scientific reports

Check for updates

OPEN

# Characterization of transcriptome diversity and in vitro behavior of primary human high-risk breast cells

Sahar J. Alothman[1,5], Keunsoo Kang[2,5], Xuefeng Liu[1,3,4,5], Ewa Krawczyk[3,4,5], Redha I. Azhar[1,5], Rong Hu[1,5], David Goerlitz[1,5], Bhaskar V. Kallakury[3,5] & Priscilla A. Furth[1,4,5]✉

**Biology and transcriptomes of non-cancerous human mammary epithelial cells at risk for breast cancer development were explored following primary isolation utilizing conditional reprogramming cell technology from mastectomy tissue ipsilateral to invasive breast cancer. Cultures demonstrated consistent categorizable behaviors. Relative viability and mammosphere formation differed between samples but were stable across three different mammary-specific media. E2F cell cycle target genes expression levels were positively correlated with viability and advancing age was inversely associated. Estrogen growth response was associated with Tissue necrosis factor signaling and Interferon alpha response gene enrichment. Neoadjuvant chemotherapy exposure significantly altered transcriptomes, shifting them towards expression of genes linked to mammary stem cell formation. Breast cancer prognostic signature sets include genes that in normal development are limited to specific stages of pregnancy or the menstrual cycle. Sample transcriptomes were queried for stage specific gene expression patterns. All cancer samples and a portion of high-risk samples showed overlapping stages reflective of abnormal gene expression patterns, while other high-risk samples exhibited more stage specific patterns. In conclusion, at-risk cells preserve behavioral and transcriptome diversity that could reflect different risk profiles. It is possible that prognostic platforms analogous to those used for breast cancer could be developed for high-risk mammary cells.**

Breast cancer is a heterogeneous disease with contributions from different pathophysiologic mechanisms including lifestyle, environment, physiological and genetic factors[1,2]. Late breast cancer recurrence estimates vary but range around 15% with hormone positive cancers being the most likely to recur[3]. A better understanding of underlying biology has been proposed as a means to improve individual risk calculation for development of secondary breast cancers[4]. Breast cancer exhibits field cancerization, meaning that normal-appearing breast cells ipsilateral to a known cancer are at higher risk than normal for development into a secondary breast cancer[5]. Risk factors predisposing an individual to breast cancer may also be related to recurrence of breast cancer. For example, normal developmental programs that go awry can contribute to breast cancer formation[6]. Expansion and differentiation of normal mammary stem cells with mammary gland transitions induced by hormonal cycling such as the menstrual cycle and pregnancy may also serve to generate cancer precursor stem cells[7–9]. Previously, menstrual cycle induced gene expression changes in women were defined from dissected mammary epithelium by bulk RNA sequencing[9] and single-cell RNA profiling used to develop a cellular blueprint of normal human breast epithelium[10]. Because mice and humans share mammary gland developmental gene expression patterns during pregnancy[11], transcriptional profiles developed using mice[8] can provide a comparator for changes in gene expression that are more difficult to characterize using human tissue. Reduction mammoplasty tissue has been used to define age associated changes in gene expression[12] and ductal lavage cells for transcriptome assessment of individuals at high versus normal breast cancer risk[13]. Several transcriptome-based breast cancer prognostic

[1]Department of Oncology, Georgetown University, 3970 Reservoir Rd NW, Washington, DC 20057, USA. [2]Department of Microbiology, College of Science and Technology, Dankook University, Cheonan 31116, Republic of Korea. [3]Department of Pathology, Georgetown University, Washington, DC 20057, USA. [4]Center for Cell Reprogramming, Georgetown University Medical Center, Washington, DC, USA. [5]Department of Medicine, Georgetown University, Washington, DC 20057, USA. ✉email: paf3@georgetown.edu

platforms are validated through clinical trials[14], but profiles for breast cancer risk prediction are not yet fully realized[3,4,15,16]. One goal of the present study was to assess diversity and pattern formation within transcriptional profiles of human high-risk mammary cells utilizing RNAseq.

A second goal was to address the biology of high-risk human mammary epithelial cells. A pillar of cancer-related biological investigations are cell culture models[17]. In this study we focused on primary cells, more immediately reflective of in vivo disease in people, to address our questions. While primary epithelial cells can be challenging to acquire, we utilized a time and cost-effective method, the epithelial-specific conditionally reprogrammed cells (CRC) technique, for efficient isolation. This technique enables maintenance of intact gene expression profiles through early passage[18–21]. It can be used in its original formulation with a fibroblast feeder layer or as CRC conditioned media ($CRC^{CM}$), enabling culture without the fibroblast feeder layer[20]. CRC-isolated mammary cultures have been reported to maintain expression of Estrogen Receptor 1 (*ESR1*) through early passage, a known challenge in studying the biology of mammary epithelial cells[22,23]. Because maintenance of *ESR1* expression in vitro is also facilitated by growth as 3D mammospheres[24], we chose matrix-free scaffold based nano-culture plates permitting both 2D monolayer and 3D sphere growth for a relatively high throughput analysis[25]. Cultures were moved into a secondary media lacking both phenol-red and serum for hormonal testing[26,16]. Three commercially available mammary-specific media are available, each developed for different purposes. Serum-free Mammary Epithelial Growth Medium (MEGM) (Lonza, Walkersville, MD, USA) maintains primary mammary epithelial cell heterogeneity through early passage. EpiCult™ and MammoCult™ Mammary Cell Culture Media (Stemcell Technologies, Vancouver, CA) support monolayer mammary precursor cell versus mammosphere growth, respectively[19,27]. Parallel replicative testing in each media was performed to assess within and between sample stability of viability and mammosphere formation in different medium to judge if viability and mammosphere formation were intrinsic to each sample or imposed by the medium. Because neither $CRC^{CM}$ or any of the mammary specific media had been evaluated in conjunction with the matrix-free scaffold-based nano-culture plates, this was considered an essential variable to appraise for interpretation of sample differences. Mycoplasma infection of the primary cultures was evaluated by both biochemical testing and screening of the RNAseq data for mycoplasma RNA sequences[28,29].

Transcriptional profiling contributes to molecular understanding and provides leads for mechanistic investigations[30]. Because our study combined transcriptional profiling with biological behavior in primary cell culture, we utilized Gene Set Enrichment Analysis (GSEA)[31] and the Molecular Signatures Database[32,33] for associations and patterns between gene expression and behavior. Previous research shows that individual gene expression patterns can be linked to specific cell behaviors including viability[34–39] and estrogen response[40,41].

## Results

**Relative viability and mammosphere formation were preserved properties across different media conditions inherent to individual samples.** Ipsilateral high-risk CRC-viable cultures were derived from women with a range of breast cancer types. Advancing age significantly reduced likelihood of initial CRC isolation with no viable samples isolated from the six women ≥ age 70 years (Table 1, Fig. 1a). No effect from exposure to prior neoadjuvant chemotherapy was found. Eight matched invasive tumor samples available were added to the study for comparison. Secondary passage in MEGM was used to expand cultures in serum- and phenol-red-free media in preparation for hormonal response testing. Age did not impact likelihood of growth in MEGM but showed an inverse relationship with measured viability scores (Fig. 1b,c). Because a range of viability was evident in MEGM, two additional mammary-specific media (EpiC and MammoC) along with CRC conditioned media (CM) were tested to determine the extent of media-specific behaviors. Although media differences within samples were identified, viability differences between samples were largely preserved across media (Fig. 1d). Viability was not significantly higher in $CRC^{CM}$ even though this modeled the initial isolation media, albeit without the feeder cell layer. Extent of mammosphere formation was similarly assessed in different media. Samples showed generally consistent behavior across media even with some within sample differences being present (Fig. 1e). A few samples were available for testing across passage. This showed variability between samples was preserved across passage (Suppl. Figure 1a, b). Because the matrix-free scaffold-based nano-culture plates permit both 2D and 3D growth, cell growth patterns were analyzed in each media. The majority of samples showed mixed mammosphere-monolayer cell growth across all media (Suppl. Figure 1c). Only two samples exhibited mammosphere-only growth across all tested media. Phenol-red and serum-free MEGM supported mammosphere growth in all samples. Based on these results, variability in behavior in culture was assessed as a preserved biological property inherent to each sample.

**Relative expression levels of cell cycle related target genes of E2F transcription factors paralleled relative viability.** Because cell viability was a preserved factor in samples, we asked if expression of genes positively regulating cell proliferation would correlate with relative viability. The HALLMARK_E2F_TARGETS gene set includes cell cycle related target genes of E2F transcription factors. While the highest viability samples showed relative higher expression of many E2F target genes, some samples demonstrated a more selective pattern (Fig. 2). Examples of genes that were recurrently expressed at higher levels in different viable samples included Aurora Kinase A *(AURKA)*, Kinesin Family Member 4A (*KIF4A*) Stromal Antigen 1 (*STAG1*), CCCTC-Binding Factor (*CTCF*), Replication Protein A1 (RPA1), and Tissue Necrosis Factor (TNF). One sample showed higher expression of Serpin Family B Member 2 (*SERPINB2*). This analysis showed that differences in sample viability on secondary passage were linked to transcriptome differences present in the cells upon their initial isolation in CRC.

| Sample code | Cell culture number | Sample type | Breast cancer type | ER | PR | HER2 | Elston score | pTNM stage | Sex | Age | *BRCA* mutation status | Second sample from same patient |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IPSI1 | 1067 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 90% | 2+ | 6 | pT2pN1c | F | 35 | No test | T1 |
| IPSI2 | 1028 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 10% | 2+ | 9 | pT3pN3 | F | 66 | Tested negative | T2 |
| IPSI3 | 987 | Non-Cancer Ipsilateral to IDC | IDC | 20% | 0 | negative | 8 | pT1cpN0 | F | 67 | No test | T3 |
| IPSI4 | 1015 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 70% | 1+ | 7,8 | pT2apN2a | F | 60 | No test | T4 |
| IPSI5 | 1018 | Non-Cancer Ipsilateral to IDC | IDC | 50% | 10% | FISH amplified | 8 | pT1c pN0 | F | 42 | Tested negative | |
| IPSI6 | 1111 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 0 | negative | 9 | pT2pN0 | F | 55 | *BRCA1* mutation | |
| IPSI7 | 1037 | Non-Cancer Ipsilateral toILC | ILC | 90% | 0 | 1+ | 5 | pT1cpN2a | F | 63 | No test | |
| IPSI8 | 1020 | Non-Cancer Ipsilateral to IDC | IDC | 0 | 0 | 3+FISH positive | II of III | ypTxpN1 | F | 41 | Tested negative | |
| IPSI9 | 1164 | Non-Cancer Ipsilateral to TNBC | TNBC | 0 | 0 | negative | 9 | ypT2pN1 | F | 44 | No test | |
| IPSI10 | 978 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 40% | 2+equivocal FISH | 9 | ypT1cpN0(sn) | F | 39 | Tested negative | T10 |
| IPSI11 | 973 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 80% | negative | 6 | pT1pN0 | F | 49 | No test | |
| IPSI12 | 1167 | Non-Cancer Ipsilateral to IBC | IBC | 90% | 30% | 1+ | 9 | mypT4d pN3a | F | 64 | No test | |
| IPSI13 | 1077 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 80% | 2+FISH negative | 8 | pT1cpN1 | F | 41 | *BRCA2* mutation | T13 |
| IPSI14 | 1074 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 90% | 3+ | 9 | pT1cpN1 | F | 43 | No test | |
| IPSI15 | 1000 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 30% | 3+ | 7 | ypT0pN1(mi) | F | 43 | Tested negative | |
| IPSI16 | 957 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 80% | 1+ | 5 | pT2pN0 | F | 45 | Tested negative | T16 |
| IPSI17 | 1112 | Non-Cancer Ipsilateral to IDC | IDC | 80% | 10% | 1+ | 5 | pT1bpN0 | F | 55 | Tested negative | |
| IPSI18 | 1191 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 10% | 3+ | 8 | pT2pN0 | F | 42 | Tested negative | |
| IPSI19 | 1092 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 90% | 3+ | 6 | yT2pN1 | F | 52 | Tested negative | |
| IPSI20 | 971 | Non-Cancer Ipsilateral to IDC | IDC | 90% | 80% | negative | 6 | pT1pN0 | F | 49 | No test | |
| IPSI21 | 1120 | Non-Cancer Ipsilateral to IDC | IDC | 50% | 60% | 3+ | 9 | ypT0pN1 | F | 34 | Tested negative | |
| IPSI22 | 1136 | Non-Cancer Ipsilateral to IDC | IDC | 0 | 0 | 3+ | 9 | pT1cpN0 | F | 32 | Tested negative | |
| IPSI23 | 1219 | Non-Cancer Ipsilateral to IDC | IDC | na | na | na | na | yTxpNx | F | 31 | *BRCA1* mutation | |

