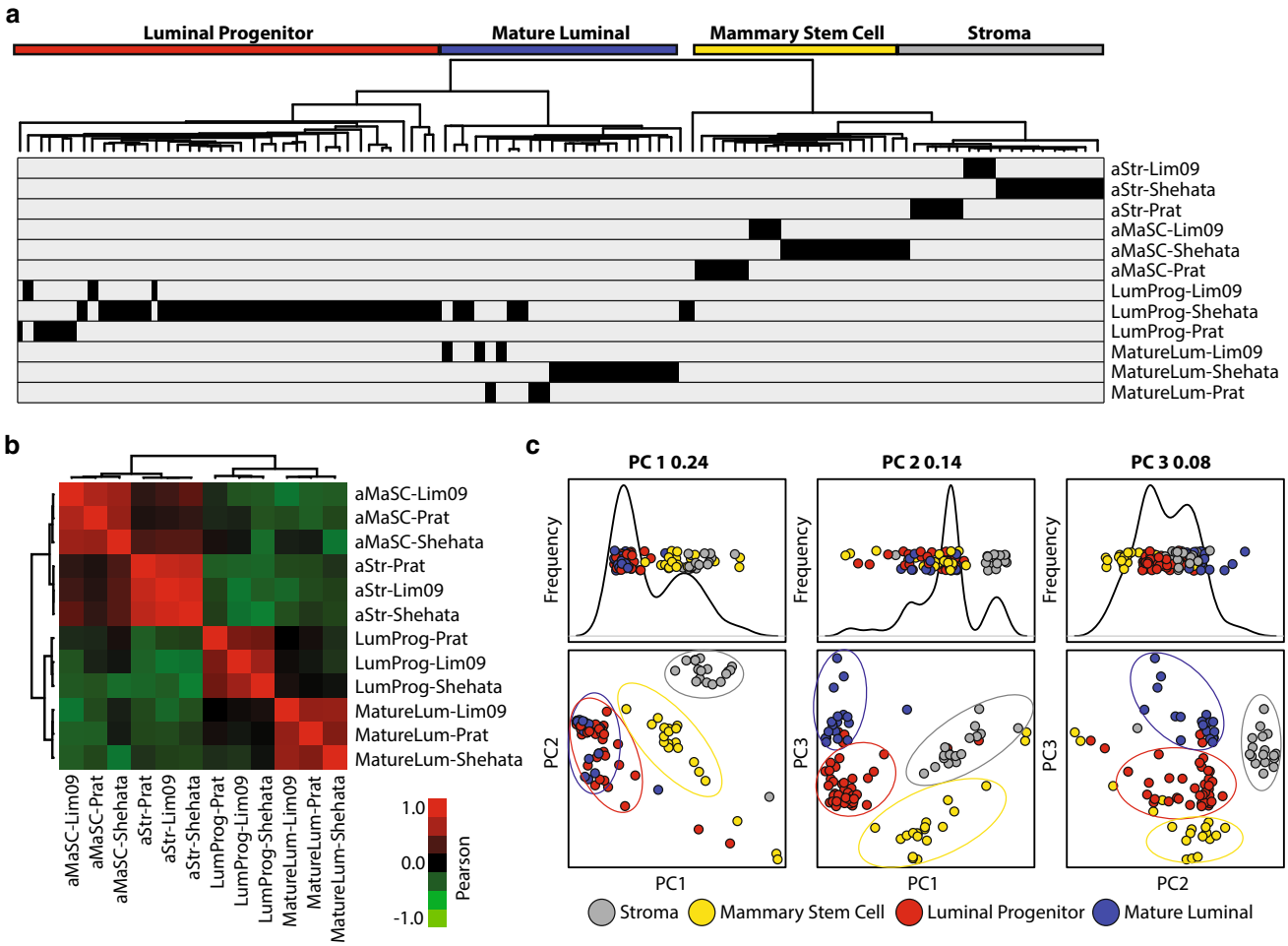


Fig. 2 Comparison of mammary subpopulations across studies. **a** Unsupervised hierarchical clustering was performed with the normal human mammary subpopulation dataset using any gene that had a $\log_2$ absolute expression value greater than three in at least four samples. **b** Pearson correlations were determined between the average expressions of each study's subpopulations using all genes. **c** The first three principle components were determined across the human mammary subpopulation dataset

determined using a 10-fold cross validation for each normal mammary subpopulation sample to centroids created using either the HsEnriched-derived gene signatures or to centroids created using the gene signatures derived separately from each human study (Supplemental Fig. 1). The HsEnriched centroids had a significantly reduced Euclidean distance ($\sim$70 %) to each mammary subpopulation (t test $p < 0.0001$), indicating greater specificity for the consensus HsEnriched signatures when compared with any individual dataset's subpopulation signature.

We next evaluated the utility of these signatures for distinguishing human tumor subtypes. Figure 3b displays the standardized average expression of each HsEnriched signature across the human intrinsic breast tumor subtypes [7, 9] using over 3,000 tumors [9, 31, 32]. The aStr-HsEnriched signature was highest in claudin-low and normal-like tumors. Interestingly, claudin-low tumors also highly express the aMaSC-HsEnriched signature. High expression of the aMaSC-HsEnriched signature in claudin-low tumors is unlikely an artifact of stromal cells in these tumors since the Pearson correlation between the aStr-HsEnriched and aMaSC-HsEnriched signatures was −0.19 across the normal human mammary samples. The LumProg and MatureLum-HsEnriched signatures were most highly expressed in basal-like and luminal subtype tumors, respectively (Fig. 3b).

