Differential expression of integrin β4 (ITGB4) mRNA in epithelial and mesenchymal-like mammary epithelial cells.

We began with the hypothesis that bulk mesenchymal-like epithelial cell populations were comprised of a heterogeneous spectrum of cells that differed in their relative epithelial versus mesenchymal cell states. We reasoned that, by characterizing non-neoplastic human mammary epithelial cells (MECs) residing in various cell states along the epithelial-mesenchymal spectrum, we might identify gene expression that was enriched in cells residing in one or another of these states; once identified, such genes might be useful for resolving distinct subpopulations of cells present in bulk mesenchymal-like carcinoma cell populations. Implicit in this analysis was the notion, supported by our own repeated experience, that the EMT programs of non-neoplastic MECs are closely paralleled by such programs operating in derived neoplastic cells (1, 2). Accordingly, we analyzed RNAseq data derived from well-characterized, immortalized but non-tumorigenic HMLE MECs (mammary epithelial cells; (3)), and their derivatives, a spontaneously arising, highly mesenchymal cell population termed NAMECs (naturally arising mesenchymal MECs; (4)). HMLE cells exhibit an outwardly epithelial phenotype, forming a clustered cobblestone morphology in monolayer culture, whereas the NAMECs exhibit a scattered, mesenchymal morphological phenotype (Figure S1A).

Initially, we focused on the expression of integrins, based on our observations that the more epithelial HMLE cells adhered more tightly to the cell culture dish than did their highly mesenchymal NAMEC derivatives; this suggested greater adhesion of epithelial cells via integrins to components of the extracellular matrix laid down by these cells in monolayer culture. The integrins were also selected because they had the potential to be exploited as cell-surface markers that would be useful in flow cytometry analyses. We elected to use a polyclonal NAMEC population for initial analyses, doing so in order to increase the likelihood of success in resolving distinct mesenchymal cell sub-populations within these NAMECs if they did indeed exist.

To begin, RNAseq analyses were used to compare the outwardly epithelial HMLE cells to a derived, more mesenchymal polyclonal NAMEC cell population termed NAMEC8 (Supplemental Dataset 1) (5). We focused on loss of integrin expression as a potential contributor to the previously observed, reduced substrate adhesion of the more mesenchymal MECs. Integrin mRNAs that were both readily detected in the highly epithelial cells and expressed at significantly lower levels in the more mesenchymal cells included *ITGA2*, *ITGA6*, *ITGB4*, and *ITGB6* (Figure S1B). Among these, *ITGB4*, which encodes a receptor for the laminin basement membrane protein (6), was abundantly expressed by the epithelial HMLE cells and decreased in the more mesenchymal MECs by approximately 10-fold, (Figure S1B). Importantly, this protein had not been previously tested for its ability to serve as a useful marker for the stratification of TNBC cells.

Meta-analyses of RNAseq data generated previously by others studying 50 common human breast cancer cell lines (7) revealed that *ITGB4* mRNA expression did not show a significant correlation with other previously characterized epithelial and mesenchymal cell-surface markers, including *PROM1* (CD133), *ITGA6* (CD49f), *EpCAM* (ESA), *MUC1* (CD227), *THY1* (CD90) and *PROCR* (CD201) (Supplemental Dataset 2). The unique gene expression profile associated with *ITGB4* suggested that it could be employed as a non-redundant marker to resolve distinct subsets of mammary carcinoma cells that had not been previously characterized. Further characterization revealed that decreased *ITGB4* mRNA expression was a generalizable property of an EMT program induced by a variety of stimuli (Figure S1C)(8).

Use of Integrin β4 (ITGB4) cell surface abundance to stratify basal-like epithelial and mesenchymal-like mammary epithelial cells.

We proceeded to determine whether the *ITGB4* mRNA levels reflected corresponding changes in protein abundance. To do so, we used FACS (fluorescence-activated cell sorting) to analyze cell-surface expression profiles of CD44, CD24 and ITGB4 using the parental, largely epithelial HMLE cells (Figure S1D). Consistent with previous work (9), the HMLE cells exhibited a predominantly epithelial $CD44^{lo}CD24^{hi}$ marker phenotype with a small minority subpopulation exhibiting a more mesenchymal $CD44^{hi}CD24^{lo}$ marker combination (Figure S1D, blue arrow). The profile generated using the combination of CD44 and ITGB4 markers revealed a small population of $CD44^{hi}$ cells that exhibited reduced levels of ITGB4 when compared with the $CD44^{lo}$ population (Figure S1D, red arrows). To determine whether ITGB4 cell-surface abundance was indeed altered in cells passing through an EMT program – thereby reflecting the behavior of cells residing in a naturally arising mesenchymal cell state – we used expression vectors to induce an EMT in the HMLE cells via constitutive expression of either the SNAIL or TWIST EMT-TFs. Following induction of an EMT, ITGB4 was significantly decreased, as determined by FACS analyses (HMLE, HMLE-SNAIL, and HMLE-TWIST histograms; Figure S1E).

The relative degree of EMT induction in the $ITGB4^{hi}$ and $ITGB4^{lo}$ HMLE cells that were engineered to express either SNAIL or TWIST was determined after segregating the cells via FACS (Figure S1F and S1G). Using ITGB4 as a single marker for cell sorting, the HMLE cells expressing SNAIL or TWIST could be segregated into subpopulations that, in monolayer culture, either retained an epithelial morphological phenotype or had gained an outwardly mesenchymal morphology (Figure S1F). Western blot analyses of the bulk HMLE cells and their sorted SNAIL- or TWIST-expressing derivatives confirmed that the acquisition of mesenchymal markers correlated inversely with the level of ITGB4 expression (Figure S1G), representing a further indication that expression of ITGB4 was aligned more closely with the epithelial state.

The traditionally used CD24 and CD44 markers have been effective in segregating highly epithelial cells from an apparently homogeneous population of highly mesenchymal cells. Replacement of the CD24 marker by ITGB4, however, revealed a broad range of ITGB4 abundance in the $CD44^{hi}$ fraction, indicating the presence of heterogeneity within the mesenchymal population (Figure S1D). This suggested, in turn, that ITGB4 might be used to more effectively resolve different subtypes of the more mesenchymal epithelial cells. Indeed, from the perspective of their $CD44^{hi}CD24^{lo}$ profile, the polyclonal NAMEC8 cells – mesenchymal derivatives of the HMLE cells – appeared to constitute a relatively homogeneous population (Figure S2A). However, upon employing the ITGB4 marker, the same NAMEC8 cells exhibited a broad spectrum of expression that spanned four orders-of-magnitude as determined by FACS analyses (Figure S2A).

To determine if the broad range of ITGB4 expression observed within the NAMEC8 population represented the persistent presence of distinct sub-types of mesenchymal cells, we used FACS to segregate $ITGB4^{hi}$ and $ITGB4^{lo}$ NAMEC8 cells. This revealed that indeed, distinct epithelial and mesenchymal populations with $ITGB4^{hi}$ and $ITGB4^{lo}$ phenotypes could be isolated (Figure S2B) and propagated separately thereafter for as long as a month in continuous culture without an apparent interconversion between the $ITGB4^{hi}$ and $ITGB4^{lo}$ states. Hence, the cells in these two distinct subpopulations stably maintained their residence in distinct phenotypic states over an extended period of time in vitro.

In order to determine in an unbiased manner, how these two NAMEC8 CD44^{hi} subpopulations – ITGB4^{hi} and ITGB4^{lo} – differed from one another with respect to their expression of epithelial versus mesenchymal genes, we performed RNAseq analyses (Figure S2C, Supplemental Dataset 3). These analyses were superior to using morphological phenotypes or the expression of canonical EMT marker genes as measures of their relative epithelial versus mesenchymal traits, since neither of these commonly used characteristics could be used to resolve the ITGB4^{hi} and ITGB4^{lo} cell populations from one another (Figure S2D-E). Wishing to further test the notion that ITGB4^{hi} cells were actually more epithelial than their ITGB4^{lo} counterparts, we filtered the differentially expressed genes to focus on those that we had previously found to be differentially expressed when comparing the epithelial HMLE and mesenchymal NAMEC8 cell lines with one another (Supplemental Dataset 1) (5). This resulted in a NAMEC8 ITGB4$^{hi/lo}$ expression signature comprised of genes that were also differentially expressed when comparing the epithelial HMLE cells to their more mesenchymal NAMEC8 counterparts (Supplemental Dataset 4).

We then subjected this expression signature to unsupervised hierarchical clustering and performed a pairwise comparison with the HMLE versus NAMEC8 RNAseq data (Figure S2F, and Supplemental Dataset 4). These analyses confirmed that the NAMEC8 ITGB4^{hi} mesenchymal cells exhibited a more epithelial gene expression profile than their ITGB4^{lo} counterparts. Notably, the genes that were able to classify the ITGB4^{hi} and ITGB4^{lo} mesenchymal cells as being more epithelial or more mesenchymal also demonstrated relatively large differences in expression when comparing the epithelial HMLE cells to the mesenchymal NAMEC cells (Figure S2G-H). Together, these results also indicated that ITGB4 could be used to resolve immortalized, outwardly mesenchymal epithelial cells into distinct subpopulations, doing so in a manner that could not be achieved with the commonly used CD44 and CD24 markers.