Continued

| Sample code | Cell culture number | Sample type | Breast cancer type | ER | PR | HER2 | Elston score | pTNM stage | Sex | Age | *BRCA* mutation status | Second sample from same patient |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IPSI24 | 1300 | Non-Cancer Ipsilateral to ILC | ILC | 95–100% | 95–100% | 1+ | Notting-ham 6 | mpT1bpN0(i-)(sn) | F | 42 | Tested negative | |
| IPSI25 | 549 | Non-Cancer Ipsilateral to IDC | IDC | 95% | 70% | 1+ | 6 | pT1bpN0 | F | 46 | *BRCA* mutation | |
| T1 | 1068 | Invasive Breast Cancer | IDC | 90% | 90% | 2+ | 6 | pT2pN1c | F | 35 | No test | IPSI1 |
| T2 | 1029 | Invasive Breast Cancer | IDC | 90% | 10% | 2+ | 9 | pT3pN3 | F | 66 | Tested negative | IPSI2 |
| T3 | 988 | Invasive Breast Cancer | IDC | 20% | 0 | negative | 8 | pT1cpN0 | F | 67 | No test | IPSI3 |
| T4 | 1016 | Invasive Breast Cancer | IDC | 80% | 70% | 1+ | 7,8 | pT2apN2a | F | 60 | No test | IPSI4 |
| T5 | 1013 | Invasive Breast Cancer | IDC | 90% | 70% | 1+ | 6 | pT1cpN0 | F | 61 | No test | |
| T10 | 980 | Invasive Breast Cancer | IDC | 80% | 40% | 2+equivocal FISH | 9 | ypT1cpN0(sn) | F | 39 | Tested negative | IPSI10 |
| T13 | 1078 | Invasive Breast Cancer | IDC | 80% | 80% | 2+FISH negative | 8 | pT1cpN1 | F | 41 | *BRCA2* mutation | IPSI13 |
| T16 | 958 | Invasive Breast Cancer | IDC | 80% | 80% | 1+ | 5 | pT2pN0 | F | 45 | Tested negative | IPSI16 |

**Table 1.** Table of samples.

**Hormonal responsiveness correlated with transcriptomic profile.** To determine if hormonal response would also show a link to the transcriptome, samples were subjected to relative viability measurements in the absence and presence of 17 beta-estradiol (E2) (10 nM) and the ER antagonist 4-hydroxy-tamoxifen (4-OHT) (1 μM) for samples with sufficient numbers of cells for concurrent testing. Of the 18 samples available for concurrent testing, eight showed statistically significant higher viability in E2 (Fig. 3a). Of the nine samples available for concurrent testing in 4-OHT, viability was unchanged in six. Three samples showed higher viability than in the control condition but each had a different pattern. There were no consistent differences between E2 responsive and non-responsive samples in patterns of mammosphere growth (Fig. 3b). One hundred and ninety-four genes were expressed at significantly higher levels and one hundred and forty-eight genes at significantly lower levels in the E2 responsive samples (Fig. 3c). Genes with at least two-fold differences were analyzed by GSEA for significant gene set enrichment. The HALLMARK_TNFA_SIGNALING_VIA_NFKB and HALLMARK_INTERFERON_ALPHA_RESPONSE gene sets were enriched in the E2 responsive samples (Fig. 3d) while HALLMARK_OXIDATIVE_PHOSPHORYLATION, HALLMARK_P53_PATHWAY and HALLMARK_CHOLESTEROL_HOMEOSTASIS gene sets were enriched in the E2 non-responsive samples (Fig. 3e). Differences in expression levels of specific genes within these enriched gene sets between samples responsive and non-responsive to E2 are illustrated using log scales (Fig. 3f,g). In summary, transcriptome differences between samples associated with absence or presence of increased viability in response to E2.

**Genes linked to mammary stem cell formation were expressed at higher levels in samples from women previously exposed to neoadjuvant chemotherapy.** To explore the hypothesis that exposure to neoadjuvant chemotherapy might alter gene expression in the ipsilateral non-cancer samples, we tested the number and character of differentially expressed genes (DEGs) in all ipsilateral samples available from individuals with Erb-B2 Receptor Tyrosine Kinase 2/HER2 positive (HER2+) breast cancer, half of whom had received neoadjuvant chemotherapy. The analysis was limited to this subgroup because this was the only breast cancer subtype that was consistently associated with neoadjuvant chemotherapy administration in our sample set. Limiting the analysis to samples from individuals all of whom had HER2+ invasive breast cancer enabled us to match for any theoretical differences in the ipsilateral high-risk cells that might be associated with development of HER2+ breast cancer. GSEA analyses was performed to identify C2 gene sets enriched in samples from individuals exposed to neoadjuvant chemotherapy and gene sets enriched in samples from individuals not-exposed to chemotherapy. Significantly, the 328 up-regulated genes from samples exposed to neoadjuvant chemotherapy showed enrichment of the LIM_MAMMARY_STEM_CELL_UP gene set while the 189 down-regulated genes, identified due to their higher expression in the absence of neoadjuvant chemotherapy exposure, showed enrichment of the LIM_MAMMARY_STEM_CELL_DOWN gene set (Fig. 4a–c). Relative expression levels of individual mammary stem cell related genes differentially expressed in samples from individuals with and without exposure to neoadjuvant chemotherapy from these two gene sets are shown using a heat map (Fig. 4d). Identification of a pattern of differentially expressed genes in the neoadjuvant-exposed samples that includes genes that are consistently up- and down-regulated in mammary stem cells suggests that mammary stem cell populations could be enriched in the ipsilateral breast tissue of individuals exposed to neoadjuvant chemotherapy.