We noted a considerable degree of signature variation within a subtype, indicating that it is not necessarily the case that all tumors of a given subtype share features with the same normal cell type. A nearest centroid predictor with a 10-fold cross validation error rate of 4.8 % was created to individually determine which normal mammary epithelial subpopulation is most similar to each tumor. Samples with positive silhouette widths [27] were considered to have a strong association with their particular subpopulation, with all other tumors being categorized as

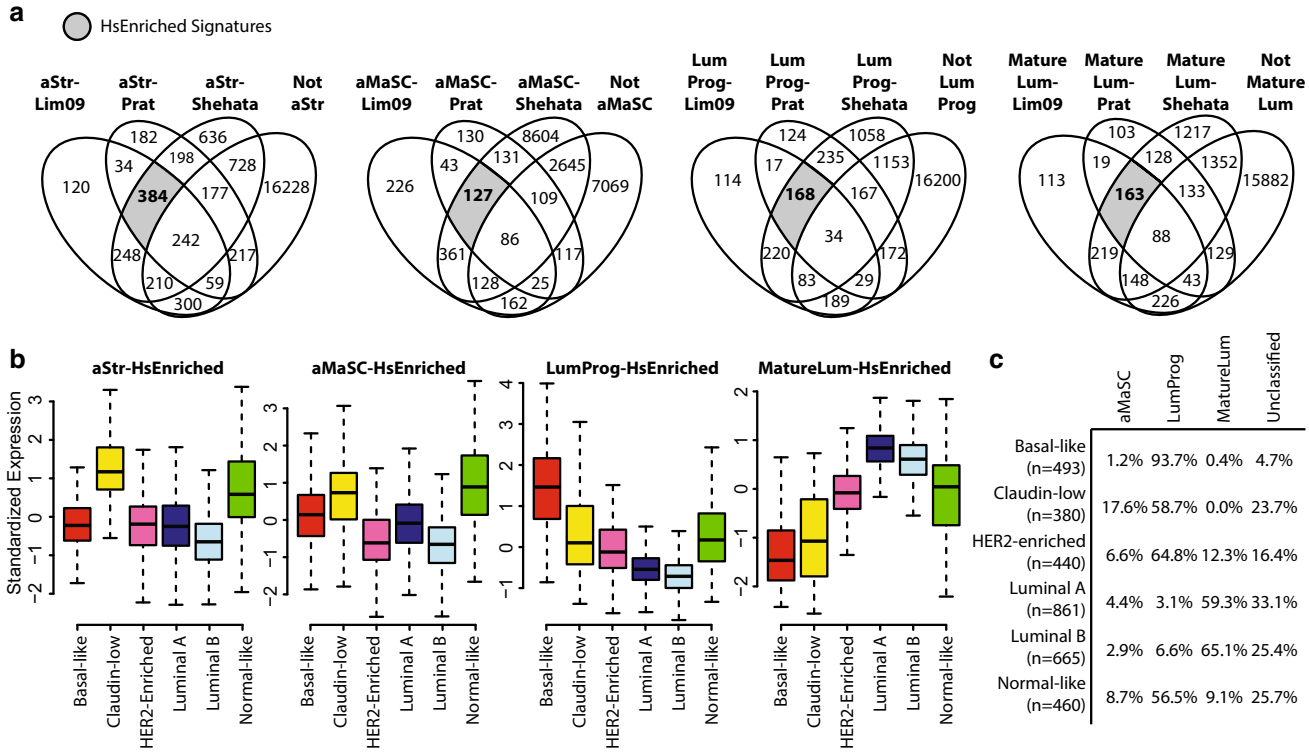


Fig. 3 *Homo sapiens*-enriched gene signatures. **a** HsEnriched gene signatures were identified for each mammary subpopulation. First, the overlap of genes highly expressed within each subpopulation across studies was determined. This overlapping gene set was further filtered to remove genes also identified as enriched in another subpopulation to limit the signature to genes specific to an individual subpopulation. The remaining genes comprised the HsEnriched gene signature for that subpopulation, as indicated by the *shaded box*. **b** The standardized average expression of the four HsEnriched gene signatures was calculated across three human datasets and displayed by intrinsic tumor subtype. **c** A nearest centroid predictor using the HsEnriched gene signatures was used to determine which epithelial features each tumor most represented. To reduce spurious findings, any tumor with a negative silhouette width was considered to have a weak association and was labeled as 'unclassified'

'unclassified' [33] (Fig. 3c). Specifically, 94 % of basal-like tumors had LumProg expression profiles. The claudin-low subtype had the highest percentage of tumors classified as aMaSC (18 %), although most claudin-low tumors were classified as having LumProg features (59 %). The HER2-enriched subtype was predominantly classified as having LumProg expression features. The luminal A and B subtypes were most similar to the MatureLum subpopulation.

Murine mammary cell subpopulation enriched gene signatures

Several groups have also profiled normal murine mammary cell subpopulation expression features using FACS [22, 23] (Table 2). In addition to highlighting conserved expression features across species [22], murine studies are uniquely positioned to enable comparisons with developmental states not easily accessed in humans, including early fetal development [23]. We were particularly interested in fetal mammary stem cells (fMaSC) [23], which is a distinct cell population not captured in any human study performed thus far (Table 3). Using the same approach that we used to derive the HsEnriched signatures, we created *Mus musculus*-enriched (MmEnriched) signatures for each murine mammary subpopulation (Fig. 4a) [22, 23].