HMLE and NAMEC8 gene expression profiles cluster respectively with those obtained from more epithelial and more mesenchymal triple negative breast cancer cells

In order to determine which subtypes of breast cancer were most relevant for validation and extension of our studies with the HMLE and NAMEC8 cells described above, we performed unsupervised hierarchical clustering analyses of RNAseq data in order to compare the expression patterns of these two cell lines to those of 49 frequently studied human breast cancer cell lines, using data reported by others (Figure S3; (10)). These analyses revealed that the HMLE and NAMEC8 cells were more closely related in their gene expression profiles to human triple-negative breast cancer (TNBC) cell lines than to those of the more luminal breast cancer subtypes (Figure S3). Further, following a previously described classification of TNBC cell lines (11, 12), we determined that the HMLE gene expression profile was associated with the more epithelial basal-like (BL1/BL2) TNBC cell lines, while that of the NAMEC8 cells was more closely associated with the mesenchymal and mesenchymal stem-like (M/MSL) TNBC lines (Figure S3). Importantly, this subtype of breast cancer has been previously shown to be enriched with carcinoma cells exhibiting certain mesenchymal traits (13-18). Moreover, TNBCs are often associated with a poor clinical prognosis (15, 19, 20). Based on these results, TNBC was selected as the focus of our subsequent analyses.

Selected epithelial vs mesenchymal gene expression in SUM159 ITGB4^{hi} and ITGB4^{lo} cells

RNAseq and quantitative real-time PCR analyses conducted using SUM159 ITGB4^{hi} and ITGB4^{lo} cells revealed only minor alterations in canonical EMT-assciated gene expression (Figure S5A and S5B). However, further analyses revealed that ITGB4^{hi} cells exhibited far higher levels of an epithelial marker,

TAp63α, while the ITGB4^{lo} cells expressed higher levels of AXL, which has been previously associated with the more mesenchymal state (Figure S5C) (21-24). Notably, using ITGB4 to isolate mesenchymal subtype mammary epithelial cells revealed a common set of EMT-associated genes that were enriched in both SUM159 ITGB4^{hi} and NAMEC8 ITGB4^{hi} cell populations when compared with their ITGB4^{lo} counterparts (Figure S5D and S5E).

In addition to TAp63α and AXL expression, which have been differentially associated with the epithelial and mesenchymal cell states, the SUM159 ITGB4^{hi} and ITGB4^{lo} cells also differed in their expression of the aldehyde dehydrogenase family members which have been previously associated with cancer stem cell state (25, 26). ALDH1A1 and ALDH7A1 were enriched in the ITGB4^{hi} and ITGB4^{lo} populations, respectively (Figure S6A). However, the expression of ALDH1A3, which is the predominant ALDH family member found to correlate with poor prognosis in TNBC (27), was nearly undetectable in both populations (Figure S6A). This was consistent with results reported by others indicating that ALDH activity is higher in the more epithelial CSC populations (25, 26). This observation led to further analyses, in which we confirmed the expression of ALDH1A3 as an epithelial marker in the ten TNBC populations we analyzed (Figure S6B and S6C)(7). Indeed, the more mesenchymal TNBC cell lines failed to exhibit a significant degree of ALDH activity, as gauged by the Aldefluor assay when directly compared with the more epithelial TNBCs (Figure S6D), indicating that ITGB4 may have utility in identifying and isolating unique subpopulations of the more mesenchymal TNBCs that are difficult to obtain using methods that are dependent on ALDH-activity.

Extended Materials and Methods

Cell lines and culture conditions

HMLE and NAMEC based cell lines were cultured essentially as previously described using MEGM medium. MEGM medium was produced using the MEGM Bullet Kit from Lonza (500ml, product #CC-3150; with the exception of gentamicin/amphotericin-B which was excluded). The MEGM was further supplemented with DME (250ml), F12 (250ml), 500ul of insulin (10mg/ml), 500ul EGF (10ug/ml) and 500ul of hydrocortisone (1mg/ml), and pen/strep. MDA-MB-157 and MDA-MB-231 cells were cultured in DME with 10% fetal bovine serum (FBS) and pen/strep. HCC38, HCC1806, and HCC1143 cells were cultured in RPMI with 10% fetal bovine serum (FBS) and pen/strep. HS578T cells were cultured in DME with 10% fetal bovine serum (FBS), insulin (100ng/ml), and pen/strep. BT549 cells were cultured in RPMI with 10% fetal bovine serum (FBS), insulin (100ng/ml), and pen/strep. MDA-MB-468 cells were cultured in RPMI with 10% fetal bovine serum (FBS), L-glutamine, and pen/strep. SUM159 and SUM149 cell lines were cultured in F12 with 5% inactivated calf serum (IFS), 1ug/ml hydrocortisone, insulin (5ug/ml), and pen/strep. All cell lines were maintained in sub-confluent conditions and media was replenished every 48 hours.

Plasmid constructs and virus production

pLenti-CRISPR-Cas9 V2 (Addgene 52961) constructs were produced as previously described (28). Spacer guide sequences used for the constructs were: sgZEB1 (GAGCACTTAAGAATTCACAG; kindly provided by Yun Zhang, Whitehead Institute for Biomedical Research, Cambridge, MA), sgTP63 (CCGTGACGCTGTTCTGCGCG), non-cutting controls sgNC1 and sgNC2 (GTGTCGGATTCCGCCGCTTA and CTATCTCGAGTGGTAATGCG, respectively; kindly provided by Jordan A. Krall, Whitehead Institute for Biomedical Research, Cambridge, MA).

FACS analyses and sorting

Cells were prepared for sorting following trypsinization and quenching in DME supplemented with 10% IFS. Briefly cells were counted and washed with ice-cold PBS^- (no calcium or magnesium) containing 2% IFS. For FACS analyses, cells were resuspended in ice-cold PBS^- + 2%IFS at 1x106 cells per 100ul. FACS antibodies were added with a 1:100 dilution, mixed gently and incubated in the dark on ice for a minimum of 30 minutes. 800ul of ice-cold PBS^- + 2%IFS was added to each tube and mixed gently then centrifuged at 300g for 5 minutes. Cells were resuspended in 500ul of ice-cold PBS^- + 2%IFS and passed through a 40μM filter prior to analysis. Cells were analyzed on a BD Biosciences LSRII or LSRFortessa instrument using FACSDiva software (BD) for data capture and FlowJo (FlowJo, LLC) software for analysis. FACS sorting was performed using the same protocol for cell preparation and then separated using a BD Biosciences FACSAria instrument with FACSDiva software. After sorting, cells were centrifuged and cultured in their respective medium. All FACS sorted cell populations can be obtained by starting with the top and bottom 2.5% of the ITGB4 histograms for the first round, followed by 5% and 25% cutoffs for the second and third rounds of sorting, respectively. After the third sort, the populations described in the results section readily maintained their ITGB4 status and were used for the indicated analyses. Cell lines were allowed a period of at least two passages after the final sort to minimize non-specific differences between the populations attributed to the process of sorting. Antibodies used for FACS sorting and analyses: ITGB4-efluor 660 (Affymetrix; 50-1049-82), CD44-efluor 450 (Affymetrix; 48-0441-82), CD24-FITC (BioLegend; 327806), CD24-PE-Cy7 (Affymetrix; 25-0247-42), EpCAM-PE (Affymetrix; 12-9326-42), ITGA6-FITC (Affymetrix; 11-0495-82). ANXAV-FITC (eBioscience 88-8005-72), DAPI, or cell scatter profiles were used for live-dead analyses. Aldefluor assays to measure ALDH activity were conducted essentially as described by the manufacturer (Stemcell Technologies; 01700), with the exception of using 2.5μl of activated Aldefluor reagent per 250μl reaction and 2.5μl of 100μM DEAB diluted in DMSO per 250μl reaction as a negative control. Reactions were allowed to proceed for 30 minutes at 37C then washed, placed on ice, and immediately analyzed.

Proliferation and tumorsphere assays

Proliferation assays were conducted in 96-well plates using CyQuant (Thermo Fisher, Inc.; C7026), according to the manufacturer recommendations, to measure DNA content in each well during a four-day time course. The first day after seeding was counted as T=0 and used for normalization of values obtained from plates collected at subsequent time points. At each time point, media was discarded and 96-well plates were frozen at -80C until analyzed. All plates were analyzed at the same time to minimize experimental variation. Tumorsphere assays were conducted using the MammoCult Medium Kit (Stemcell Technologies; 05620) supplemented with 4ug/ml heparin, 0.48ug/ml hydrocortisone, pen/strep, and 1% methylcellulose. 100 cells were seeded per replicate with 10 replicates per condition and spheres were counted on day ten.