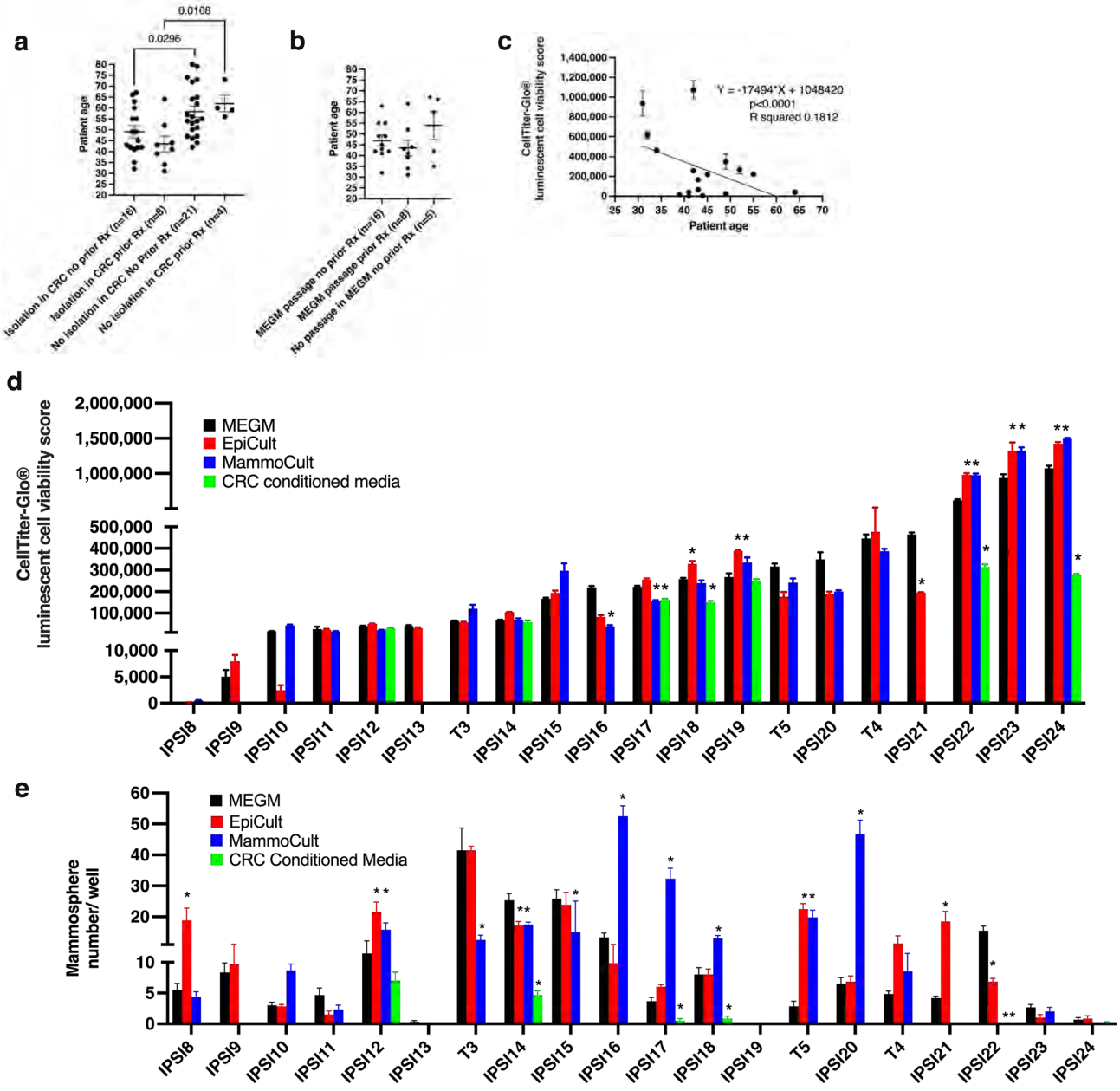



**Figure 1.** Viability and mammosphere formation in different media. (**a**) Scatter plot illustrating age distribution of samples isolated and not isolated in CRC with and without prior neoadjuvant therapy (Rx). Sample numbers, mean and standard error of the mean (SEM), and *p* values < 0.05 shown. Ordinary one-way ANOVA, $p = 0.0035$ F = 5.218, Brown-Forsythe test F = 0.6113, DFn = 3 DFd = 45, Sidak's multiple comparisons test padj = 0.0296 [Isolation in CRC No prior Rx n = 20, No isolation in CRC No prior Rx n = 21], padj = 0.0168 [Isolation in CRC Prior Rx n = 7, No isolation in CRC Prior Rx n = 4]. (**b**) Scatter plot illustrating age distribution of samples passaged and not passaged in MEGM with and without prior neoadjuvant therapy (Rx). Sample numbers and mean and SEM shown. (**c**) Regression scatter plot: patient age (years) and viability measured in MEGM. n = 6 replicates. Regression equation, *p* value and R squared shown. (**d**) Bar graphs presenting cell viability of primary cells measured seven days following plating in different media (n = 6 replicates/condition with exceptions IPSI9 n = 3 replicates/EpiCult, IPSI24 n = 5 replicates/CRC Conditioned Media). (**e**) Bar graphs presenting numbers of mammospheres formed counted seven days following plating in different media (n = 6 replicates/condition with exceptions IPSI9 n = 3 replicates/EpiCult, IPSI24 n = 5 replicates/CRC Conditioned Media). $^*p < 0.05$. 2way ANOVAs demonstrated an interactive effect between medium and samples accounting for 17.03% of the variance ($p < 0.0001$ F = 193.21 DFn = 21 DFd156) in viability for samples tested in MEGM, EpiC, MammoC and CRC Conditioned Media with sample accounting for 66.54% ($p < 0.0001$ F = 2264.44 DFn = 7 DFd = 156) and medium 12.33% ($p < 0.0001$ F = 979.30 DFn = 3 DFd = 156) with 7.535% of the variance for samples tested in MEGM, EpiC and MammoC due to an interactive effect ($p < 0.0001$ F = 6.43 DFn = 26 DFd = 204) with sample accounting for 82.71% of the variance ($p < 0.0001$ F = 141.05 DFn = 13 DFd = 204) and medium 0.5213% ($p < 0.0036$ F = 5.78 DFn = 2 DFd = 204). 2way ANOVAs showed an interactive effect accounting for 32.94% of variance ($p < 0.0001$ F = 21.85 DFn = 21 DFd = 157) in mammosphere numbers for samples tested in MEGM, EpiC, MammoC and CRC Conditioned Media with sample accounting for 35.11% ($p < 0.0001$ F = 69.89 DFn = 7 DFd = 157) and medium 20.51% ($p < 0.001$ F = 95.28 DFn = 3 DFd = 157) with 34.01% of the samples tested in MEGM, EpiC and MammoC due to an interactive effect ($p < 0.001$ F = 14.67 DFn = 24 DFd = 195) with sample variance accounting for 44.58% of the variance ($p < 0.0001$ F = 38.46 Dfn = 12 Dfd = 195) and medium 2.581% ($p < 0.0001$ F = 13.36 DFn = 2 DFd = 195). Cell viability measured utilizing CellTiter-Glo® 3D with relative viability expressed as CellTiter-Glo luminescent cell viability score. Color coding: Black: Phenol red-free MEGM™. Red: EpiCult™. Blue: MammoCult™. Green: CRC Conditioned Media. *IPSI* ipsilateral non-cancer, *T* tumor cancer.

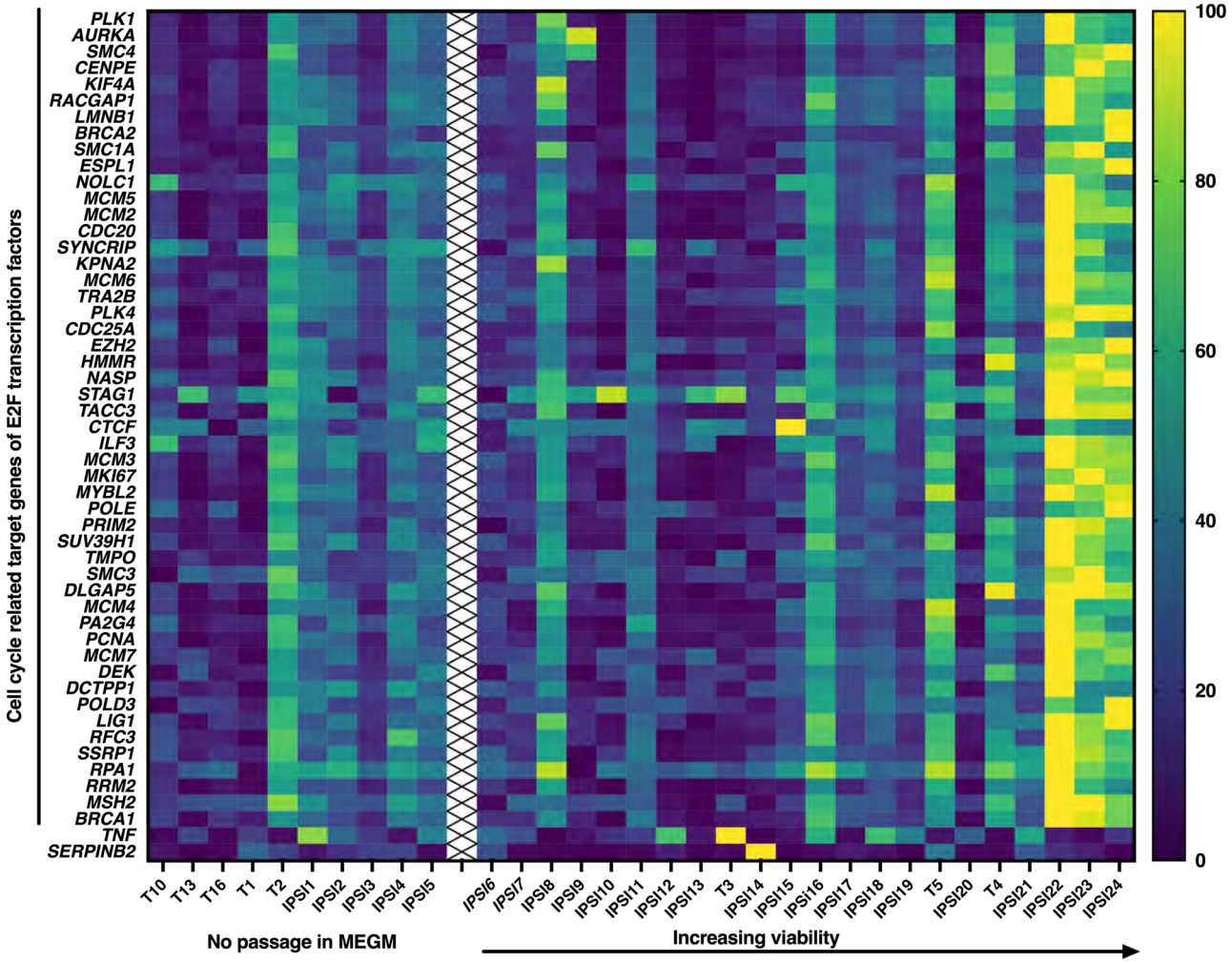



**Figure 2.** Relative expression levels of cell cycle related target genes of E2F transcription factors. Relative expression levels of cell cycle related target genes of E2F transcription factors (HALLMARK_E2F_TARGETS) are shown with samples arrayed left to right according to increasing viability. Italicized samples (*IPSI6, IPSI7, IPSI8*) were mycoplasma negative on sequence evaluation and grew in MEGM but had no material for biochemical mycoplasma testing and were therefore excluded from presentation of viability and mammosphere numbers. *IPSI* ipsilateral non-cancer, *T* tumor cancer. Color coding: Dark blue to yellow with increasing expression.

**Both ipsilateral non-cancer and cancer samples demonstrated aberrant patterns of pregnancy-related gene transcription.** Pregnancy-associated mammary gland development is characterized by stages of stereotypic gene expression changes regulating normal mammary cell proliferation and differentiation. Abnormal expression patterns of these genes could underlie unregulated cell proliferation associated with breast cancer. To explore this question, sample transcriptomes were queried for adherence to normal pregnancy-associated gene expression patterns. A heat map illustrating changes in gene expression during four different stages of pregnancy (A–D) was developed utilizing mouse transcriptomic data with genes that are members of established breast cancer prognostic profiles indicated by asterisks (Fig. 5a). Relative expression patterns of pregnancy-associated genes in cancer samples were similarly arrayed in a heat map and sorted by expression pattern to determine if there were any links to patterns during pregnancy. Links were found, but patterns invariably overlapped different pregnancy stages, consistent with the notion that cancer cells demonstrate abnormal patterns of genes that normally are not expressed together (Fig. 5b). To address the question of whether or not the limited number of pregnancy-associated genes that are members of an established breast cancer prognostic platform are associated together in established pathways, they were subjected to GSEA analysis (Fig. 5c). This demonstrated that they show significant enrichment in cancer pathways, including breast cancer specific ones, illustrating the potential significance of up-regulated expression in high-risk samples. Finally, relative expression patterns of pregnancy-associated genes in the high-risk samples were arrayed in a heat map and sorted by expression patterns (Fig. 5d). Expression patterns arrayed differently than the cancer samples but still showed links to different pregnancy stages. One-third of the samples showed an overlapping pattern of pregnancy-associated gene expression (A/B/D), similar to that seen in some of the cancer samples. IPSI10 and T10 samples from the same individual showed the same pattern but other cancer samples showed divergent patterns from