We calculated the standardized average expression of each MmEnriched signature across the murine intrinsic subtypes/classes (Fig. 4b) [14]. As in human tumors, the Str-MmEnriched signature was most highly expressed in Normal-like^{Ex} and Claudin-low^{Ex}; this common feature was anticipated given the high similarity of these two classes to their human subtype counterparts and their known enrichment for stroma-associated genes [14, 23]. The aMaSC-MmEnriched signature was most highly expressed in Class14^{Ex} and to a slightly lesser extent in Wnt1-Late^{Ex}, Wnt1-Early^{Ex}, p53null-Basal^{Ex}, and Squamous-like^{Ex}. The fMaSC-MmEnriched signature was most highly expressed in WapINT3^{Ex}, which is consistent with the finding that *Int3* (*Notch4*) inhibits mammary cell differentiation [34, 35]. The LumProg-MmEnriched signature was highest in PyMT^{Ex} and Neu^{Ex}. This finding was unexpected given that these two mouse classes have been shown to resemble

Table 2 Murine FACS-enriched normal mammary cell subpopulation studies

Enriched population	FACS markers	Species	Source	Abbreviation	Reference
Stroma	Cd24neg/low/med	Mouse	Fetal	fStr-Spike	Spike et al. [23]
	Cd29neg, Cd24neg	Mouse	Adult	aStr-Lim10	Lim et al. [22]
Stem cell	Cd49fhi, Cd24hi	Mouse	Fetal	fMaSC-Spike	Spike et al. [23]
	Cd49fhi, Cd24med	Mouse	Adult	aMaSC-Spike	Spike et al. [23]
	Cd29pos, Cd24pos, Cd61pos	Mouse	Adult	aMaSC-Lim10	Lim et al. [22]
Luminal progenitor	Cd29neg, Cd24pos, Cd61pos	Mouse	Adult	LumProg-Lim10	Lim et al. [22]
Mature luminal	Cd29neg, Cd24pos, Cd61neg	Mouse	Adult	MatureLum-Lim10	Lim et al. [22]

Table 3 Gene set analysis of human and murine cell subpopulations

Murine subpopulation	Human subpopulation			
	Str	aMaSC	LumProg	MatureLum
Str	**0.044**	–	–	–
fMaSC	–	–	0.4395	0.4395
aMaSC	–	**0.044**	–	–
LumProg	–	–	**0.042**	0.386
MatureLum	–	0.464	0.306	**0.004**

A comparative analysis of each human subpopulation versus each murine subpopulation was performed using GSA. The FDR is displayed for all comparisons with a positive association. Statistically significant associations (FDR < 0.05) are bolded

luminal human tumors [13, 14]. Lastly, the MatureLum-MmEnriched signature was most highly expressed in Stat1^{Ex} and Class14^{Ex}. Both the *Stat1*$^{-/-}$ and *Pik3ca*-H1047R mouse models, which define these two classes respectively, are often ER positive [36, 37], and these data suggest that they have MatureLum features. Class14^{Ex} also exhibited significant expression of the aMaSC-MmEnriched signature, indicating that these tumors contain a mixture or share features of multiple cell types.

Consistent with Fig. 4b, 91 % of WapINT3^{Ex} tumors were classified as having fMaSC features in a nearest centroid predictor analysis. Mouse luminal classes of breast carcinoma (Erbb2-like^{Ex}, Myc^{Ex}, PyMT^{Ex}, and Neu^{Ex}) were most similar to LumProg cells, which again were unexpected but consistent with previous findings [22, 38]. Wnt1-Early^{Ex}, p53null-Basal^{Ex}, and Squamous-like^{Ex} tumors had primarily aMaSC features. Interestingly, Claudin-low^{Ex} and to a lesser extent C3-Tag^{Ex} tumors also had aMaSC features. All Stat1^{Ex} tumors had MatureLum features, consistent with being ER positive [36].

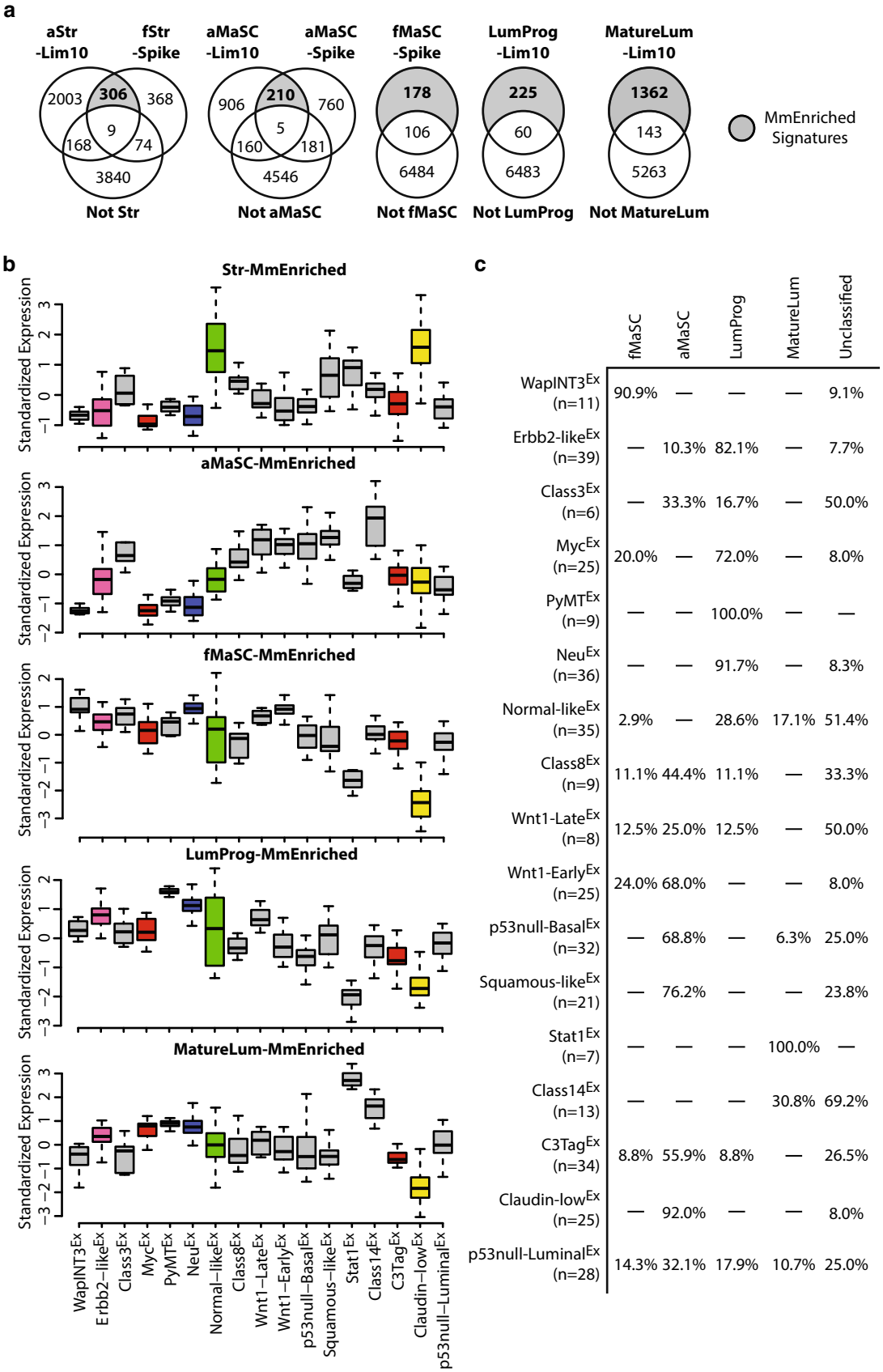
LumProg and fMaSC features predict neoadjuvant chemotherapy response