Western blot analyses

To prepare protein, cells were washed twice with ice-cold PBS^- and placed on ice. Ice-cold lysis buffer [50mM Tris pH7.5, 150mM NaCl, 10mM EDTA pH8.0, 0.2% Sodium Azide, 50mM NaF, 0.5% NP40, proteinase inhibitor cocktail (1:100; Sigma P8340), phosphatase cocktail 2 (1:100; Sigma P5726), and phosphatase inhibitor cocktail 3 (1:100; Sigma P0044)] was added to each plate with a volume of 450ul per 15cm dish. Cells were scraped into the lysis buffer on ice then flash frozen on dry ice. Prior to analysis,

cell lysates were centrifuged at 13,000g for 15 minutes and supernatants were used for western blot analyses. Western blots were run using 1XMOPS buffer and NuPAGE Novex 4-12% Bis-Tris Gels as described by the manufacturer (Thermo Fisher Scientific, Inc) and transferred to PVDF membranes. Membranes were blocked with 5% non-fat milk for 45 minutes, washed three times in 0.1% TBST pH7.4 prior to overnight incubation with primary antibodies diluted in 0.1% TBST pH7.4 with 5% bovine serum albumin. Blots were washed and incubated with HRP conjugated secondary antibodies, washed, and visualized using Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific, Inc) on autoradioraphy film (LabScientific, Inc). Antibodies and conditions used for analysis: ITGB4 (Sigma HPA036348, 1:1000), CDH1 (Cell Signaling Technologies (CST) 3195, 1:1000), CDH2 (Fisher Scientific BDB610921, 1:1000), CDH3 (CST 4061, 1:1000), TWIST1 (Abcam ab50887, 1:500), SNAIL (CST 3879, 1:1000), (ZEB1 (CST 3396, 1:1000), FN1 (BD Biosciences 610078, 1:1000), VIM (CST 5741, 1:1000), β-catenin (Sigma C2206, 1:1000), p63 4A4 (Santa Cruz Biotechnology sc-8431, 1:1000), pan-AKT (CST 4685, 1:1000), phospho-Akt ser473 (CST 4060), ERK1/2 (CST 4695, 1:1000), phospho-ERK1/2 (CST 4370, 1:1000), AXL (CST 8661, 1:1000), FGFR1 (CST 9740, 1:1000), ALDH1A3 (Fisher Scientific PA5-29188, 1:5000), COXIV (CST 4850, 1:5000) GAPDH (CST 8884, 1:5000).

Quantitative real-time PCR primers

Primers used for analysis: TWIST1 F 5'-TGCGGAAGATCATCCCCACG and R 5'-GCTGCAGCTTGCCATCTTGGA, TWIST2 F 5'-GCAAGATCCAGACGCTCAAGCT and R 5'-ACACGGAGAAGGCGTAGCTGAG, SNAI1 F 5'-CTGGGTGCCCTCAAGATGCA and R 5'-CCGGACATGGCCTTGTAGCA, SNAI2 F 5'-TACCGCTGCTCCATTCCACG and R 5'-CATGGGGGTCTGAAAGCTTGG, ZEB1 F 5'- TGCACTGAGTGTGGAAAAGC and R 5'-TGGTGATGCTGAAAGAGACG, ZEB2 F 5'-AATGCACAGAGTGTGGCAAGGC and R 5'-CTGCTGATGTGCGAACTGTAGG, CDH1 F 5'-TTGCACCGGTCGACAAAGGAC and R 5'-TGGATTCCAGAAACGGAGGCC, CDH2 F 5'-TGTCGGTGACAAAGCCCCTG and R 5'-AGGGCATTGGGATCGTCAGC, CDH3 F 5'-CAGGTGCTGAACATCACGGACA and R 5'-CTTCAGGGACAAGACCACTGTG, FN1 F 5'-GAGAATGGACCTGCAAGCCCA and R 5'-AGTGCAAGTGATGCGTCCGC, VIM F 5'-ACCCGCACCAACGAGAAGGT and R 5'-ATTCTGCTGCTCCAGGAAGCG, HPRT1 F 5'-CTCCGTTATGGCGACCC and R 5'-CACCCTTTCCAAATCCTCAG, GUSB F 5'-CTCATTTGGAATTTTGCCGATT and R 5'-CCGAGTGAAGATCCCCTTTTTA.

Bioinformatic analyses

Primary patient survival correlations for TNBC and molecular basal subtype breast cancer were performed using normalized gene expression data from The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (29) was obtained from the publicly available European Genome-phenome Archive (IDs EGAD00010000210 and EGAD0001000021) (30). Paired METABRIC clinical feature data including age, grade, stage, and chemotherapy treatment from Synapse: METABRIC Data for Use in Independent Research (syn1688369, 2014). Overall survival was capitated at 5 years as a censored event and all data was included for all analyses. Patients with 'triple-negative' breast cancer were defined as study-reported negative for the estrogen receptor immunohistochemistry (IHC), progesterone receptor expression, and HER2-receptor IHC or SNP data when IHC was unavailable; those patients with one or more missing components were excluded. Patients with 'basal-like' breast cancer were defined by study-provided PAM50 determination. All microarray data processing and statistical analyses were performed in R version 3.1.3. To evaluate the prognostic association of ITGB4 in TNBC and basal-like breast cancer,

we stratified samples into quartiles based on ITGB4 expression and compared the highest expression quartile with the remaining quartiles. Cox proportional hazards model was calculated for ITGB4 highest expression quartile versus remaining quartiles alone ('Univariate') or in a combined model with three clinicopathologic characteristics (age, tumor stage at diagnosis, tumor grade; 'Multivariate') were determined using the 'coxph' package. Kaplan-Meier curves and log likelihood ratio statistic were calculated using the plot_km function in the 'packHV' package.

Secondary validation of METABRIC data was performed using the kmplot tool (31). Comparisons for ITGB4 and relapse-free survival were performed using affymetrix probe 204990_s_at, capitated at 5 years (v2014). In the dataset, there was 4142 total cases, of which chemotherapy status was available for 173 TNBC and 399 molecular basal patients that were selected for comparison. ITGB4 high and ITGB4 low patient cohorts were assigned using an unbiased self-optimizing algorithm to determine the threshold necessary to observe greatest degree of separation for survival probabilities between patients in the ITGB4 high and ITGB4 low subgroups. Lung adenocarcinoma, ovarian and gastric carcinoma analyses were performed using the same tool, probeset, self-optimizing algorithm for ITGB4 high and ITGB4 low cohort selection, and 5-year capitation for progression-free survival parameters. In the lung adenocarcinoma dataset (v2015, 866 cases) (32), 461 patients for whom progression-free survival was known were used for the reported analyses. In the serous ovarian cancer dataset (v2015, 1144 cases) (33), only stage 4 patients (143) demonstrated a statistically significant difference in progression-free survival and were thus used for the reported analyses. In the gastric cancer dataset (34), patients with known progression-free survival data (641) were selected for the reported analyses.

SI Appendix Figure Legends

Figure S1. Identification of integrin-beta4 (ITGB4) as a marker that can segregate epithelial and mesenchymal mammary epithelial cells. **A.** Morphological appearance of the more epithelial HMLE and more mesenchymal NAMEC8 cell lines. **B.** Fold-change in RNAseq values comparing the expression of integrins a2 (ITGA2), a6 (ITGA6), b4 (ITGB4), b6 (ITGB6), and b8 (ITGB8) in the HMLE and NAMEC8 cell lines. **C.** Meta-analysis of ITGB4 mRNA expression from Taube, et al. in HMLE cells and HMLE cells induced to undergo an EMT in response to TGF-β stimulation, ectopic Twist1, Goosecoid (Gsc), or Snail expression constructs or siRNA targeting E-cadherin (CDH1). **D.** CD44, CD24 and ITGB4 FACS profiles of HMLE cells. Arrows indicate naturally arising mesenchymal subpopulations. **E.** ITGB4 FACS histograms of HMLE cells before or after integration of ectopic TWIST1 or SNAIL expression constructs. **F.** Morphological appearance of HMLE-TWIST or HMLE-SNAIL cells sorted for high or low levels of ITGB4. **G.** Western blots for EMT markers in HMLE cells compared to HMLE-TWIST or HMLE-SNAIL cells sorted for high or low levels of ITGB4. CDH1, E-cadherin; CDH2, N-cadherin; CDH3, P-cadherin; BCAT, b-catenin.

Figure S2. Segregation of distinct mesenchymal mammary epithelial cells and identification of non-canonical EMT-associated genes that were differentially expressed by the isolated populations. **A.** CD44, CD24 and ITGB4 FACS profiles of the more mesenchymal NAMEC8 cells. **B.** Overlayed FACS histograms of ITGB4 after isolation of NAMEC8 ITGB4^{hi} and ITGB4^{lo} populations. **C.** Heatmap of genes differentially expressed NAMEC8 ITGB4^{hi} and ITGB4^{lo} populations. **D.** Morphological appearance of parental NAMEC8 cells and sorted ITGB4^{hi} and ITGB4^{lo} subpopulations. **E.** Fold-change for canonical EMT-associated gene expression as determined by comparison of RNAseq data from the NAMEC8 ITGB4^{hi} and ITGB4^{lo} subpopulations. **F.** Heatmap of the more epithelial HMLE vs more mesenchymal NAMEC8 RNAseq values for a subset of non-canonical EMT-associated genes which reflect the changes

observed in more epithelial NAMEC8 ITGB4^{hi} vs more mesenchymal NAMEC8 ITGB4^{lo} RNAseq comparisons. **G and H.** Log2 values for the top 15 epithelial (HMLE) and mesenchymal (NAEMC8) enriched genes represented in panel F.

Figure S3. Unsupervised hierarchical clustering of previously reported RNAseq data from the more epithelial HMLE and more mesenchymal NAMEC8 cell lines with 49 common breast cancer cell lines. Triple-negative breast cancer (TNBC) basal-like (BL1/BL2) and mesenchymal/mesenchymal stem-like (M/MSL) lines are indicated.