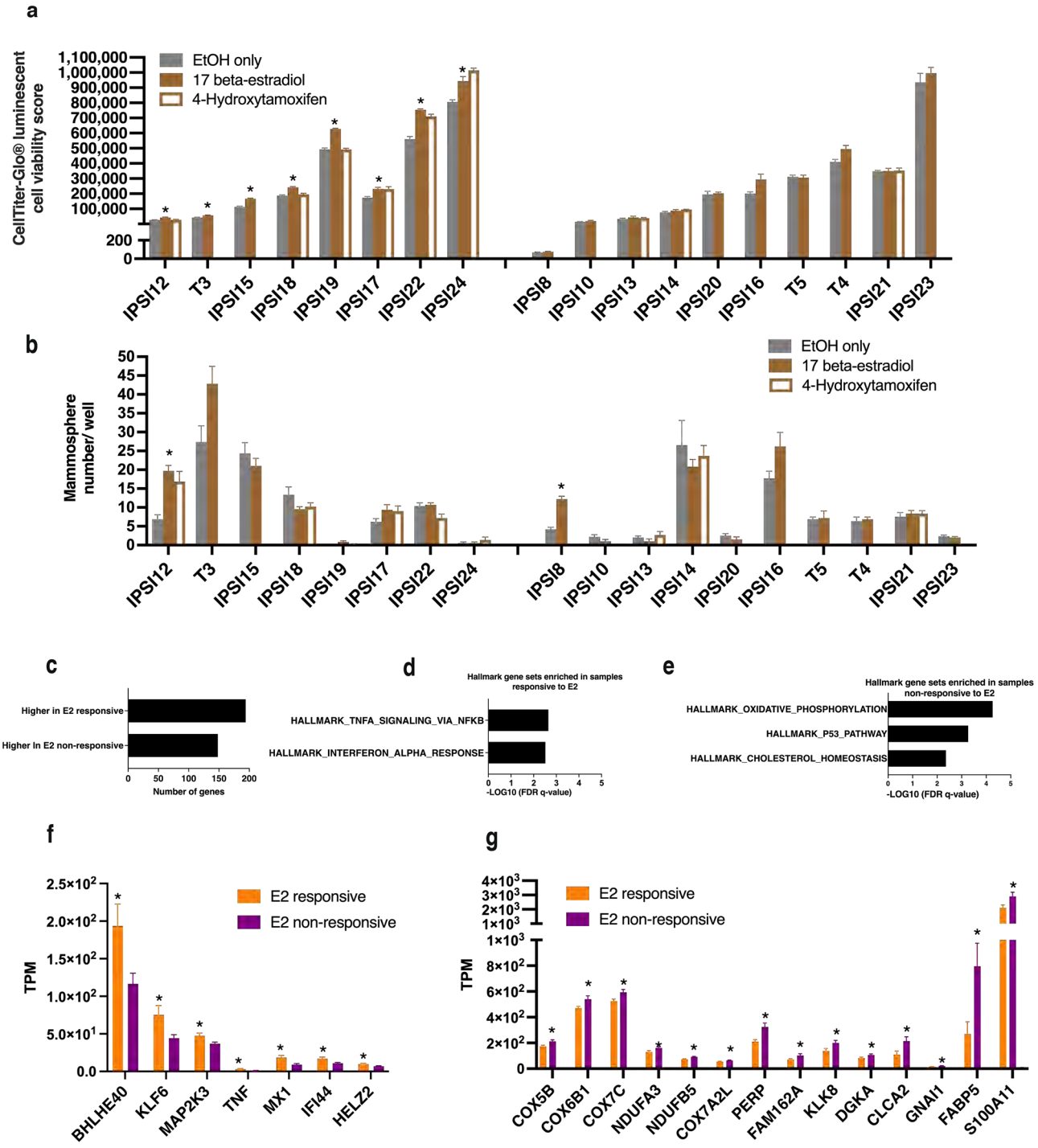
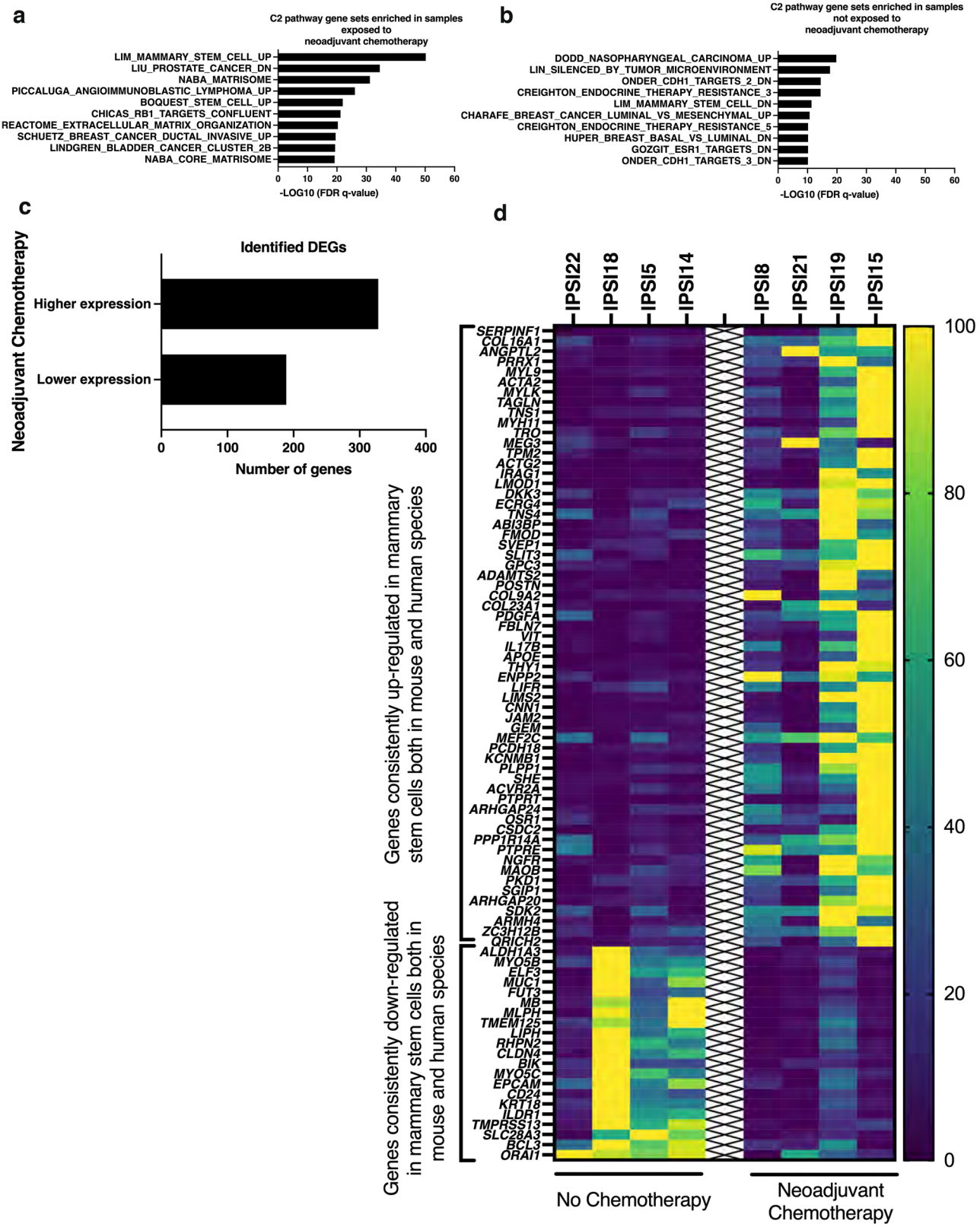



**Figure 3.** Comparative viability and mammosphere formation following hormonal exposure. (**a**) Bar graph illustrating relative viability following seven-day exposure to 17 beta-estradiol (E2,10 nM), 4-Hydroxytamoxifen (4-OHT, Tamoxifen) (1 μM) or EtOH (n = 6 replicates/condition with exceptions T3 (n = 5/E2), IPSI14 (n = 5/E2), IPSI12(n = 4/4-OHT). (**b**) Bar graph illustrating mammosphere numbers/well following seven-day exposure to 17 beta-estradiol (E2,10 nM), 4-Hydroxytamoxifen (4-OHT, Tamoxifen) (1 μM) or EtOH (n = 6 replicates/condition with exceptions T3 (n = 5/E2), IPSI14 (n = 5/E2), IPSI12(n = 4/4-OHT). (**c**) Bar graph illustrating numbers of genes expressed at significantly higher or lower levels in samples responsive to E2 as compared to samples non-responsive to E2. *q ≤ 0.004, unpaired t tests with Welch correction, Variance assumption: Individual variance for each group, Multiple comparisons: False Discovery Rate, Two-stage step-up (Benjamini, Krieger, and Yekutieli), Desired FDR (Q):1.00%. (**d**) Bar graph showing significantly enriched HALLMARK gene sets for genes expressed at higher levels in E2 respoonsive samples. (**e**) Bar graph showing significantly enriched HALLMARK gene sets for genes expressed at lower levels in E2 responsive samples. (**f**) Bar graph illustrating sample-specific expression levels (TPM) of genes included in HALLMARK_TNF_SIGNALING_VIA_NFKB and HALLMARK_INTERFERON_ALPHA_RESPONSE gene sets. Mean and SEM indicated. (**g**) Bar graph illustrating sample-specific expression levels (TPM) of genes included in HALLMARK_OXIDATIVE_PHOSPHORYLATION, HALLMARK_P53_PATHWAY, and HALLMARK_CHOLESTEROL_HOMEOSTASIS gene sets. Mean and SEM indicated. Cell viability measured utilizing CellTiter-Glo® 3D with relative viability expressed as CellTiter-Glo luminescent cell viability score. *EtOH* Ethanol, *E2* 17 beta-estradiol, *4-OHT* 4-Hydroxytamoxifen, *TPM* transcripts per million, *FDR* false discovery rate, *IPSI* ipsilateral non-cancer, *T* tumor cancer. Color coding: Gray: EtOH. Solid brown: E2. Brown outline: 4-OHT. Orange: E2 responsive. Purple: E2 non-responsive.


a
C2 pathway gene sets enriched in samples exposed to neoadjuvant chemotherapy
LIM_MAMMARY_STEM_CELL_UP
LIU_PROSTATE_CANCER_DN
NABA_MATRISOME
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_UP
BOQUEST_STEM_CELL_UP
CHICAS_RB1_TARGETS_CONFLUENT
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP
LINDGREN_BLADDER_CANCER_CLUSTER_2B
NABA_CORE_MATRISOME
0 10 20 30 40 50 60
-LOG10 (FDR q-value)
b
C2 pathway gene sets enriched in samples not exposed to neoadjuvant chemotherapy
DODD_NASOPHARYNGEAL_CARCINOMA_UP
LIN_SILENCED_BY_TUMOR_MICROENVIRONMENT
ONDER_CDH1_TARGETS_2_DN
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_3
LIM_MAMMARY_STEM_CELL_DN
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5
HUPER_BREAST_BASAL_VS_LUMINAL_DN
GOZGIT_ESR1_TARGETS_DN
ONDER_CDH1_TARGETS_3_DN
0 10 20 30 40 50 60
-LOG10 (FDR q-value)
c
Neoadjuvant Chemotherapy
Identified DEGs
Higher expression
Lower expression
0 100 200 300 400
Number of genes
d
IPSI22 IPSI18 IPSI5 IPSI14 IPSI8 IPSI21 IPSI19 IPSI15
SERPINF1 COL16A1 ANGPTL2 PRRX1 MYL9 ACTA2 MYLK TAGLN TNS1 MYH11 TRO MEG3 TPM2 ACTG2 IRAG1 LMOD1 DKK3 ECRG4 TNS4 ABI3BP FMOD SVEP1 SLIT3 GPC3 ADAMTS2 POSTN COL9A2 COL23A1 PDGFA FBLN7 VIT IL17B APOE THY1 ENPP2 LIFR LIMS2 CNN1 JAM2 GEM MEF2C PCDH18 KCNMB1 PLPP1 SHE ACVR2A PTPRT ARHGAP24 OSR1 CSDC2 PPP1R14A PTPRE NGFR MAOB PKD1 SGIP1 ARHGAP20 SDK2 ARMH4 ZC3H12B ORICH2
ALDH1A3 MYO5B ELF3 MUC1 FUT3 MB MLPH TMEM125 LIPH RHPN2 CLDN4 BIK MYO5C EPCAM CD24 KRT18 ILDR1 TMPRSS13 SLC28A3 BCL3 ORAI1
Genes consistently up-regulated in mammary stem cells both in mouse and human species
Genes consistently down-regulated in mammary stem cells both in mouse and human species
100
80
60
40
20
0
No Chemotherapy
Neoadjuvant Chemotherapy

◂**Figure 4.** Neoadjuvant chemotherapy exposure associated with gene expression changes linked to mammary stem cell formation in ipsilateral high-risk non-cancer samples. (**a**) Bar graph presenting the top ten C2 gene sets with the lowest significant FDR q-values identified from the MSigDB Collection utilizing genes expressed at significantly higher levels in samples from individuals exposed to neoadjuvant chemotherapy. (**b**) Bar graph presenting the top ten C2 gene sets with the lowest significant FDR q-values identified from the MsigDB Collection utilizing genes expressed at significantly lower levels in samples from individuals exposed to neoadjuvant chemotherapy. (**c**) Bar graph indicating numbers of up-regulated and down-regulated DEGs identified in non-cancer ipsilateral samples from women with ER+/HER2+or HER2+ cancer that were exposed or not exposed to neoadjuvant chemotherapy. (**d**) Heat map illustrating relative expression levels of identified DEGs enriched in LIM_MAMMMARY_STEM_CELL_UP (Genes consistently up-regulated in mammary stem cells both in mouse and human species) and LIM_MAMMMARY_STEM_CELL_DOWN (Genes consistently down-regulated in mammary stem cells both in mouse and human species) in samples from individuals with no chemotherapy exposure and those with neoadjuvant chemotherapy exposure. *DEG* differentially expressed gene at Padj ≤ 0.05, *FDR* false discovery rate, *MsigDB* Molecular Signatures Database v7.5.1, *C2* Curated gene sets. N = 4 no neoadjuvant chemotherapy (aged 40 ± 3 years, mean ± SEM), N = 4 neoadjuvant chemotherapy (aged 42 ± 4 years). Color coding: Dark blue to yellow with increasing expression.

their ipsilateral counterparts. A portion of samples more closely resembled single pregnancy stage profiles. In summary, a portion of the high-risk cells studied exhibited expression profiles similar to a pregnancy profile.