Breast tumors respond heterogeneously to neoadjuvant chemotherapy treatment [15]. We hypothesized that cellular features of normal mammary subpopulations may identify tumors most likely to respond to neoadjuvant chemotherapy. To test this, we compiled a dataset of 702 neoadjuvant anthracycline and taxane chemotherapy-treated patients (Supplemental Table 2).

Although genes within each 'enriched signature' are highly correlated within their respective normal cell subpopulation, it does not necessarily follow that all genes within a given normal cell signature would be as coordinately regulated in tumors. Therefore, we subdivided each signature into smaller features (feature1, feature2, etc.) that are coordinately expressed in tumors, reasoning that such refined 'features' may be more clinically robust. All 'enriched' and refined 'features' were tested for their ability to predict pCR to neoadjuvant chemotherapy in a UVA (Supplemental Table 3). UVA significant signatures ($p < 0.05$) were then considered in a MVA with age, ER status, PR status, HER2 status, tumor stage, PAM50 subtype [39], and PAM50 proliferation score [39] to determine if any mammary subpopulation 'features' added novel information for predicting pCR (Supplemental Table 4).

Six normal mammary gene signatures were UVA and MVA significant (Supplemental Tables 3 and 4), with the 95 % UVA odds ratio of these six signatures and all other 'enriched signatures' displayed in Fig. 5a. Interestingly, the LumProg-HsEnriched and LumProg-HsEnriched-feature1 signatures, both of which were highly correlated (Fig. 5b), were significant in the UVA and MVA analyses, indicating that tumors with LumProg features are more likely to respond to neoadjuvant treatment. Importantly, this response was independent of proliferation, as highlighted by their low correlation to the PAM50-Proliferation gene signature (Fig. 5b).

Interestingly, the fMaSC-MmEnriched signature refined into two distinctly opposite, highly significant signatures in both the UVA and MVA (Supplemental Table 3, 4; Fig. 5b, c). While the fMaSC-MmEnriched signature was highest in basal-like tumors, the refined signatures varied,


a
aStr -Lim10
fStr -Spike
aMaSC -Lim10
aMaSC -Spike
fMaSC -Spike
LumProg -Lim10
MatureLum -Lim10
2003
306
368
9
168
74
3840
Not Str
906
210
760
5
160
181
4546
Not aMaSC
178
106
6484
Not fMaSC
225
60
6483
Not LumProg
1362
143
5263
Not MatureLum
MmEnriched Signatures
b
Str-MmEnriched
aMaSC-MmEnriched
fMaSC-MmEnriched
LumProg-MmEnriched
MatureLum-MmEnriched
Standardized Expression
WapINT3Ex
Erbb2−likeEx
Class3Ex
MycEx
PyMTEx
NeuEx
Normal−likeEx
Class8Ex
Wnt1−LateEx
Wnt1−EarlyEx
p53null−BasalEx
Squamous−likeEx
Stat1Ex
Class14Ex
C3TagEx
Claudin−lowEx
p53null−LuminalEx
c
fMaSC
aMaSC
LumProg
MatureLum
Unclassified
WapINT3Ex (n=11) 90.9% — — — 9.1%
Erbb2-likeEx (n=39) — 10.3% 82.1% — 7.7%
Class3Ex (n=6) — 33.3% 16.7% — 50.0%
MycEx (n=25) 20.0% — 72.0% — 8.0%
PyMTEx (n=9) — — 100.0% — —
NeuEx (n=36) — — 91.7% — 8.3%
Normal-likeEx (n=35) 2.9% — 28.6% 17.1% 51.4%
Class8Ex (n=9) 11.1% 44.4% 11.1% — 33.3%
Wnt1-LateEx (n=8) 12.5% 25.0% 12.5% — 50.0%
Wnt1-EarlyEx (n=25) 24.0% 68.0% — — 8.0%
p53null-BasalEx (n=32) — 68.8% — 6.3% 25.0%
Squamous-likeEx (n=21) — 76.2% — — 23.8%
Stat1Ex (n=7) — — — 100.0% —
Class14Ex (n=13) — — — 30.8% 69.2%
C3TagEx (n=34) 8.8% 55.9% 8.8% — 26.5%
Claudin-lowEx (n=25) — 92.0% — — 8.0%
p53null-LuminalEx (n=28) 14.3% 32.1% 17.9% 10.7% 25.0%