Figure S4. CD44/CD24 versus CD44/ITGB4 FACS profiles and morphological appearance of HMLE, NAMEC8, and a panel of common triple-negative breast cancer cell lines. A. CD44 and CD24 FACS profiles for eight TNBC cell lines. **B.** CD44 vs ITGB4 FACS profiles for eight TNBC cell lines. Basal-like (BL1/BL2); mesenchymal/mesenchymal stem-like (M/MSL). **C.** Morphological appearance of ten TNBC cell lines and the more epithelial HMLE versus more mesenchymal NAMEC8 cell lines.

Figure S5. Selected canonical and non-canonical EMT-associated gene expression in SUM159 ITGB4^{hi} and ITGB4^{lo} cells and comparison of differentially expressed EMT-associated genes with those identified by comparing NAMEC8 ITGB4^{hi} and ITGB4^{lo} cells. A. Fold-change of RNAseq values for EMT-associated genes in the SUM159 ITGB4^{hi} and ITGB4^{lo} cell populations. **B.** Quantitative real-time PCR validation of results for EMT-associated gene expression comparing the SUM159 ITGB4^{hi} vs ITGB4^{lo} cells. **C.** Western blots demonstrating differential expression of TAp63α and AXL in the SUM159 ITGB4^{hi} vs ITGB4^{lo} cells with GAPDH used as a loading control. **D-E.** Log2 ratio values of EMT-associated genes (differentially expressed in HMLE vs NAMEC8 RNA-seq) that were commonly upregulated in the more epithelial and more mesenchymal subpopulations from SUM159 ITGB4^{hi} vs ITGB4^{lo} and NAMEC8 ITGB4^{hi} vs ITGB4^{lo} comparisons, respectively.

Figure S6. Characterization of ALDH1A1, ALDH1A3, and ALDH7A1 expression in ITGB4^{hi} and ITGB4^{lo} SUM159 cells and in a panel of more epithelial and more mesenchymal TNBC cells and comparison of functional abilities of ITGB4^{hi} and ITGB4^{lo} SUM159 cells. A. Normalized RNAseq read counts for cancer stem cell-associated aldehyde dehydrogenase (ALDH1A1, ALDH1A3, and ALDH7A1) gene expression. **B.** Meta-analysis of ALDH1A3 gene expression in basal-like (BL1/BL2) and mesenchymal/mesenchymal stem-like (M/MSL) TNBC cells. **C.** Western blots for ALDH1A3 and COXIV in ten TNBC cell lines and the more epithelial HMLE versus more mesenchymal NAMEC8 (N8) cell lines. **D.** Aldefluor assay FACS analyses directly comparing the relative aldehyde dehydrogenase (ALDH) activity within more epithelial (top row; BL1/BL2) and more mesenchymal (bottom row; M/MSL) TNBC cell lines at the same time point. N,N-diethylaminobenzaldehyde (DEAB), was used as an inhibitor of ALDH activity. **E.** Proliferation rates and tumorsphere forming efficiency of two independently derived SUM159 ITGB4^{hi} and ITGB4^{lo} cell line pairs.

Figure S7. Reduction of ITGB4 cell surface abundance as a result of HRAS-dependent induction of EMT and characteristics of the NAMEC populations used for tumor initiation assays. A. Morphological appearance of HMLE cells expressing low (HMLER^{lo}), medium (HMLER^{med}), or high (HMLER^{hi}) levels of HRAS^{G12V}. **B-D**. Quantitative real-time PCR results for EMT-associated gene expression comparing the HMLE, HMLER^{lo}, and HMLER^{med} cell lines. CDH1, E-cadherin; CDH2, N-cadherin; CDH3, P-cadherin. **E.** FACS profiles of HMLE, HMLER^{lo}, and HMLER^{med} cell lines comparing the level of HRAS-IRES-GFP expression and ITGB4 cell-surface abundance. **F and G.** CD44, CD24, and ITGB4 FACS profiles of NAMEC1 and NAMEC5

cell lines. **H.** Log2 values for NAMEC5/NAMEC1 RNAseq comparisons of ITGB4 and four genes that were also correlated inversely with ITGB4 in HMLE vs NAMEC8, NAMEC8 ITGB4^{hi} vs ITGB4^{lo}, and SUM159 ITGB4^{hi} vs ITGB4^{lo} RNAseq analyses. **I.** Morphological appearance of NAMEC1R, NAMEC5R and NAMEC8R cells in vitro.

Figure S8. Characteristics and gene expression profiles of ITGB4^{hi} and ITGB4^{lo} MDA-MB-231 cells. A. ITGB4, CD44, and CD24 FACS profiles of MDA-MB231 cells and their ITGB4^{hi} and ITGB4^{lo} subpopulations. **B.** Morphological appearance of MDA-MB-231 cells and isolated ITGB4^{hi} and ITGB4^{lo} subpopulations in vitro. **C.** Normalized RNAseq fold-change values for canonical EMT-associated gene expression comparing the ITGB4^{hi} and ITGB4^{lo} MDA-MB-231 cells. **D.** RNAseq Log2 values for comparisons of genes upregulated in the more epithelial HMLE or more mesenchymal NAMEC8 cells that were differentially expressed in the comparison of RNAseq values for MDA-MB-231 ITGB4^{hi} and ITGB4^{lo} cells.

Figure S9. Clinical correlations between ITGB4 mRNA expression and patient relapse- and progression-free survival in triple-negative subtype breast cancer, lung adenocarcinoma, stage 4 ovarian cancer, and gastric cancer. Kaplan-Meier survival curves indicating correlations between ITGB4 expression and relapse- or progression-free survival in (A) TNBC, (B) lung adenocarcinoma, (C) stage 4 ovarian, and (D) gastric cancer patients. HR, hazard ratio; p, logrank p-value. The number of patients for each analysis and those remaining at risk during the time course are shown below each survival curve.

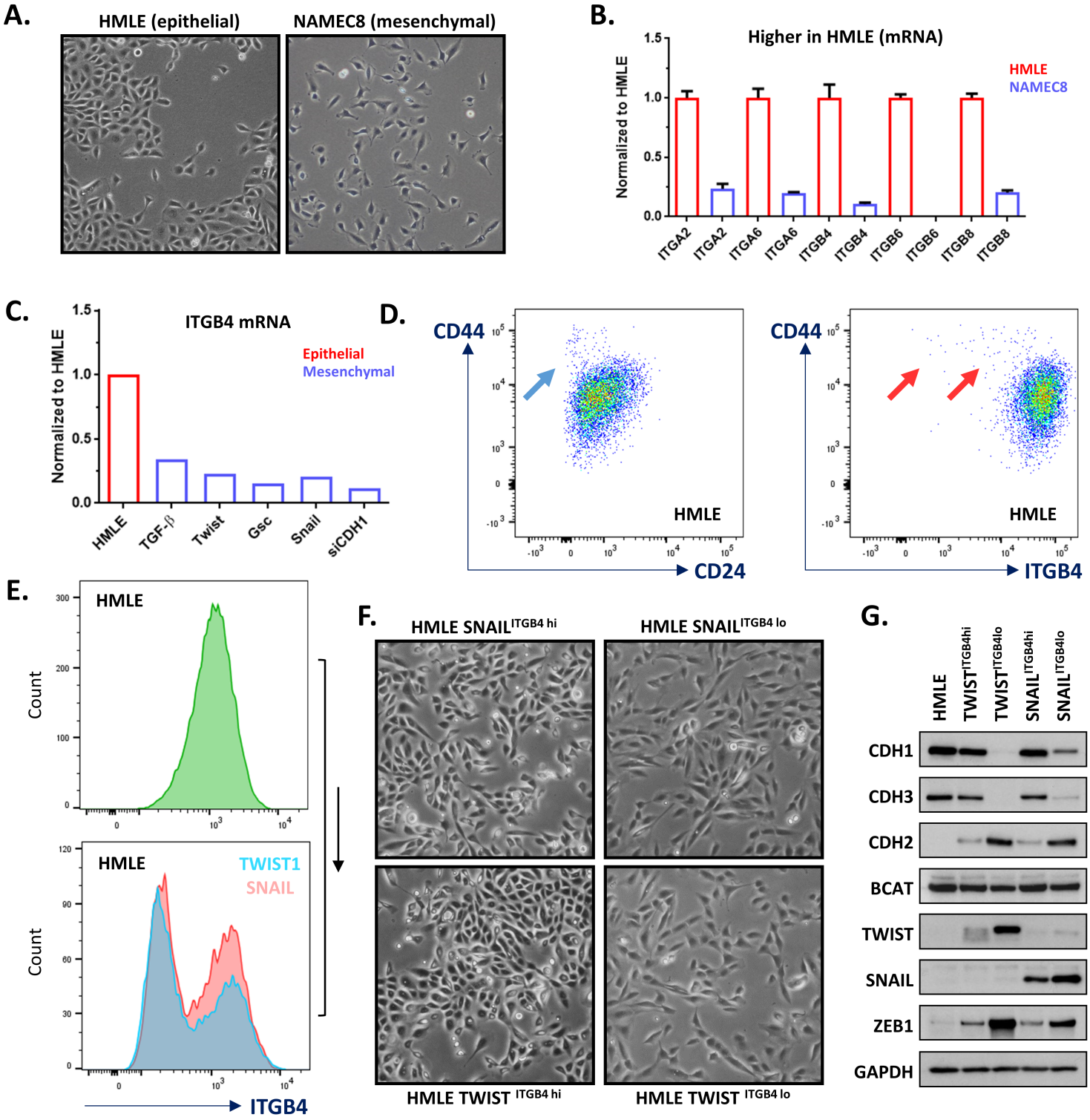
Figure S10. FACS and western blot analyses of SUM159 ITGB4^{hi} and ITGB4^{lo} cells that were transduced with CRISPR Cas9-sg or constitutive ITGB4 expression constructs. A-C. FACS histograms of ITGB4 on SUM159 ITGB4^{hi} (High; Hi) and ITGB4^{lo} (Low; B4^{lo}) cells and their derivatives expressing Cas9 and spacer guides (sg) for ZEB1, TP63, or non-cutting controls (NC) as indicated, and ITGB4^{lo} cells harboring a constitutive ITGB4 expression construct (low B4^{OE}; B4^{lo}B4^{OE}). **D**. Western blots for EMT- and RAS-signaling-associated markers in the cell lines represented in C.