## Discussion

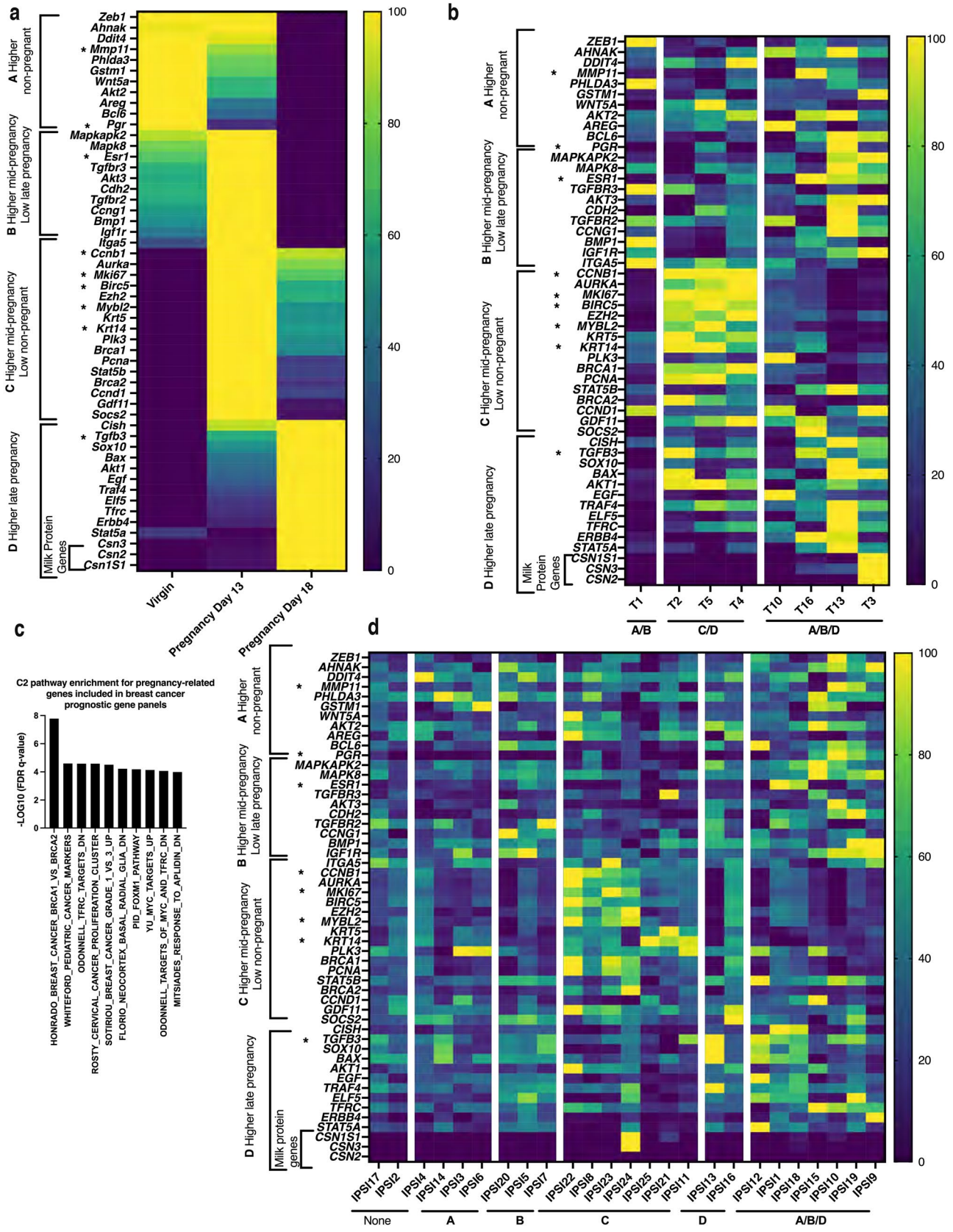
The overall goal of the study was to study the biology and transcriptional profiles of human mammary epithelial cells at higher than normal risk for breast cancer development. Because women who develop one breast cancer are at risk for secondary breast cancer development, this was approached through isolation of primary mammary epithelial cells from non-cancerous breast tissue obtained from mastectomies performed for treatment of invasive breast cancer. The conditionally reprogrammed cells (CRC) technique was utilized for initial isolation for its efficiency in isolating primary epithelial cells. Transcriptomic studies were performed using cells from this initial isolation condition. Biological studies were performed in mammary-specific media to maintain appropriate differentiation status and provide a suitable environment for testing hormonal response. All studies were performed using low passage number cells to minimize changes that can come with extended passage.

There is reason to consider that the sample diversity identified in this study could be reflective of different risk probabilities for future cancer development. Inspection of the data provides different behavioral and transcriptional profiles for individual samples. This can range significantly. For examples, some samples showed the combination of high viability and E2 responsiveness with a transcriptome including higher E2F target and proliferative mid-pregnancy gene expression that included known breast cancer prognostic genes. Other samples were equivalently E2 responsive but showed an overlapping pattern of different pregnancy stage gene expression. Others showed low viability with or without E2 responsiveness. Without long-term clinical follow-up one cannot define if these different patterns are predictive of different outcomes but they are measurable differences between samples. Only subsets of women develop either primary or secondary breast cancer. The approach utilized here provides additional understanding of at-risk mammary epithelial cell biology that could be further exploited for improvement of secondary breast cancer risk profiling[3,4].

The study demonstrated associations between behavior and gene expression. Retention of an estrogen growth response correlated with expression of genes enriched in TNF alpha and Interferon alpha signaling pathways. Both of these pathways are established contributors to estrogen signaling[40,41]. Relatively high TNF alpha expression was found in one of the viable samples that lacked high expression of E2F cell cycle target genes. It is not uncommon for cancer cells to rely on a specific growth factors such as TNF alpha for viability[36]. Cancer growth can also be reliant on high expression of specific E2 cell cycle target genes including *AURKA*, *KIF4A*, *STAG1*, *CTCF*, and RPA1, all of which were identified highly expressed in at least one of the at-risk samples studied here[34–39]. One sample with low overall E2F target genes expression demonstrated high levels of *SERPINB* expression, which is associated with maintenance of cell viability[37]. These findings can be interpreted as evidence that high risk cells are poised for future cancer development due to transcriptional changes that favor cell growth and survival. However, whether or not the high-risk cultures could be contaminated with cancer cells has to be considered. This possibility cannot be absolutely excluded but the experimental design included all possible controls for this variable. Ipsilateral samples were taken from breast geographically distinct from the invasive tumor and assessed as non-cancerous on gross examination with mirrored tissue samples from all submitted samples examined histologically for any microscopic evidence of cancer cells. Contaminated samples were excluded from the study.

One can speculate on whether or not the in vitro differences in E2 and 4-OHT response would reflect in vivo response. Anti-hormonals are the mainstay for prevention of secondary cancers for individuals with ER+ cancer, but there is variability in response and resistance can develop[42,43]. Whether or not prospective in vitro testing of hormonal response would ever be clinically useful is an open question. The overall goal for this study, to test a possible platform for evaluating hormonal response of primary mammary epithelial cells and evaluate transcriptome associations was completed. Replication of these results with associated clinical follow-up of anti-hormonal response would be a possible next step.

Neoadjuvant chemotherapy exposure was associated with enrichment of normal mammary gland stem cell gene sets in the high-risk cells[11]. Specific genes relevant to mammary stem cells that were found up-regulated include leukemia inhibitory factor receptor (*LIFR*), reported as promoting breast cancer stem cell renewal[44] and Thy-1 Cell Surface Antigen (*Thy1*), a gene associated with serial transplantation of mammary epithelial cells[45].


a
A Higher non-pregnant
B Higher mid-pregnancy Low late pregnancy
C Higher mid-pregnancy Low non-pregnant
D Higher late pregnancy
Milk Protein Genes
Zeb1
Ahnak
Ddit4
* Mmp11
Phlda3
Gstm1
Wnt5a
Akt2
Areg
Bcl6
* Pgr
Mapkapk2
Mapk8
* Esr1
Tgfbr3
Akt3
Cdh2
Tgfbr2
Ccng1
Bmp1
Igf1r
Itga5
* Ccnb1
Aurka
* Mki67
* Birc5
Ezh2
* Mybl2
Krt5
* Krt14
Plk3
Brca1
Pcna
Stat5b
Brca2
Ccnd1
Gdf11
Socs2
Cish
* Tgfb3
Sox10
Bax
Akt1
Egf
Traf4
Elf5
Tfrc
Erbb4
Stat5a
Csn3
Csn2
Csn1S1
100
80
60
40
20
0
Virgin
Pregnancy Day 13
Pregnancy Day 18
b
A Higher non-pregnant
B Higher mid-pregnancy Low late pregnancy
C Higher mid-pregnancy Low non-pregnant
D Higher late pregnancy
Milk Protein Genes
ZEB1
AHNAK
DDIT4
* MMP11
PHLDA3
GSTM1
WNT5A
AKT2
AREG
BCL6
* PGR
MAPKAPK2
MAPK8
* ESR1
TGFBR3
AKT3
CDH2
TGFBR2
CCNG1
BMP1
IGF1R
ITGA5
* * CCNB1
AURKA
* * MKI67
* * BIRC5
EZH2
* * MYBL2
KRT5
* KRT14
PLK3
BRCA1
PCNA
STAT5B
BRCA2
CCND1
GDF11
SOCS2
CISH
* TGFB3
SOX10
BAX
AKT1
EGF
TRAF4
ELF5
TFRC
ERBB4
STAT5A
CSN1S1
CSN3
CSN2
100
80
60
40
20
0
T1
T2
T5
T4
T10
T16
T13
T3
A/B
C/D
A/B/D
c
C2 pathway enrichment for pregnancy-related genes included in breast cancer prognostic gene panels
-LOG10 (FDR q-value)
8
6
4
2
0
HONRADO_BREAST_CANCER_BRCA1_VS_BRCA2
WHITEFORD_PEDIATRIC_CANCER_MARKERS
ODONNELL_TFRC_TARGETS_DN
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN
PID_FOXM1_PATHWAY
YU_MYC_TARGETS_UP
ODONNELL_TARGETS_OF_MYC_AND_TFRC_DN
MITSIADES_RESPONSE_TO_APLIDIN_DN
d
A Higher non-pregnant
B Higher mid-pregnancy Low late pregnancy
C Higher mid-pregnancy Low non-pregnant
D Higher late pregnancy
Milk protein genes
ZEB1
AHNAK
DDIT4
* MMP11
PHLDA3
GSTM1
WNT5A
AKT2
AREG
BCL6
* PGR
MAPKAPK2
MAPK8
* ESR1
TGFBR3
AKT3
CDH2
TGFBR2
CCNG1
BMP1
IGF1R
ITGA5
* * CCNB1
AURKA
* * MKI67
BIRC5
EZH2
* * MYBL2
KRT5
* KRT14
PLK3
BRCA1
PCNA
STAT5B
BRCA2
CCND1
GDF11
SOCS2
CISH
* TGFB3
SOX10
BAX
AKT1
EGF
TRAF4
ELF5
TFRC
ERBB4
STAT5A
CSN1S1
CSN3
CSN2
100
80
60
40
20
0
IPSI17
IPSI2
IPSI4
IPSI14
IPSI3
IPSI6
IPSI20
IPSI5
IPSI7
IPSI22
IPSI8
IPSI23
IPSI24
IPSI25
IPSI21
IPSI11
IPSI13
IPSI16
IPSI12
IPSI1
IPSI18
IPSI15
IPSI10
IPSI19
IPSI9
None
A
B
C
D
A/B/D