◀**Fig. 4** *Mus musculus*-enriched gene signatures. **a** MmEnriched gene signatures were identified for each mammary subpopulation. First, the overlap of genes highly expressed within each subpopulation across studies was determined. This overlapping gene set was further filtered to remove genes also identified as enriched in another subpopulation to limit the signature to genes specific to an individual subpopulation. The remaining genes comprised the MmEnriched gene signature for that subpopulation, as indicated by the *shaded box*. **b** The standardized average expression of the five MmEnriched gene signatures was calculated across a murine dataset and displayed by intrinsic tumor class. **c** A nearest centroid predictor using the MmEnriched gene signatures was used to determine which epithelial features each tumor most represented. To reduce spurious findings, any tumor with a negative silhouette width was considered to have a weak association and was labeled as 'unclassified'

with fMaSC-MmEnriched-feature1 (Fig. 5d) being highest in basal-like tumors and fMaSC-MmEnriched-feature2 (Fig. 5e) expressed in luminal tumors. Tumors with fMaSC-MmEnriched-feature1 expression were more likely to respond to neoadjuvant chemotherapy, while those tumors with fMaSC-MmEnriched-feature2 were more resistant. The fMaSC-MmEnriched-feature1 signature was very highly correlated with the LumProg-HsEnriched signatures (Fig. 5b), sharing four genes in common (Fig. 5d). These results support the hypothesis that subsets of genes within the larger 'enriched signature' are likely regulated by different biological mechanisms.

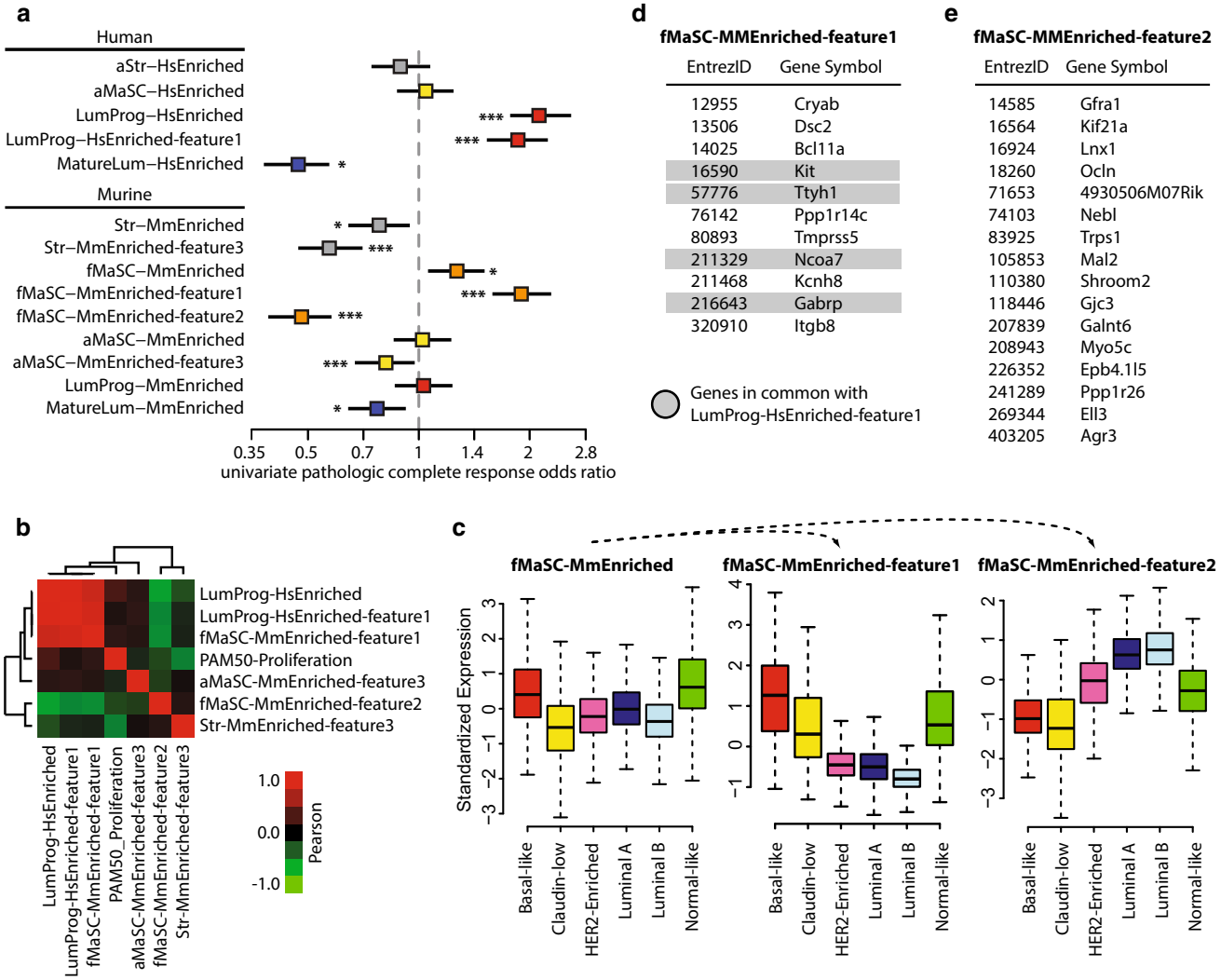


Fig. 5 fMaSC-enriched gene signatures. **a** The univariate logistic regression odds ratio predicting pathologic complete response to neoadjuvant anthracycline and taxane chemotherapy was determined using a 702 patient dataset, with the 95 % confidence interval shown as a forest plot. A single '*' indicates that the signature was univariate significant, while '***' indicates that the signature was both univariate and multivariate significant ($p < 0.05$). **b** Pearson correlations of multivariate significant gene signatures and proliferation were determined. **c** The standardized average expression of the fMaSC-MmEnriched signature and its two refined signatures were calculated across three human datasets and displayed by intrinsic tumor subtype. **d** Genes in the fMaSC-MmEnriched-refined1 signature. **e** Genes in the fMaSC-MmEnriched-refined2 signature