References

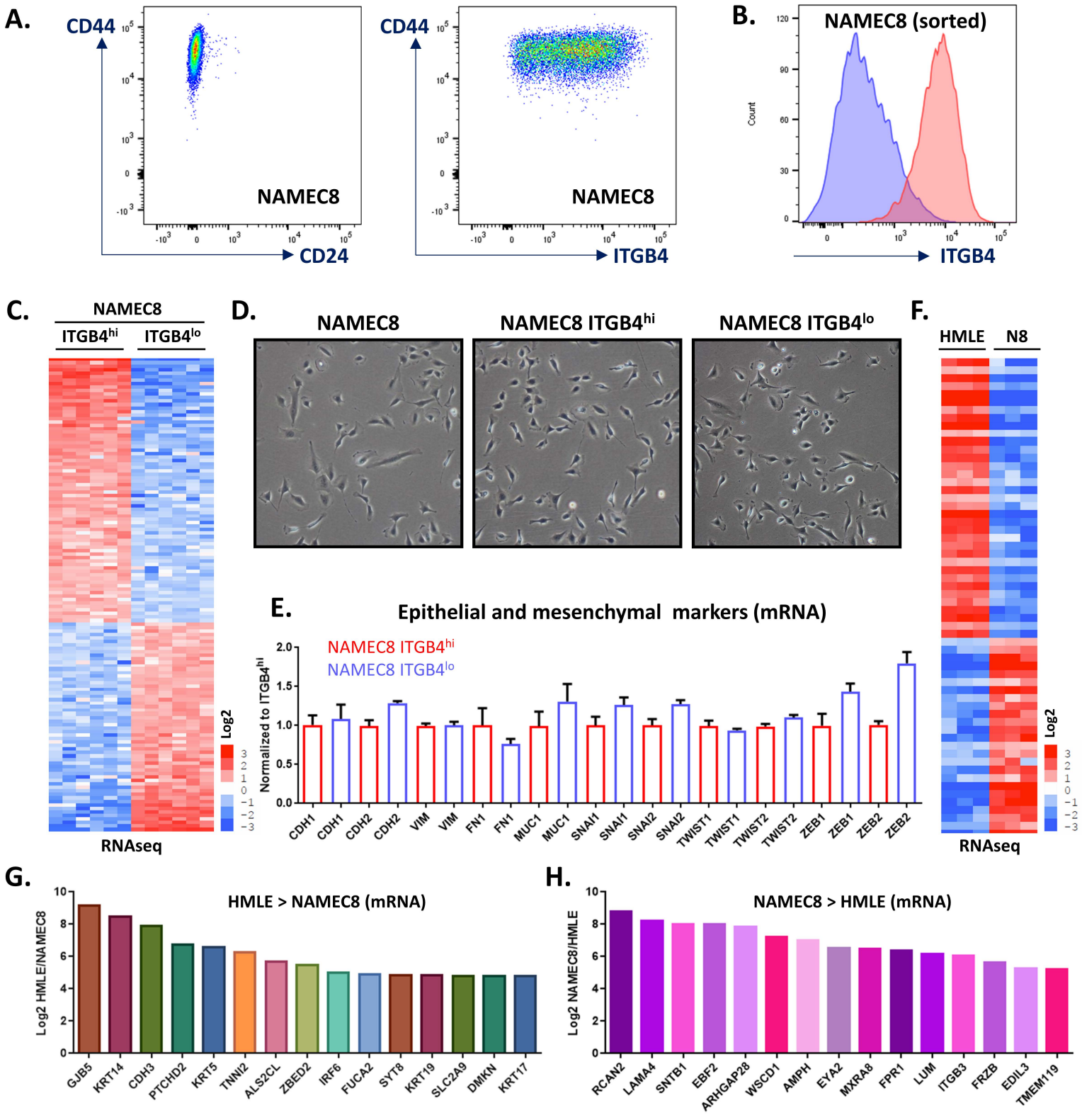
1. Guo W*, et al.* (2012) Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* 148(5):1015-1028.
2. Ye X*, et al.* (2015) Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature* 525(7568):256-260.
3. Elenbaas B*, et al.* (2001) Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev* 15(1):50-65.
4. Tam WL*, et al.* (2013) Protein kinase C alpha is a central signaling node and therapeutic target for breast cancer stem cells. *Cancer cell* 24(3):347-364.
5. Pattabiraman DR*, et al.* (2016) Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. *Science* 351(6277):aad3680.
6. Hynes RO (2002) Integrins: bidirectional, allosteric signaling machines. *Cell* 110(6):673-687.
7. Cerami E*, et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* 2(5):401-404.
8. Taube JH*, et al.* (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences of the United States of America* 107(35):15449-15454.

9. Mani SA*, et al.* (2008) The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* 133(4):704-715.
10. Barretina J*, et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391):603-607.
11. Lehmann BD*, et al.* (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation* 121(7):2750-2767.
12. Lehmann BD*, et al.* (2016) Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PloS one* 11(6):e0157368.
13. Honeth G*, et al.* (2008) The CD44+/CD24- phenotype is enriched in basal-like breast tumors. *Breast cancer research : BCR* 10(3):R53.
14. Chaffer CL*, et al.* (2013) Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. *Cell* 154(1):61-74.
15. Bianchini G, Balko JM, Mayer IA, Sanders ME, & Gianni L (2016) Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol*.
16. Prat A*, et al.* (2013) Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast cancer research and treatment* 142(2):237-255.
17. Prat A*, et al.* (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research : BCR* 12(5):R68.
18. Sarrio D*, et al.* (2008) Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer research* 68(4):989-997.
19. Haffty BG*, et al.* (2006) Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. *J Clin Oncol* 24(36):5652-5657.
20. Dent R*, et al.* (2007) Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research : an official journal of the American Association for Cancer Research* 13(15 Pt 1):4429-4434.
21. Barbieri CE & Pietenpol JA (2006) p63 and epithelial biology. *Exp Cell Res* 312(6):695-706.
22. Antony J*, et al.* (2016) The GAS6-AXL signaling network is a mesenchymal (Mes) molecular subtype-specific therapeutic target for ovarian cancer. *Science signaling* 9(448):ra97.
23. Gjerdrum C*, et al.* (2010) Axl is an essential epithelial-to-mesenchymal transition-induced regulator of breast cancer metastasis and patient survival. *Proceedings of the National Academy of Sciences of the United States of America* 107(3):1124-1129.
24. Wang C*, et al.* (2016) Gas6/Axl Axis Contributes to Chemoresistance and Metastasis in Breast Cancer through Akt/GSK-3beta/beta-catenin Signaling. *Theranostics* 6(8):1205-1219.
25. Vivanco MdM (2015) *Mammary stem cells : methods and protocols* (Humana Press, New York) pp x, 275 pages.
26. Ginestier C*, et al.* (2007) ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell stem cell* 1(5):555-567.
27. Opdenaker LM*, et al.* (2014) Immunohistochemical analysis of aldehyde dehydrogenase isoforms and their association with estrogen-receptor status and disease progression in breast cancer. *Breast Cancer (Dove Med Press)* 6:205-209.
28. Sanjana NE, Shalem O, & Zhang F (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* 11(8):783-784.
29. Curtis C*, et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486(7403):346-352.
30. Lappalainen I*, et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* 47(7):692-695.