◂**Figure 5.** Both ipsilateral high-risk non-cancer and cancer samples demonstrated aberrant expression patterns of pregnancy-related genes. (**a**) Heat map illustrating relative expression patterns of 52 genes known to be expressed in mammary cancer that exhibit pregnancy-stage related changes in gene expression. Values shown from mouse transcriptomic data for virgin, days 13 and 18 pregnancy timepoints (mouse transcriptomic data). A, B, C, and D indicate four stages of distinct patterns of gene expression that change through pregnancy. Asterisks: Genes that are members of established breast cancer prognostic gene profiles. (**b**) Heat map illustrating relative expression patterns of the pregnancy-stage related genes in breast cancer samples sorted by similarity in gene expression pattern. Resemblance to normal patterns (A–D) during pregnancy indicated below x axis. Patterns overlapping more than one stage indicated. (**c**) Bar graph presenting the top 10 C2 gene sets with the lowest significant FDR q-values identified from the MSigDB Collection utilizing pregnancy-stage related genes that are members of established breast cancer prognostic profiles. (**d**) Heat map illustrating relative expression patterns of pregnancy-stage related genes in ipsilateral high-risk non-cancer samples sorted by similarity in gene expression pattern. Correspondance to normal patterns (A–D) during pregnancy indicated below x axis. Patterns overlapping more than one stage indicated. *MsigDB* Molecular Signatures Database v7.5.1, *C2* Curated gene sets. Color coding: Dark blue to yellow with increasing expression.

Specific stem-cell related genes that were down-regulated included CD24, a gene which when down-regulated is associated with a breast cancer stem cell phenotype[46]. While we posit that expansion of the mammary stem cell population may be a normal homeostatic restoration of breast tissue post-chemotherapy given chemotherapy can induce lobular atrophy[47], the character of the surviving stem cells merits additional study. It is known that breast cancer stem cells can survive chemotherapy and the expression pattern identified does have links to both normal and neoplastic mammary stem cells. Whether or not the enriched population found here is reflective of a protective healing response or might provide precursors for future cancer stem cell development is unknown[48]. Hormone receptor expression is a marker for secondary breast cancer risk[49]. The analysis presented here included three hormone receptor positive samples with neoadjuvant chemotherapy exposure (IPSI21, IPSI19, IPSI15) and three hormone receptor positive samples without receipt of neoadjuvant chemotherapy (IPSI18, IPSI5, IPSI14).

Transcriptome data was used to approach a long-standing hypothesis in the field, that is, aberrant expression patterns of genes expressed in normal development might serve as risk factors for cancer generation[50]. Nine of the 52 pregnancy-associated genes in the presented profile are members of at least one established breast cancer prognostic screen[14]. There are pregnancy-related genes known to have clear links to human breast cancer such as *BIRC5*[14] and *STAT5A*[51], which were found increased in a subset of samples here. A question we considered is whether or not the patterns found related to different menstrual cycle phases[9]. While the study did include women under the age of 50, differences were found between patterns seen here and the luteal phase patterns. Samples with higher expression of five of the 10 genes expressed during the luteal phase were found but expression was limited to these five genes (*BRCA1, CCNB1, MKI67, BIRC5, EZH2, PCNA*) and the full luteal pattern was not seen. Future studies would be recommended to include menstrual cycle history at the time of collection to directly address this question. Identification of both cancer and high-risk samples with combinations of gene expression that are not found during normal pregnancy development is consistent with the possibility of a link between deregulated expression of pregnancy genes and cancer risk.

From a technical aspect, CRC-technology was a cost-effective approach to develop a collection of high-risk human breast cells that could be effectively exploited for comparison of growth characteristics under different mammary-specific media and hormonal conditions. Use of matrix-free scaffold-based nano-culture plates enabled us to process replicate samples in parallel at reasonably high efficiency for simultaneously scoring of viability and mammosphere growth in different media and under different conditions[14]. Consistent with previous reports that CRC technology supports preservation of stem/progenitor cells[21], we found uniform ability to form mammospheres upon secondary culture.

In conclusion, at-risk cells preserved behavioral and transcriptome diversity that could reflect different risk profiles, providing a baseline for further development of possible prognostic platforms for breast cancer risk. It is clear that platforms would be challenging to build. For example, they would require clinical validation that might take decades to develop, given the long time frame for breast cancer recurrence. However, utilization of a CRC-based approach would make this feasible as it enables bio-banking of primary breast at-risk cells that can be renewed as needed. This allows them to be available for recurrent experimentation as new technologies and new questions arise over a long-term follow-up study.

Limitations of the study include inadequate sample numbers from men and individuals over age 70. Previously, we used CRC technology to successfully isolate primary squamous cell carcinoma metastatic to salivary gland from individuals aged 70–80[20], but metastatic epithelial cancer cell and non-cancer breast epithelial cell biology are significantly different and alternative approaches for non-cancerous mammary epithelial cells from aging individuals may be needed[19]. Only one site was utilized due to funding limitations and to facilitate successful handling procedures but future studies could utilize different sites to help expand diversity. Samples from women with ER+/HER2+ breast cancer were relatively over-represented (22%) and samples from triple-negative under-represented (7%) compared to U.S. population frequency (13.4% and 13%, respectively)[52]. The distribution towards successful isolation from younger age individuals contributed to this under-representation because the majority of triple negative samples submitted were from older women. Because the study was designed to characterize in vitro cell behavior at the lowest passage number possible to limit media-induced behavioral and gene expression changes[19], few samples were studied across different passages. Passage five was the highest passage number evaluated. Because the study focused on primary cell behavior, establishment of cell lines was not attempted. A limitation was that it was not possible to obtain primary cell cultures of the same sample at the same

passage at different points in time so each sample could be studied in temporally distinct experiments, rather than with multiple replicates at a single timepoint as was performed here. However, while it was not feasible to test all samples in more than one experiment, a subset of samples that passaged well were repetitively studied in two or three different experiments across passage. Results showed general consistency across the individual experiments (Supplementary Figure 1). Funding was insufficient to include a limiting dilution analysis for mammosphere formation. Protein expression validation was also not able to be included, however, in previous studies we have validated concordance between quantitative and specific RNA and qualitative protein measurements[19,20]. The pregnancy gene expression profile was developed from RNAseq of whole mammary gland tissue, which would also contain populations of stromal and other cells that could influence gene expression levels shown[53]. To help address this issue, the pregnancy-related gene profile was limited to genes known to be expressed in mammary adenocarcinoma cell lines. Although both ipsilateral and cancer sample tissues were evaluated to confirm histological identity, the possibility remains that either an ipsilateral sample would contain some infiltrating tumor cells or, conversely that a tumor sample might contain non-cancerous cells that could affect results.

## Methods

**Research involving human participants.** This study was approved by the Institutional Review Board (IRB) of the Office of Research Oversight/Regulatory Affairs, Georgetown University. It was determined to impose minimal risk on participants. Informed consent was obtained. All research was performed in accordance with relevant guidelines/regulations. This included deidentification of all samples and medical records with assignment of unique GUMC identifier.

**Human sample acquisition and RNA sequencing.** Experiments were designed to ethically collect deidentified human breast tissue from living individuals with a diagnosis of breast cancer or at high-risk for breast cancer, with biospecimens obtained from surgically excised tissue not needed for pathologic diagnosis by a board-certified pathologist at room temperature, stabilized by immersion in CRC-compatible media, with mammary epithelial cells isolated from minced tissue incubated in media containing a collagenase/hyaluronidase/dispase solution followed by culture at 37 °C utilizing CRC conditions[18–21], and viably stored in liquid nitrogen (− 196 °C) for < 1 to 4 years. Ipsilateral non-cancerous breast tissue (volume 2 $cm^3$) were procured from grossly unremarkable fibroadipose mammary parenchyma, at least 6–8 cm away from any tumor. Breast cancer tissue (volume 2 $cm^3$) was included for comparative evaluation when consent provided. Tissue was processed into two "mirrored" (volume $0.5^3$–1 $cm^3$) samples: one for CRC technology and one for formalin-fixation. Deidentified pathology reports provided clinical and pathology diagnoses including age, breast cancer subtype, pathologic stage, *BRCA* gene mutation status, lymph node or other metastasis, and history of neoadjuvant chemotherapy. Numbers of ipsilateral tissue samples for acquisition were determined prior to the experiment (n = 50). Fifty-six samples were obtained. The only study inclusion factor was that the individual was undergoing mastectomy for breast cancer treatment. Final classification of samples as invasive tumor or non-cancer was determined after review of H&E sections of the formalin-fixed "mirrored" tissue by a board-certified pathologist. One ipsilateral non-cancer specimen was re-classified as invasive tumor. All samples very processed equivalently and no ipsilateral or invasive tumor submitted sample was excluded from the study. Unique Georgetown University Medical Center (GUMC) primary cell culture identifiers were assigned following initial isolation[20]. Epithelial cell growth was separated from the underlying fibroblast feeder layer by differential trypsin treatment, divided, and viably frozen as Passage 0 (P0). The majority (83%) of ipsilateral samples yield 4 tubes with a minority yielding only two tubes. RNA sequencing (RNAseq) was performed on P0 tubes when > 2 tubes available with remaining samples being secondarily passaged using CRC technology and a P1 tube used. Total RNA isolated from cell pellets (RNeasy Mini Kit, Qiagen, Gaithersburg, MD) was quantified, analyzed for quality (Nanodrop, ThermoFisher Scientific, Wilmington, DE Bioanalyzer 2100, Agilent Technologies, Santa Clara, CA). For 86% of the ipsilateral samples, 1 µg ribosome-depleted RNA was used for stranded-specific paired-end library preparation (Illumina TruSeq Stranded mRNA Library Preparation Kit (polyA cDNA synthesis) (San Diego, CA) and sequenced (Illumina HiSeq4000 machine, 150 bp pair-ended lane, minimum reads ≥ 100 M per sample). For 14% of the ipsilateral samples, 1 µg ribosome-depleted RNA was used for shotgun library construction (200 bp insert) and sequenced (Illumina HiSeq2000, 91 bp pair-ended lane generating 2 Gb/sample)[20]. Quality check (FastQC), quality trimming (Trim galore) and alignment (STAR) were performed[54] according to library preparation method with batch effect normalization[55]. Normalized expression levels were estimated by means of transcripts per million (TPM) using RSEM[56]. Sequences were mapped to a merged human + mouse genome file (HG38 + mm10) to assess for mouse fibroblast feeder contamination of epithelial cell pellets. Samples with > 10% mouse sequence contamination were excluded from further analysis. Sequences were then mapped to a mycoplasma genome file for detection of mycoplasma contamination[29]. Samples with detectable mycoplasma sequences were excluded from further analysis. Sample numbers retained following screening for human and mycoplasma sequences and confirmed diagnosis of invasive tumor in the individual from whom samples were obtained were ipsilateral (IPSI) n = 25 (97% ± 0.44 (mean ± SEM) human sequence) and invasive tumor (T) n = 8 (97% ± 0.86 human sequence) (Table 1). Differentially expressed genes were identified using DESeq2[57]. Genes were considered statistically significantly differentially expressed when Padj < 0.05.