Discussion

Normal mammary gland physiology is supported by an underlying, complex cell hierarchy [2–5]. The simplest model treats differentiation from mammary stem cells to progenitor cells to mature cells as unidirectional, but recent observations indicate that bidirectional processes are also possible for normal and neoplastic cells [11]. This differentiation plasticity may allow tumors to acquire cell features foreign to the initial cell-of-origin or to lose native features through the accumulation of specific genetic aberrations [40].

Regardless of how different cellular traits are acquired, it is critical to identify the 'current' normal cellular features within a tumor, and therefore, we first analyzed the expression profiles of normal human and mouse mammary epithelial cell subpopulations [19–23]. We chose to use nomenclature that maintains continuity with the literature. However, these terms should be considered provisional as the complete biological profiles of these FACS fractions are investigated [4]. Recent work by Prater et al. [41] found that mouse 'LumProg' cells ($CD49f^{+}$, $EpCAM^{+}$) have complete mammary gland repopulating potential, indicating that 'LumProg' may be a misnomer. Importantly, even if our understanding and naming of these cell subpopulations change, only the retrospective interpretation of the data presented here will be affected, not the data itself.

Using a meta-analysis approach, FACS-purified mammary epithelial cell subpopulation 'enriched' gene signatures were derived and a nearest centroid predictor was developed to identify which normal mammary subpopulation each human and mouse tumor most represented using over three thousand human patients and 27 mouse models of mammary carcinoma [14]. While these analyses imply a cell-of-origin for a given tumor, additional experiments (e.g., lineage tracing) will be required to unequivocally determine this. Nevertheless, these associations at the very least identify which normal mammary subpopulation a given tumor most represents in its current state.

With this in mind, several associations between both the human and mouse intrinsic subtypes and specific normal cell subpopulations were observed. First, human basal-like tumors have been referred to as 'undifferentiated', which is consistent with their exhibiting LumProg [19] and fetal MaSC features [23]. Three mouse classes have been identified to be human basal-like counterparts: Myc^{Ex}, $p53null\text{-}Basal^{Ex}$, and $C3\text{-}Tag^{Ex}$ [14]. Myc^{Ex} tumors were the most similar to the LumProg cell profile. By contrast, both $p53null\text{-}Basal^{Ex}$ and $C3\text{-}Tag^{Ex}$ tumors had adult MaSC features. These results indicate that Myc^{Ex} tumors share similar cell features as their human basal-like counterpart, making it an attractive mouse model for studying basal-like tumors with aberrant Myc signaling [10, 42]. Interestingly, neither $p53null\text{-}Basal^{Ex}$ nor $C3\text{-}Tag^{Ex}$ tumors had strong LumProgs features, indicating that their association with human basal-like tumors is more likely driven by their underlying genetics [10].

Human claudin-low tumors had heterogeneous normal cell features. While most were similar to LumProg cells, the claudin-low subtype also had the largest percentage of tumors classified as adult MaSC. Given that claudin-low tumors are enriched with epithelial-to-mesenchymal transition features [9, 43, 44], our results suggest that these tumors may originate from the LumProg population prior to acquiring adult MaSC and/or mesenchymal features. Similarly, mouse $Claudin\text{-}low^{Ex}$ tumors were also strongly associated with the adult MaSC population, indicating that such tumors may be the closest analogs of the subset of human claudin-low tumors with adult MaSC features.

Human HER2-enriched tumors were the most similar to the LumProg subpopulation. This is a novel finding and may explain why both human basal-like and HER2-enriched subtype tumors show high TP53 mutation frequencies (>70 %) and widespread chromosomal instability [10]. These data could suggest that the normal LumProg cell is somehow extremely dependent on TP53 function. The murine $Erbb2\text{-}like^{Ex}$ class has been identified as a mouse counterpart for human HER2-enriched tumors [14] and was shown here to also have LumProg features.

When analyzing the human luminal A and B subtypes, a clear association with normal MatureLum cells was observed. The murine Neu^{Ex} class is a proposed counterpart for human luminal A tumors [14], yet these mouse tumors were most similar to normal mouse LumProg cells. The Myc^{Ex} class was also identified to resemble human luminal B tumors [14]. As discussed, Myc^{Ex} tumors have LumProg features; therefore, most mouse luminal A/B tumor models do not share the same normal cell features as their human tumor counterparts. These differences may reflect limitations of model system design, as tumors within these mouse classes are primarily driven by either the WAP or MMTV promoter. These differences in cell features, however, indicate that the trans-species associations observed previously [14] are possibly driven by the genetics of each mouse model. Nevertheless, broad molecular features are conserved between these human–murine counterparts [14]. Therefore, we propose that these mouse models retain significant preclinical utility provided that shared versus distinct molecular features are taken into account.