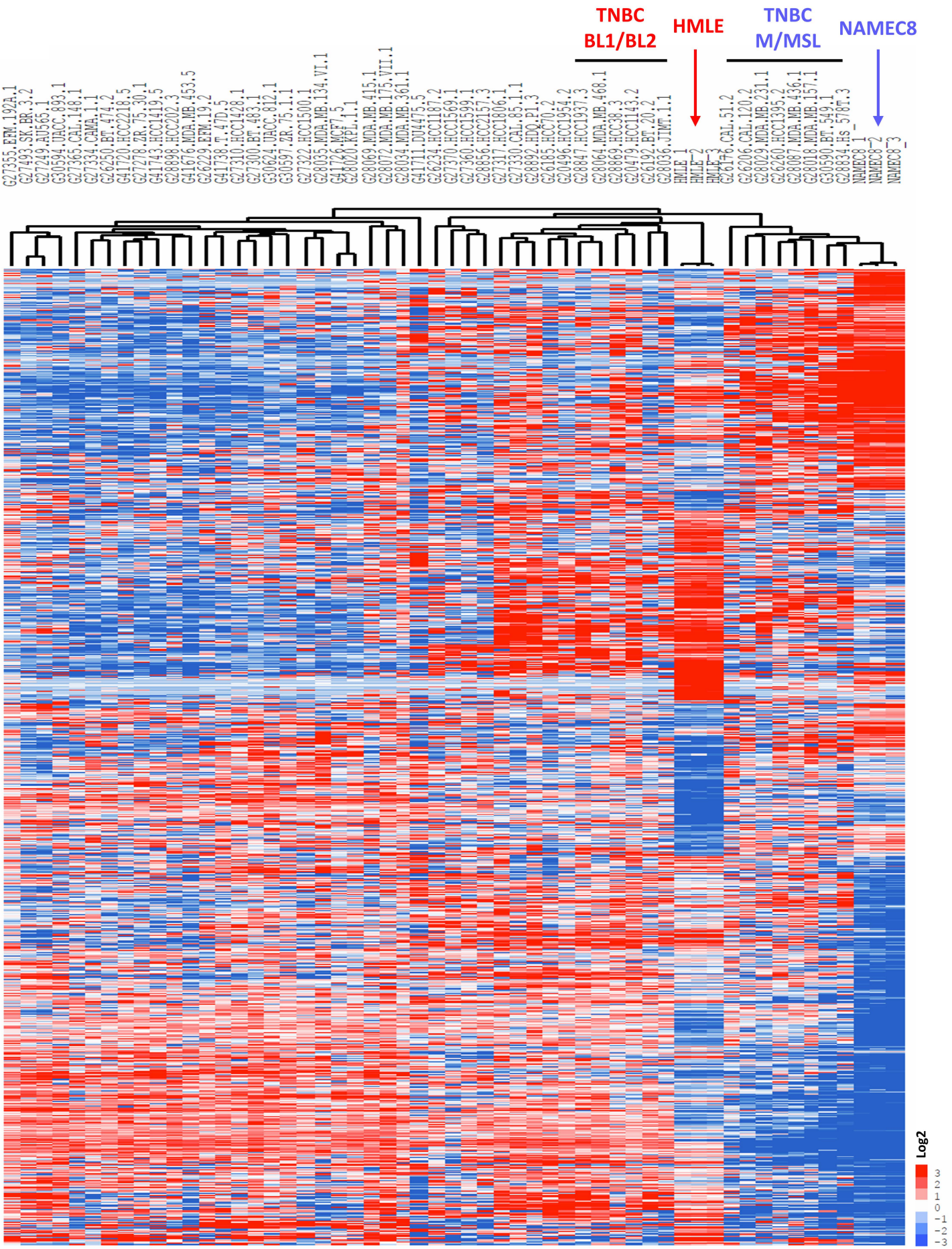
31. Gyorffy B*, et al.* (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment* 123(3):725-731.
32. Gyorffy B, Surowiak P, Budczies J, & Lanczky A (2013) Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PloS one* 8(12):e82241.
33. Gyorffy B, Lanczky A, & Szallasi Z (2012) Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr Relat Cancer* 19(2):197-208.
34. Szasz AM*, et al.* (2016) Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget*.



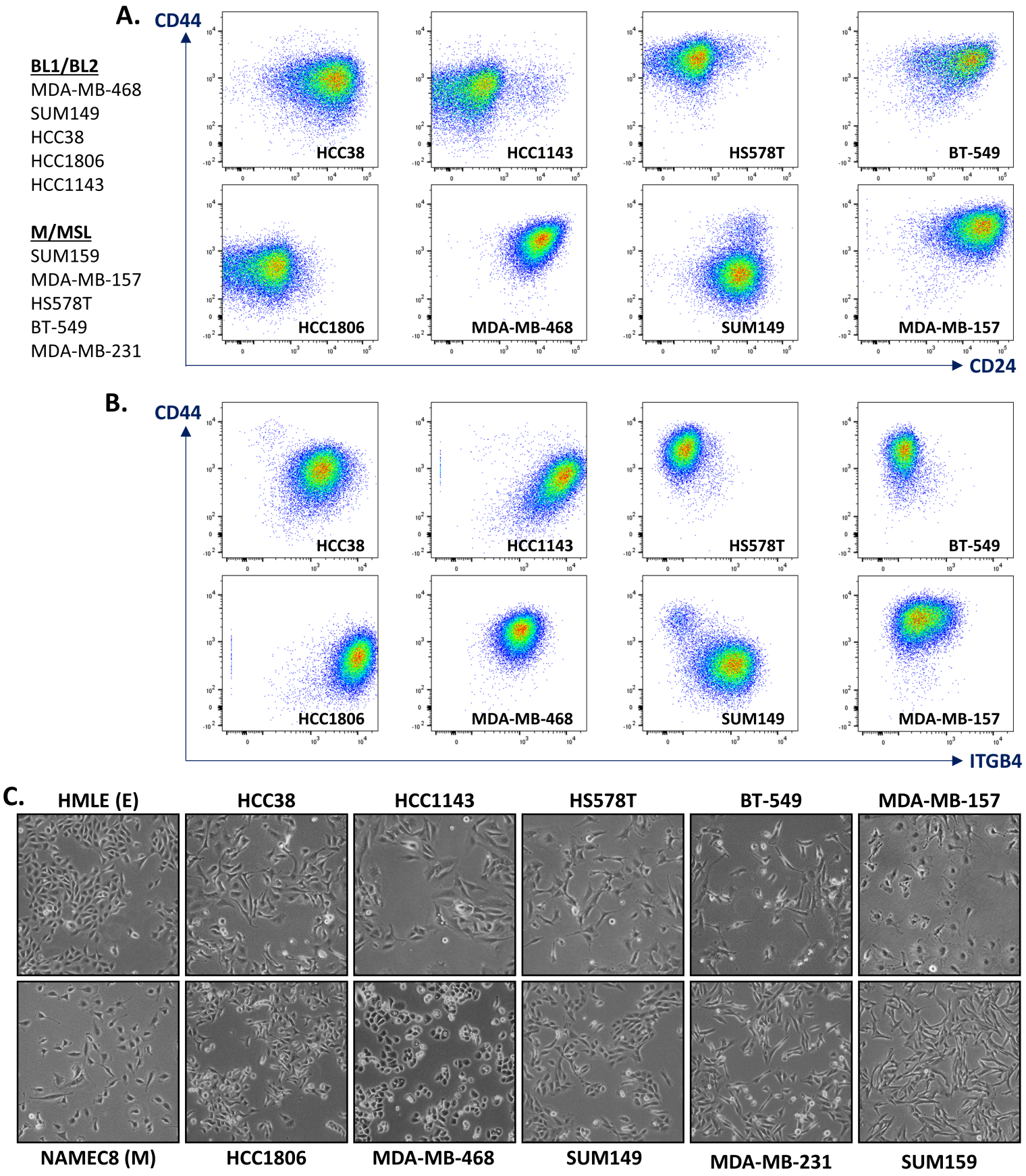
Supplemental Figure S1. Identification of integrin-beta4 (ITGB4) as a marker that can segregate epithelial and mesenchymal mammary epithelial cells.