**Passage in MEGM, comparative culture in MEGM, EpiCult™, MammoCult™, CRC$^{CM}$ , and assessment of hormonal response.** Cells were thawed from CRC pellet with the exception of one ipsilateral sample that was trypsinized (Trypsin-EDTA 0.05%, Thermofisher Scientific Waltham, MA) directly from secondary CRC culture. Cells were collected, centrifuged at 1000xRPM, reconstituted in serum-free, phenol-red free MEGM (Lonza) and counted using a TC20™ Automated Cell Counter (Bio-Rad Laboratories, Hercules, CA).

Cells were assessed for expansion in MEGM for one passage with media renewal every third day in T25 flasks (Nest Biotechnology, Jiangsu, China) at 37 °C in 5% CO2 incubator with a goal of 80% confluency to enable sufficient cell numbers for comparative media and hormonal growth studies. Biochemical mycoplasma testing was performed following MEGM passage (MycoAlertTM Mycoplasma Detection Kit, Lonza). Cultures testing positive or with inadequate testing material were excluded from further analyses. When target confluency was reached, cell cultures were trypsinized (phenol-red-free, TrypLE Express Enzyme (1x) trypsin,ThermoFisher Scientific cat no: 12604021), collected, centrifuged at 1000xRPMs, reconstituted in MEGM, counted, viability estimated (0.4% trypan blue in saline dye exclusion) and seeded into 3D NanoCulture Low-Binding Micro Honeycomb 96 well-plates (ORGANOGENIX, Japan)[25] ($10 \times 10^3$ cells/well, 0.1 mL medium/well, n = 6 replicates/condition/passage number with exceptions for insufficient cell numbers (IPSI9, n = 3 replicates (EpiC); IPSI24 (CRC Conditioned Media), T3, IPSI14 n = 5 replicates (E2), IPSI12 n = 4 replicates (4-OHT). For each sample, multiple replicates of independently plated and processed cells at a single passage within a single experiment were performed to limit possible passage induced variability. Cell viability and mammosphere formation after seven days growth were determined in three different mammary specific media; MEGM, EpiCult™-C Human Media (STEMCELL Technologies) supplemented with 10 ng/mL EGF (cat no: PHG0311) and bFGF (cat no: PHG0261) (Thermofischer Scientific) and 0.48 µg/mL Hydrocortisone (STEMCELL Technologies, MammoCult™ Human Media (MammoC) (STEMCELL Technologies) supplemented with 4 µg/mL Heparin (STEMCELL Technologies) and 0.48 µg/mL Hydrocortisone, and Conditioned Media (CM)[20]. Media was renewed daily by removing 0.05 mL media and replacing it with 0.05 mL fresh media. Hormone response was assessed in serum-free, phenol-red free, MEGM. 17β-Estradiol (E2) (cat no:50–28-2) and 4-Hydroxytamoxifen (4-OHT) (98% Z isomer, cat no: 68047–06-3) (Sigma Aldrich (St. Louis MO). Stock solutions were prepared in pure Ethanol (EtOH), stored at 20 °C in 1 mL aliquots and diluted in MEGM for use in cell culture. Cell cultures were treated with 10 nM E2/0.1% EtOH (1 mM), 1 µM 4-OHT/0.1% EtOH (1 mM) or vehicle (EtOH) alone for seven days with daily renewal of media and treatment.

**Mammosphere and cell viability measurements.** Cell aggregates were counted as mammospheres if they were equal or greater than 100 µm in diameter[58,59]. After seven days of culture at 37 °C in a 5% CO2 incubator, each well was imaged in entirety using phase contrast microscopy (4x, EVOS FL Cell Imaging System, Life Technologies, Paisley, UK), numbers of mammospheres counted in each well, and cell growth patterns recorded as monolayer only, mammosphere only, or mixed monolayer and mammosphere. Cell viability was measured (CellTiter-Glo® 3D, Promega Corporation, Madison, WI, cat no: G9681) by removing all media, adding reagent to each well, then shaking plates (five minutes). Luminescence was recorded after 30 min (room temperature) with the signal measured at 1.0.s increments using a VICTOR Multilabel Plate Reader (PerkinElmer, Waltham, MA).

**Gene set enrichment analyses and data visualizations.** Mycoplasma-free samples on sequence analysis [ipsilateral (IPSI, n = 25) and invasive tumor (T, n = 8) samples placed into initial MEGM culture were queried for TPM values of HALLMARK_E2F_TARGETS gene set (Molecular Signatures Database v7.4 C1, accessed July 2021)[31,32]. TPM values were normalized and relative expression levels visualized (Heat Map, GraphPad Prism 9.3.1, GraphPad Software, LLC. San Diego, CA). To investigate possible gene expression changes induced by neoadjuvant chemotherapy in ipsilateral breast of individuals with ER/PR/HER2+ or HER2+ invasive breast cancer, DEGs[57] were analyzed in samples from individuals that received neoadjuvant chemotherapy prior to sample acquisition (n = 4) as compared to samples from individuals that did not (n = 4). Significantly up- and down-regulated DEGs were analyzed separately (C2 gene sets, GSEA)[31,33]. Enrichment in gene sets related to mammary epithelial stem cells were identified for both (accessed July 2021) (FDR q-value < 0.05). TPM values of the genes found enriched in these two gene sets were normalized and relative expression levels visualized (Heat Map, GraphPad Prism 9.3.1). All cultures with sufficient numbers of cells for this evaluation were tested for significant differences in viability and mammosphere formation in the presence of E2 (n = 18). As a comparator, viability in 4-OHT was examined in the nine samples that had additional sufficient cells for this evaluation. The transcriptomes of samples with significantly higher viability in E2 were compared to those with equivalent viability to identify genes expressed at significantly different levels (Multiple unpaired t tests with Welch correction, Two-stage step-up, Graphpad Prism, q ≤ 0.004 considered statistically significant). Significantly differently up- and down-regulated genes were analyzed separately for enrichment in HALLMARK gene sets (GSEA, accessed February 2022, FDR q-value < 0.05). TPM values of genes found in enriched gene sets were visualized on bar graphs (GraphPad Prism 9.3.1).

**Pregnancy-linked breast cancer risk genes.** A heat map illustrating changes in relative gene expression in the mammary gland during pregnancy was generated using downloaded data generated from virgin and pregnant 2-month-old mice (GSE70440)[8] (GraphPad Prism 9.3.1). Four patterns (A–D) of gene expression changes during pregnancy were identified based on their relative expression levels in non-pregnant (virgin), day 13 and day 18 pregnancy. Pregnancy-related genes that are included in at least one validated breast cancer prognosis platform were identified[26] and subjected to GSEA analysis for C2 gene set enrichment. Heat maps of pregnancy-related genes were generated separately for invasive tumor (T) and ipsilateral (IPSI) samples (GraphPad Prism 9.3.1). Gene expression patterns of the individual samples were visually sorted by similarity. Three different patterns were recognized in the invasive tumor samples and six different patterns in the ipsilateral samples. These patterns were then compared to the four pregnancy-related patterns. Patterns were assigned to a single pregnancy-stage-related pattern or overlapping pregnancy-stage-related patterns dependent upon the gene expression data.

**Analyses for chromosomal deletions and amplifications.** Transcriptomes from four mycoplasma-free validated invasive tumor/ipsilateral pairs were available to assess for known breast cancer-related chromosomal deletions/amplifications (Suppl. Table 1)[57]. DEGs between T4/IPSI4, T13/IPSI13, T3/IPSI3 and T10/IPSI10 were identified using DESeq2 (Padj < 0.05 considered statistically significant). Chromosomal regions of differentially expressed genes were identified using C1:positional gene sets (Molecular Signatures Database v7.4 C1, nominal $p$ value < 1%) (accessed August 2019)[31,33]. cBioPortal (accessed February 2021, nine different human breast cancer databases queried) was used to identify specific percentages of chromosomal region/gene amplifications/deletions in the paired samples[60–62].

**Statistical analyses.** For RNAseq, quality check (FastQC), quality trimming (Trim galore) and alignment (STAR) were performed[54] according to library preparation method and batch effect normalization conducted[55]. Normalized expression levels were estimated by means of transcripts per million (TPM) using RSEM[56] and differentially expressed genes were identified using DESeq2 ($padj \leq 0.05$ considered statistically significant)[57]. Mean ± standard error of the mean (SEM) for patient ages, viability measurements and mammosphere counts were calculated (GraphPad Prism 9.3.1). Ordinary one-way ANOVA with Brown-Forsythe test and Sidak's multiple comparisons test was used to compare probability of CRC isolation by age and neoadjuvant chemotherapy exposure ($p \leq 0.05$ considered statistically significant, GraphPad Prism 9.3.1). Fisher's exact, two-tailed was used to compare probability of MEGM passage in cancer versus non-cancer cells ($p \leq 0.05$ considered statistically significant, GraphPad Prism 9.3.1). 2way ANOVA was used to analyze for statistically significant interactions between samples and media for viability and mammosphere numbers for groups with three and four media ($p \leq 0.05$ considered statistically significant, DF values reported in figure legend 1, GraphPad Prism 9.3.1). Multiple unpaired t-tests using FDR approach (Two-stage step-up method of Benjamin, Krieger and Yekutieli, GraphPad Prism 9.3.1) was used to examine for statistically significant differences for two media comparison of viability and mammosphere numbers, differences in hormonal response for viability, mammosphere numbers and TPM values for hormonal response ($p < 0.05$ statistically significant, t, df values reported in figure legend 3, GraphPad Prism 9.3.1). Simple linear regressions were conducted to analyze for significant relationships between MEGM viability and patient age ($p \leq 0.05$ statistically significant, GraphPad Prism 9.3.1). Regression equations, $p$ and R squared values presented on regression scatter plot graph (Fig. 1c). Scatter plots, stacked bar graphs, bar graphs and heat maps were prepared in GraphPad Prism 9.3.1.

**Ethical approval.** This study was approved by the Institutional Review Board (IRB) of the Office of Research Oversight/Regulatory Affairs, Georgetown University.

**Informed consent.** Informed consent was obtained prior to sample acquisition.

## Data availability

The data discussed in this publication were deposited in NCBI's Gene Expression Omnibus[63], accessible through GEO Series accession number GSE 185314 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE185314).