Neoadjuvant chemotherapy is a common approach for treating breast tumors, but only a relatively low percentage of patients have a pCR (~20 % overall). We tested the clinical significance of normal cellular features for predicting pCR using a combination of UVA and MVA logistic regression analyses. Human LumProg and mouse

fetal MaSC expression features were identified as predictive of pCR sensitivity across all breast cancer patients. More specifically, LumProg-HsEnriched-feature1 and fMaSC-MmEnriched-feature1 were highly expressed in basal-like tumors. This is consistent with the clinical observation that basal-like tumors have better neoadjuvant chemotherapy response rates since higher expression of these normal cell signatures was associated with a higher likelihood of pCR. Distinct from these signatures, tumors with high expression of fMaSC-MmEnriched-feature2 were more resistant to neoadjuvant chemotherapy. Not surprisingly, this signature was most highly expressed in luminal A and B tumors, consistent with the clinical observation that these subtypes have lower chemotherapy response rates. Importantly, these signatures remained significant even after controlling for intrinsic subtype, proliferation, and clinical variables in the MVA analysis; thus these normal cell signatures add information even when tumor subtype and clinical features are known. It is presently unknown whether tumors with these features arise from a LumProg or fetal MaSC cell-of-origin or acquire these features during tumorigenesis. Whether these features are acquired or inherent, the 'current' cellular traits of a tumor are likely most important as these appear to be a major determinant of chemotherapy sensitivity. The biological explanation for why LumProg and fetal MaSC expression features predict tumor responsiveness to neoadjuvant chemotherapy will need to be explored further, but it is likely linked to the common genetic features of TP53 loss [45], RB-pathway loss [46], and high proliferation status [47], as well as other inherent characteristics of these cellular states. This work highlights the efficacy of studying the normal mammary gland cell hierarchy and development to provide insights into human tumor therapy responsiveness.

Acknowledgments We would like to thank J.S. Parker, J.C. Harrell, and the Perou lab for helpful suggestions. This study was supported by funds from the following sources: NCI P50-CA58223 Breast SPORE program (CMP), RO1-CA138255 (CMP), RO1-CA148761 (CMP), Department of Defense W81XWH-12-1-0106 and W81XWH-12-1-0107 (CMP and GMW), NIEHS T32-ES007017-35S1 (ADP), and the Breast Cancer Research Foundation (CMP and GMW).

Conflict of interest C. M. P is an equity stock holder of BioClassifier LLC and University Genomics, and has filed a patent on the PAM50 subtyping assay.

References

1. Gjorevski N, Nelson CM (2011) Integrated morphodynamic signalling of the mammary gland. Nat Rev Mol Cell Biol 12:581–593
2. Van Keymeulen A, Rocha AS, Ousset M, Beck B, Bouvencourt G, Rock J, Sharma N, Dekoninck S, Blanpain C (2011) Distinct stem cells contribute to mammary gland development and maintenance. Nature 479:189–193
3. Santagata S, Thakkar A, Ergonul A, Wang B, Woo T, Hu R, Harrell JC, McNamara G, Schwede M, Culhane AC, Kindelberger D, Rodig S, Richardson A, Schnitt SJ, Tamimi RM, Ince TA (2014) Taxonomy of breast cancer based on normal cell phenotype predicts outcome. J Clin Invest 124:859–870
4. Visvader JE, Stingl J (2014) Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes Dev 28:1143–1158
5. Visvader JE (2009) Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. Genes Dev 23:2563–2577
6. Visvader JE (2011) Cells of origin in cancer. Nature 469:314–322
7. 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, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406:747–752
8. 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, Lonning PE, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS 98:10869–10874
9. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 12:R68
10. TCGA (2012) Comprehensive molecular portraits of human breast tumours. Nature 490:61–70
11. Chaffer CL, Brueckmann I, Scheel C, Kaestli AJ, Wiggins PA, Rodrigues LO, Brooks M, Reinhardt F, Su Y, Polyak K, Arendt LM, Kuperwasser C, Bierie B, Weinberg RA (2011) Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. PNAS 108:7950–7955
12. Spike BT, Wahl GM (2011) p53, stem cells, and reprogramming: tumor suppression beyond guarding the genome. Genes Cancer 2:404–419
13. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, Backlund MG, Yin Y, Khramtsov AI, Bastein R, Quackenbush J, Glazer RI, Brown PH, Green JE, Kopelovich L, Furth PA, Palazzo JP, Olopade OI, Bernard PS, Churchill GA, Van Dyke T, Perou CM (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol 8:R76
14. Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM (2013) Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol 14:R125
15. Usary J, Zhao W, Darr D, Roberts PJ, Liu M, Balletta L, Karginova O, Jordan J, Combest A, Bridges A, Prat A, Cheang MC, Herschkowitz JI, Rosen JM, Zamboni W, Sharpless NE, Perou CM (2013) Predicting drug responsiveness in human cancers