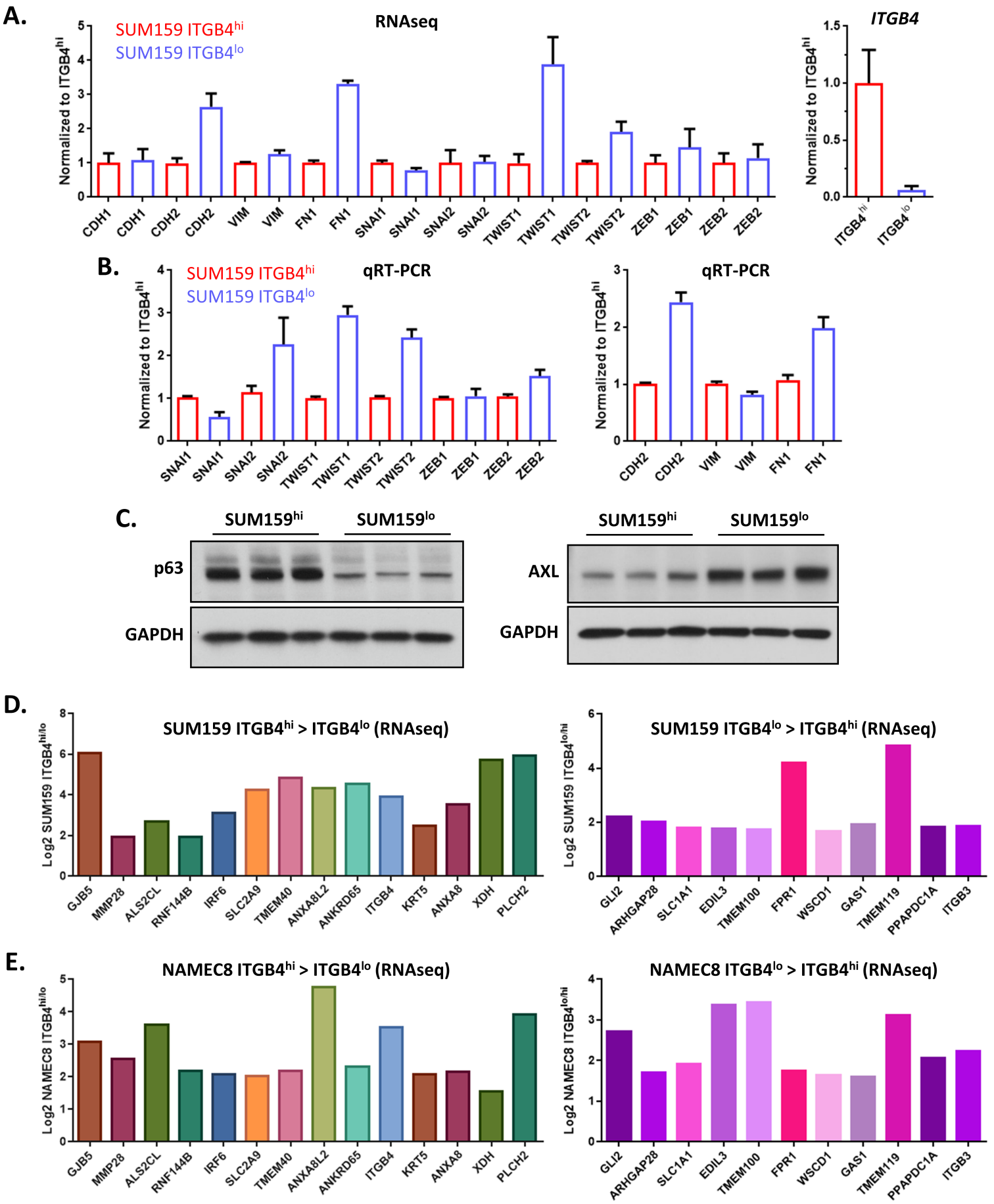
Supplemental Figure S2. Segregation of distinct mesenchymal mammary epithelial cells and identification of non-canonical EMT-associated genes that were differentially expressed by the isolated populations.



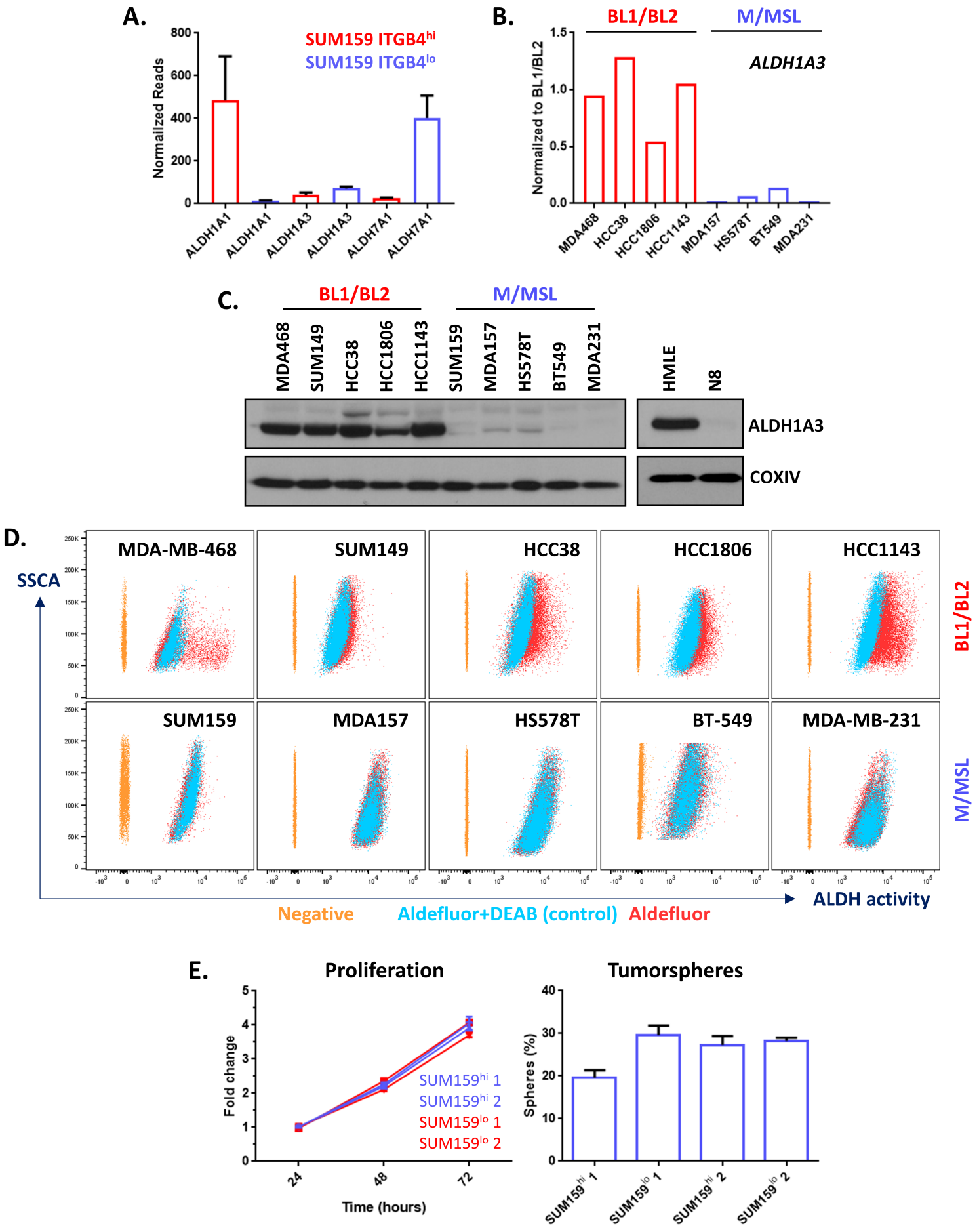
Supplemental Figure S3. Unsupervised hierarchical clustering of previously reported RNAseq data from the more epithelial HMLE and more mesenchymal NAMEC8 cell lines with 49 common breast cancer cell lines.



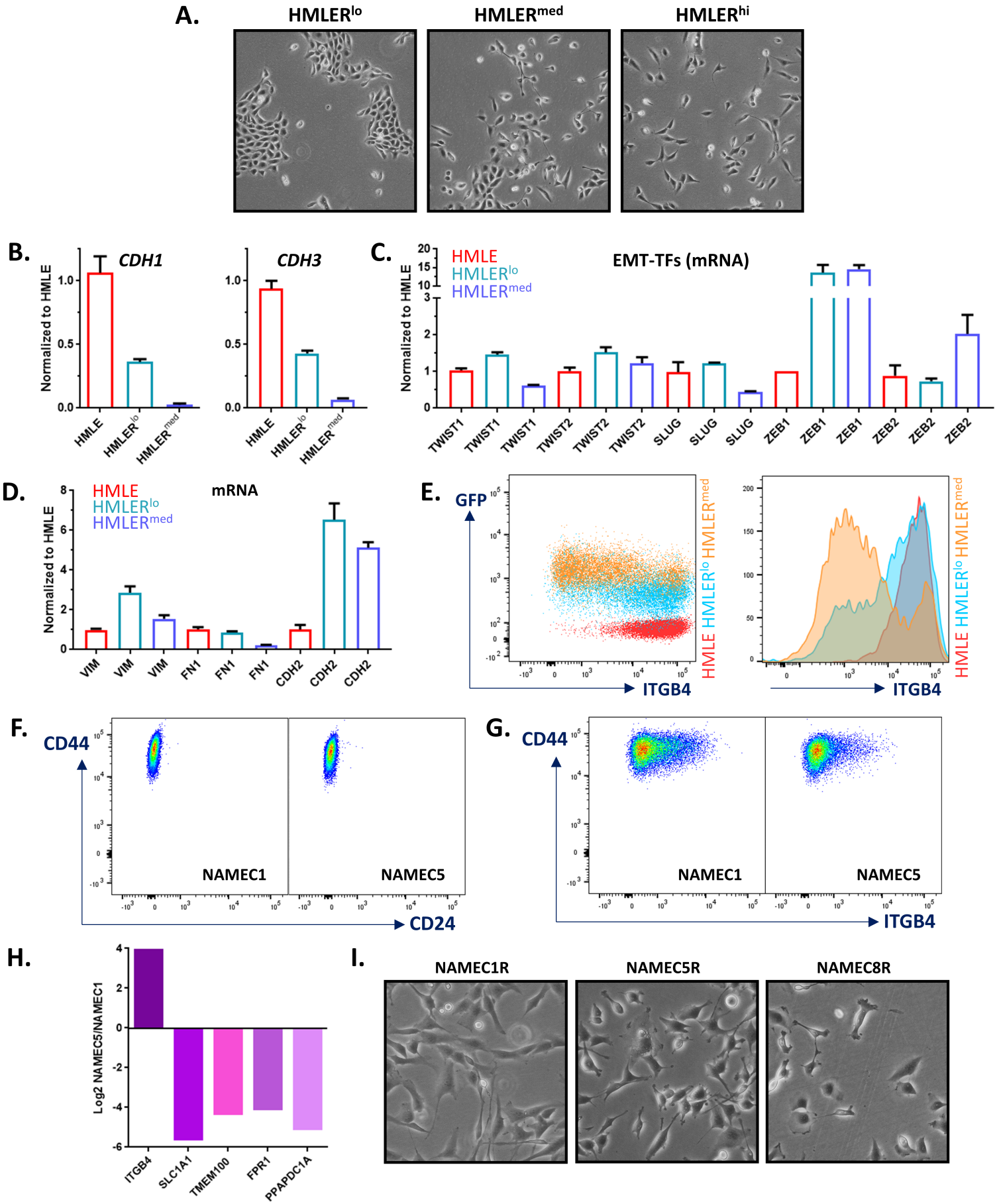
Supplemental Figure S4. CD44/CD24 versus CD44/ITGB4 FACS profiles and morphological appearance of HMLE, NAMEC8, and a panel of common triple-negative breast cancer cell lines.