Received: 30 October 2021; Accepted: 1 April 2022
Published online: 22 April 2022

## References

1. Stringer-Reasor, E. M., Elkhanany, A., Khoury, K., Simon, M. A. & Newman, L. A. Disparities in breast cancer associated with African American identity. *Am. Soc. Clin. Oncol. Educ. Book* **41**, e29–e46 (2021).
2. Feng, Y. *et al.* Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis.* **5**, 77–106 (2018).
3. Shrestha, A. *et al.* Clinical Treatment Score post-5 years as a predictor of late distant recurrence in hormone receptor-positive breast cancer: Systematic review and meta-analysis. *Br. J. Surg.* https://doi.org/10.1093/bjs/znac008 (2022).
4. Pedersen, R. N. *et al.* The incidence of breast cancer recurrence 10–32 years after primary diagnosis. *J. Natl. Cancer Inst.* https://doi.org/10.2139/ssrn.3762484 (2021).
5. Muse, M. E. *et al.* Enrichment of CpG island shore region hypermethylation in epigenetic breast field cancerization. *Epigenetics* **15**, 1093–1106 (2020).
6. Holliday, H., Baker, L. A., Junankar, S. R., Clark, S. J. & Swarbrick, A. Epigenomics of mammary gland development. *Breast Cancer Res.* **20**, 100 (2018).
7. Fu, N. Y., Nolan, E., Lindeman, G. J. & Visvader, J. E. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol. Rev.* **100**, 489–523 (2020).
8. Yoo, K. H. *et al.* Loss of EZH2 results in precocious mammary gland development and activation of STAT5-dependent genes. *Nucleic Acids Res.* **43**, 8774–8789 (2015).
9. Pardo, I. *et al.* Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank. *Breast Cancer Res.* **16**, R26 (2014).
10. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 2028 (2018).
11. Lim, E. *et al.* Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
12. Pirone, J. R. *et al.* Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1735–1744 (2012).
13. Danforth, D. N. *et al.* Characteristics of breast ducts in normal-risk and high-risk women and their relationship to ductal cytologic atypia. *Cancer Prev. Res.* **13**, 1027–1036 (2020).
14. Vieira, A. F. & Schmitt, F. An update on breast cancer multigene prognostic tests-emergent clinical biomarkers. *Front. Med. (Lausanne)* **5**, 248 (2018).

15. Behravan, H., Hartikainen, J. M., Tengström, M., Kosma, V.-M. & Mannermaa, A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci. Rep.* **10**, 11044 (2020).
16. Berthois, Y., Katzenellenbogen, J. A. & Katzenellenbogen, B. S. Phenol red in tissue culture media is a weak estrogen: Implications concerning the study of estrogen-responsive cells in culture. *Proc. Natl. Acad. Sci. USA* **83**, 2496–2500 (1986).
17. Jedrzejczak-Silicka, M. History of cell culture. In *New Insights into Cell Culture Technology* (ed. Gowder, S. J. T.) (IntechOpen, 2017).
18. Liu, *et al.* Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat. Protoc.* **12**, 439 (2017).
19. Alamri, A. M. *et al.* Primary cancer cell culture: Mammary-optimized vs conditional reprogramming. *Endocr. Relat. Cancer* **23**, 535–554 (2016).
20. Alamri, A. M. *et al.* Expanding primary cells from mucoepidermoid and other salivary gland neoplasms for genetic and chemosensitivity testing. *Dis. Model Mech.* **11**, dmm031716 (2018).
21. Suprynowicz, F. A. *et al.* Conditionally reprogrammed cells represent a stem-like state of adult epithelial cells. *Proc. Natl. Acad. Sci. USA* **109**, 20035–20040 (2012).
22. Jin, L. *et al.* Characterization of primary human mammary epithelial cells isolated and propagated by conditional reprogrammed cell culture. *Oncotarget* **9**, 11503–11514 (2017).
23. Brown, D. D. *et al.* Developing in vitro models of human ductal carcinoma in situ from primary tissue explants. *Breast Cancer Res. Treat.* **153**, 311–321 (2015).
24. Leeper, A. D. *et al.* Determining tamoxifen sensitivity using primary breast cancer tissue in collagen-based three-dimensional culture. *Biomaterials* **33**, 907–915 (2012).
25. Arai, K. *et al.* A novel high-throughput 3D screening system for EMT inhibitors: A pilot screening discovered the EMT inhibitory activity of CDK2 inhibitor SU9516. *PLoS ONE* **11**, e0162394 (2016).
26. Lee, J. K. *et al.* Different culture media modulate growth, heterogeneity, and senescence in human mammary epithelial cell cultures. *PLoS ONE* **13**, e0204645 (2018).
27. Qu, Y. *et al.* Differentiation of human induced pluripotent stem cells to mammary-like organoids. *Stem Cell Rep.* **8**, 205–215 (2017).
28. Nikfarjam, L. & Farzaneh, P. Prevention and detection of mycoplasma contamination in cell culture. *Cell J.* **13**, 203–212 (2012).
29. Olarerin-George, A. O. & Hogenesch, J. B. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Res.* **43**, 2535–2542 (2015).
30. Ferreira, M. A. *et al.* Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat. Commun.* **10**, 1741 (2019).
31. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
32. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
33. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
34. Benedetti, L., Cereda, M., Monteverde, L., Desai, N. & Ciccarelli, F. D. Synthetic lethal interaction between the tumour suppressor STAG2 and its paralog STAG1. *Oncotarget* **8**, 37619–37632 (2017).
35. Oh, S., Oh, C. & Yoo, K. H. Functional roles of CTCF in breast cancer. *BMB Rep.* **50**, 445–453 (2017).
36. Mercogliano, M. F., Bruni, S., Elizalde, P. V. & Schillaci, R. Tumor necrosis factor α blockade: An opportunity to tackle breast cancer. *Front. Oncol.* **10**, 584 (2020).
37. Lee, N.-H. *et al.* SERPINB2 is a novel indicator of cancer stem cell tumorigenicity in multiple cancer types. *Cancers (Basel)* **11**, 499 (2019).
38. Regan, J. L. *et al.* Aurora A kinase regulates mammary epithelial cell fate by determining mitotic spindle orientation in a notch-dependent manner. *Cell Rep.* **4**, 110–123 (2013).
39. Jin, Y., Yang, L., Li, X. & Liu, F. Circular RNA KIF4A promotes cell migration, invasion and inhibits apoptosis through miR-152/ZEB1 axis in breast cancer. *Diagn. Pathol.* **15**, 55 (2020).
40. Rubio, M. F. *et al.* TNF-alpha enhances estrogen-induced cell proliferation of estrogen-dependent breast tumor cells through a complex containing nuclear factor-kappa B. *Oncogene* **24**, 1367–1377 (2006).
41. Fu, X., De Angelis, C. & Schiff, R. Interferon signaling in estrogen receptor–positive breast cancer: A revitalized topic. *Endocrinology* **163**, bqab235 (2021).
42. Katzenellenbogen, B. S. *et al.* Molecular mechanisms of estrogen action: Selective ligands and receptor pharmacology. *J. Steroid Biochem. Mol. Biol.* **74**, 279–285 (2000).
43. Vogel, V. G. Role of hormones in cancer prevention. *Am. Soc. Clin. Oncol. Educ. Book* https://doi.org/10.1469/EdBook_AM.2014.34.34 (2014).
44. Woosely, N. N. *et al.* TGFβ promotes breast cancer stem cell self-renewal through an ILEI/LIFR signaling axis. *Oncogene* **38**, 3794–3811 (2019).
45. Lobo, N. A., Zabala, M., Qian, D. & Clarke, M. F. Serially transplantable mammary epithelial cells express the Thy-1 antigen. *Breast Cancer Res.* **20**, 121 (2018).
46. Vesuna, F., Lisok, L., Kimble, B. & Raman, V. Twist modulates breast cancer stem cells by transcriptional regulation of CD24 expression. *Neoplasia* **11**, 1318–1328 (2009).
47. Aktepe, F., Kapucuoğlu, N. & Pak, I. The effects of chemotherapy on breast cancer tissue in locally advanced breast cancer. *Histopathology* **29**, 63–67 (1996).
48. Lu, H. *et al.* Chemotherapy-induced S100A10 recruits KDM6A to facilitate OCT4-mediated breast cancer stemness. *J. Clin. Investig.* **130**, 4607–4623 (2020).
49. Bessanova, L., Taylor, T. H., Mehta, R. S., Zell, J. A. & Anton-Culver, H. Risk of a second breast cancer associated with hormone-rec eptor and HER2/neu status of the first breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **20**, 389–396 (2011).
50. Rubin, P., Williams, J. P., Devesa, S. S., Travis, L. B. & Constine, L. S. Cancer genesis across the age spectrum: Associations with tissue development, maintenance, and senescence. *Semin. Radiat. Oncol.* **20**, 3–11 (2010).
51. Furth, P. A., Nakles, R. E., Millman, S., Diaz-Cruz, E. S. & Cabrera, M. C. Signal transducer and activator of transcription 5 as a key signaling pathway in normal mammary gland developmental biology and breast cancer. *Breast Cancer Res.* **13**, 220 (2011).
52. Female Breast Cancer Subtypes—Cancer Stat Facts. *SEER*. https://seer.cancer.gov/statfacts/html/breast-subtypes.html (2021).
53. Henry, S. *et al.* Characterization of gene expression signatures for the identification of cellular heterogeneity in the developing mammary gland. *J. Mammary Gland Biol. Neoplasia* **26**, 43–66 (2021).
54. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
56. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
58. Wang, R. *et al.* Comparison of mammosphere formation from breast cancer cell lines and primary breast tumors. *J. Thorac Dis.* **6**, 829–837 (2014).

59. Lombardo, Y., de Georgio, A., Coombes, C. R., Stebbing, J. & Castellano, L. Mammosphere formation assay from human breast cancer tissues and cell lines. *J. Vis. Exp.* **97**, 52671 (2015).
60. De Preter, K., Barriot, R., Speleman, F., Vandesompele, J. & Moreau, Y. Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucleic Acids Res.* **36**, e43 (2008).
61. Cerami, E. *et al.* The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
62. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
63. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

## Acknowledgements

The authors gratefully acknowledge the support of Survey, Recruitment, and Biospecimen Collection, Histopathology and Tissue, Tissue Culture and Biobanking, and Genomics & Epigenomics GUMC Shared Resources, Grace Keegan for assistance in developing the pregnancy-associated developmental gene profile and Xiaogang Zhong for preliminary RNAseq analyses.

## Author contributions

S.J.A. & P.A.F. contributed study design and conception, acquisition, analysis, and interpretation of data, wrote the main manuscript text. P.A.F. prepared all figures. K.K., X.L., E.K., R.I.A., R.H., D.G. & B.V.K. contributed acquisition and/or analysis of data. All authors reviewed and approved the manuscript.

## Competing interests

Financial Competing Interests, Disclosures and Funding: X.L. declares the following competing interests: several patents for conditional reprogramming technology have been awarded to Georgetown University by the United States Patent Office. The license for this technology has been given to a Maryland-based start-up company for commercialization. The inventor, X.L., and Georgetown University receive potential royalties and payments from the company. Several organizations and companies (Propagenix, ATCC, STEMCELL Technologies, etc.) are selling CR cells, media and related reagents. Funding: NCI, NIH RO1CA112176 (P.A.F., B.V.K.), NCI, NIH P30CA051008 (P.A.F., X.L., E.K., R.H., D.G.), King Abdullah Scholarship Program, Ministry of Higher Education, Kingdom of Saudi Arabia (S.J.A.). K.K. & R.I.A. declare no competing financial interests. Non-financial competing interests: S.J.A., K.K. X.L., E.K., R.I.A., R.H., D.G., B.V.K. & P.A.F. declare no non-financial competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10246-4.

**Correspondence** and requests for materials should be addressed to P.A.F.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.