using genetically engineered mice. Clin Cancer Res 19: 4889–4899
16. Roberts PJ, Usary JE, Darr DB, Dillon PM, Pfefferle AD, Whittle MC, Duncan JS, Johnson SM, Combest AJ, Jin J, Zamboni WC, Johnson GL, Perou CM, Sharpless NE (2012) Combined PI3 K/mTOR and MEK inhibition provides broad antitumor activity in faithful murine cancer models. Clin Cancer Res 18:5290–5303
17. Bennett CN, Tomlinson CC, Michalowski AM, Chu IM, Luger D, Mittereder LR, Aprelikova O, Shou J, Piwinica-Worms H, Caplen NJ, Hollingshead MG, Green JE (2012) Cross-species genomic and functional analyses identify a combination therapy using a CHK1 inhibitor and a ribonucleotide reductase inhibitor to treat triple-negative breast cancer. Breast Cancer Res 14:R109
18. Roberts PJ, Bisi JE, Strum JC, Combest AJ, Darr DB, Usary JE, Zamboni WC, Wong KK, Perou CM, Sharpless NE (2012) Multiple roles of cyclin-dependent kinase 4/6 inhibitors in cancer therapy. J Natl Cancer Inst 104:476–487
19. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med 15:907–913
20. Shehata M, Teschendorff A, Sharp G, Novcic N, Russell A, Avril S, Prater M, Eirew P, Caldas C, Watson CJ, Stingl J (2012) Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. Breast Cancer Res 14:R134
21. Prat A, Karginova O, Parker JS, Fan C, He X, Bixby L, Harrell JC, Roman E, Adamo B, Troester M, Perou CM (2013) Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. Breast Cancer Res Treat 142:237–255
22. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE (2010) Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Res 12:R21
23. Spike BT, Engle DD, Lin JC, Cheung SK, La J, Wahl GM (2012) A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. Cell Stem Cell 10:183–197
24. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98:5116–5121
25. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM (2007) EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics 8:258
26. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS (2004) Adjustment of systematic microarray data biases. Bioinformatics 20:105–114
27. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster-analysis. J Comput Appl Math 20:53–65
28. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, Martin M, Cotrina J, Gomez H, Hubbard R, Chacon JI, Ferrer-Lozano J, Dyer R, Buxton M, Gong Y, Wu Y, Ibrahim N, Andreopoulou E, Ueno NT, Hunt K, Yang W, Nazario A, DeMichele A, O'Shaughnessy J, Hortobagyi GN, Symmans WF (2011) A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305:1873–1881
29. Miyake T, Nakayama T, Naoi Y, Yamamoto N, Otani Y, Kim SJ, Shimazu K, Shimomura A, Maruyama N, Tamaki Y, Noguchi S (2012) GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. Cancer Sci 103:913–920
30. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, Chan S, Welcher R, Liu D (2013) Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or paclitaxel in early-stage breast cancer. Clin Cancer Res 19:1587–1595
31. Harrell JC, Prat A, Parker JS, Fan C, He X, Carey L, Anders C, Ewend M, Perou CM (2012) Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse. Breast Cancer Res Treat 132:523–535
32. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486:346–352
33. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN, Cancer Genome Atlas Research N (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17:98–110
34. Gallahan D, Jhappan C, Robinson G, Hennighausen L, Sharp R, Kordon E, Callahan R, Merlino G, Smith GH (1996) Expression of a truncated Int3 gene in developing secretory mammary epithelium specifically retards lobular differentiation resulting in tumorigenesis. Cancer Res 56:1775–1785
35. Smith GH, Gallahan D, Diella F, Jhappan C, Merlino G, Callahan R (1995) Constitutive expression of a truncated INT3 gene in mouse mammary epithelium impairs differentiation and functional development. Cell Growth Differ 6:563–577
36. Chan SR, Vermi W, Luo J, Lucini L, Rickert C, Fowler AM, Lonardi S, Arthur C, Young LJ, Levy DE, Welch MJ, Cardiff RD, Schreiber RD (2012) STAT1-deficient mice spontaneously develop estrogen receptor alpha-positive luminal mammary carcinomas. Breast Cancer Res 14:R16
37. Adams JR, Xu K, Liu JC, Agamez NM, Loch AJ, Wong RG, Wang W, Wright KL, Lane TF, Zacksenhaus E, Egan SE (2011) Cooperation between Pik3ca and p53 mutations in mouse mammary tumor formation. Cancer Res 71:2706–2717
38. Li Z, Tognon CE, Godinho FJ, Yasaitis L, Hock H, Herschkowitz JI, Lannon CL, Cho E, Kim SJ, Bronson RT, Perou CM, Sorensen PH, Orkin SH (2007) ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. Cancer Cell 12:542–558
39. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27:1160–1167
40. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. Nature 501:328–337
41. Prater MD, Petit V, Russell IA, Giraddi RR, Shehata M, Menon S, Schulte R, Kalajzic I, Rath N, Olson MF, Metzger D, Faraldo MM, Deugnier MA, Glukhova MA, Sting J (2014) Mammary stem cells have myoepithelial cell properties. Nat Cell Biol 16:942–950

42. Chandriani S, Frengen E, Cowling VH, Pendergrass SA, Perou CM, Whitfield ML, Cole MD (2009) A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. PLoS One 4(8):e6693
43. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW, Hollier BG, Ram PT, Lander ES, Rosen JM, Weinberg RA, Mani SA (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. PNAS 107:15449–15454
44. Morel AP, Hinkal GW, Thomas C, Fauvet F, Courtois-Cox S, Wierinckx A, Devouassoux-Shisheboran M, Treilleux I, Tissier A, Gras B, Pourchet J, Puisieux I, Browne GJ, Spicer DB, Lachuer J, Ansieau S, Puisieux A (2012) EMT inducers catalyze malignant transformation of mammary epithelial cells and drive tumorigenesis towards claudin-low tumors in transgenic mice. PLoS Genet 8:e1002723
45. Gluck S, Ross JS, Royce M, McKenna EF Jr, Perou CM, Avisar E, Wu L (2012) TP53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine ± trastuzumab. Breast Cancer Res Treat 132:781–791
46. Herschkowitz JI, He X, Fan C, Perou CM (2008) The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. Breast Cancer Res 10:R75
47. Prat A, Lluch A, Albanell J, Barry WT, Fan C, Chacon JI, Parker JS, Calvo L, Plazaola A, Arcusa A, Segui-Palmer MA, Burgues O, Ribelles N, Rodriguez-Lescure A, Guerrero A, Ruiz-Borrego M, Munarriz B, Lopez JA, Adamo B, Cheang MC, Li Y, Hu Z, Gulley ML, Vidal MJ, Pitcher BN, Liu MC, Citron ML, Ellis MJ, Mardis E, Vickery T, Hudis CA, Winer EP, Carey LA, Caballero R, Carrasco E, Martin M, Perou CM, Alba E (2014) Predicting response and survival in chemotherapy-treated triple-negative breast cancer. Br J Cancer 111:1532–1541