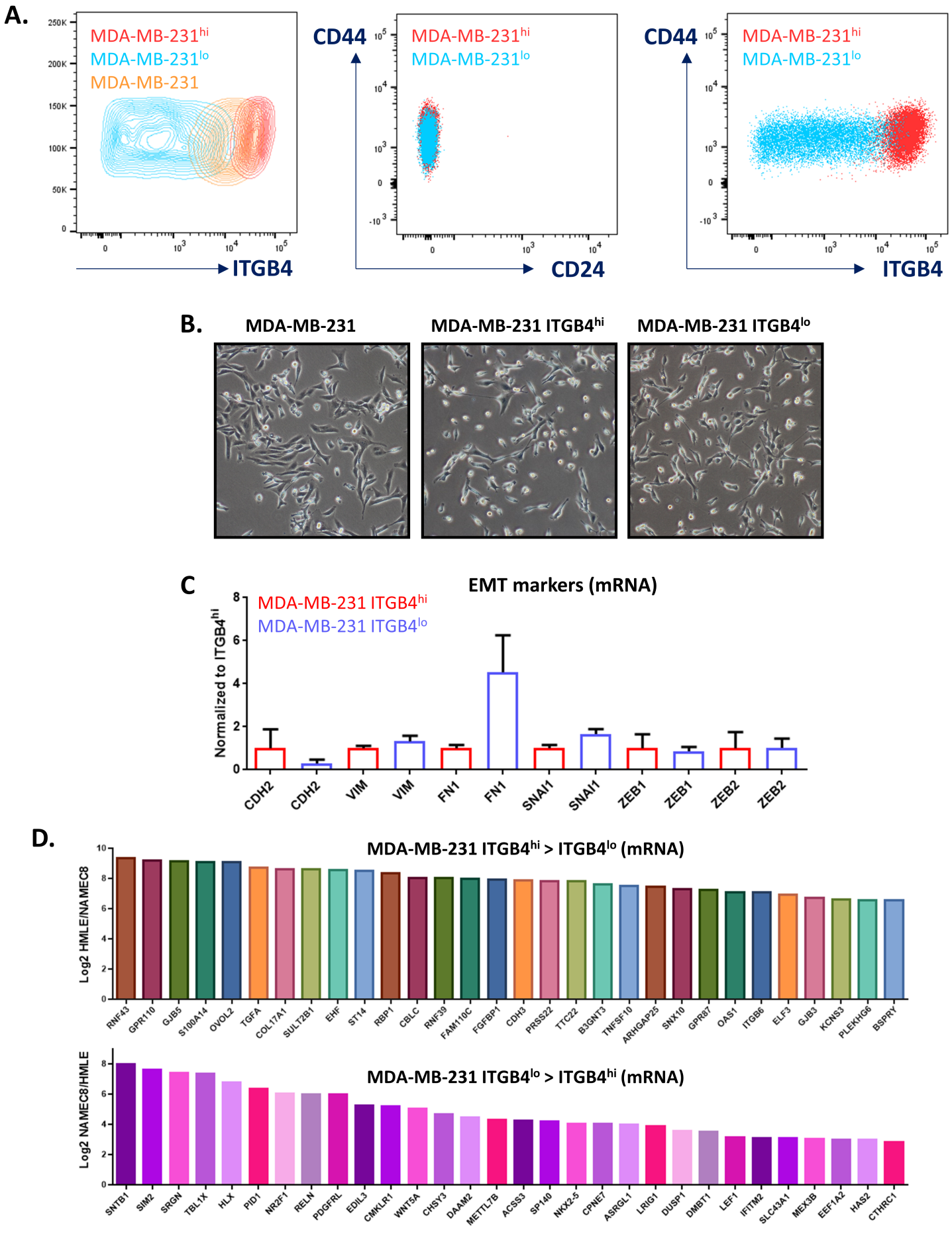
Supplemental Figure S5. Selected canonical and non-canonical EMT-associated gene expression in SUM159 ITGB4^{hi} and ITGB4^{lo} cells and comparison of differentially expressed EMT-associated genes with those identified by comparing NAMEC8 ITGB4^{hi} and ITGB4^{lo} cells.



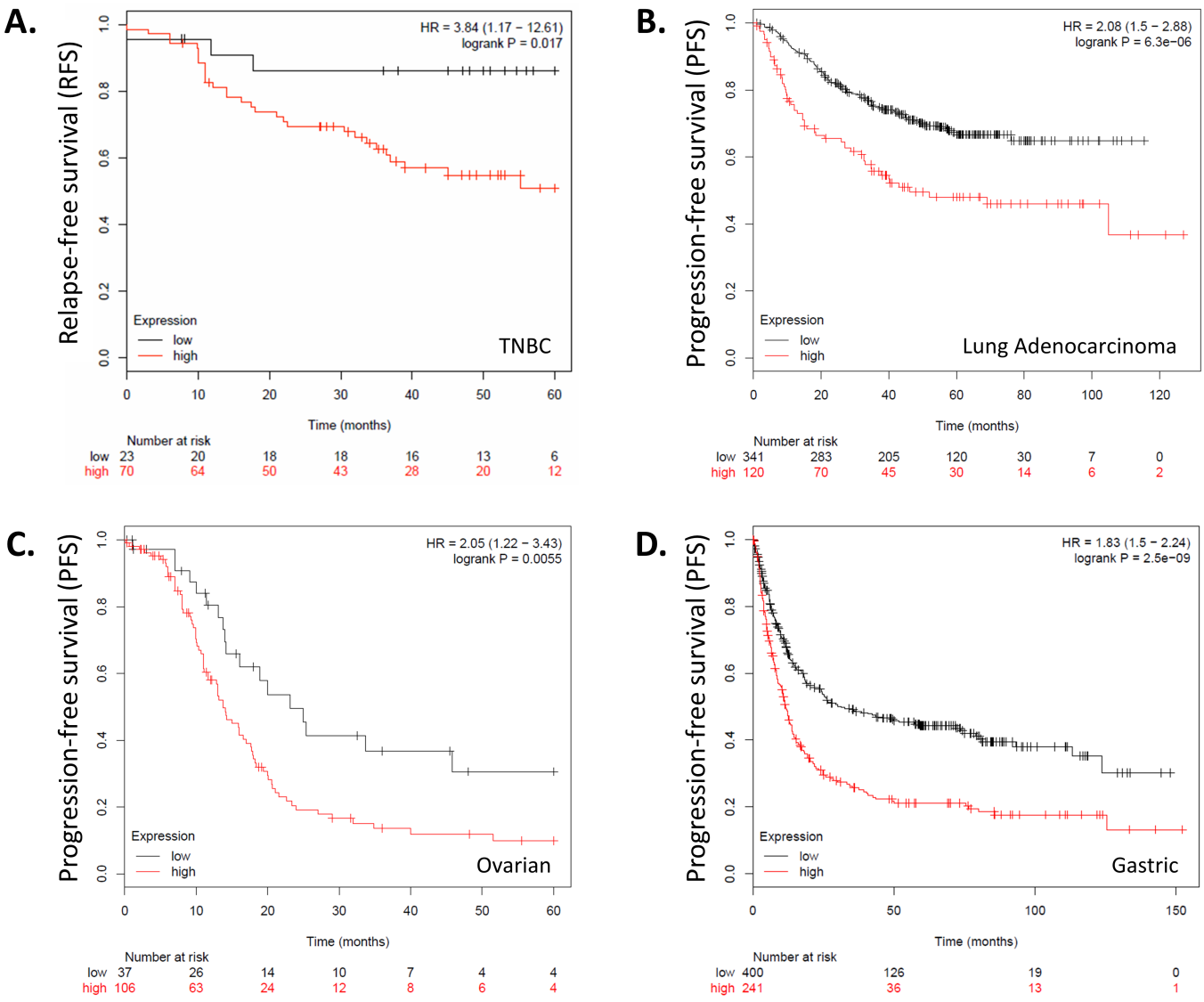
Supplemental Figure S6. Characterization of ALDH1A1, ALDH1A3, and ALDH7A1 expression in ITGB4^{hi} and ITGB4^{lo} SUM159 cells and in a panel of more epithelial and more mesenchymal TNBC cells and comparison of functional abilities of ITGB4^{hi} and ITGB4^{lo} SUM159 cells.



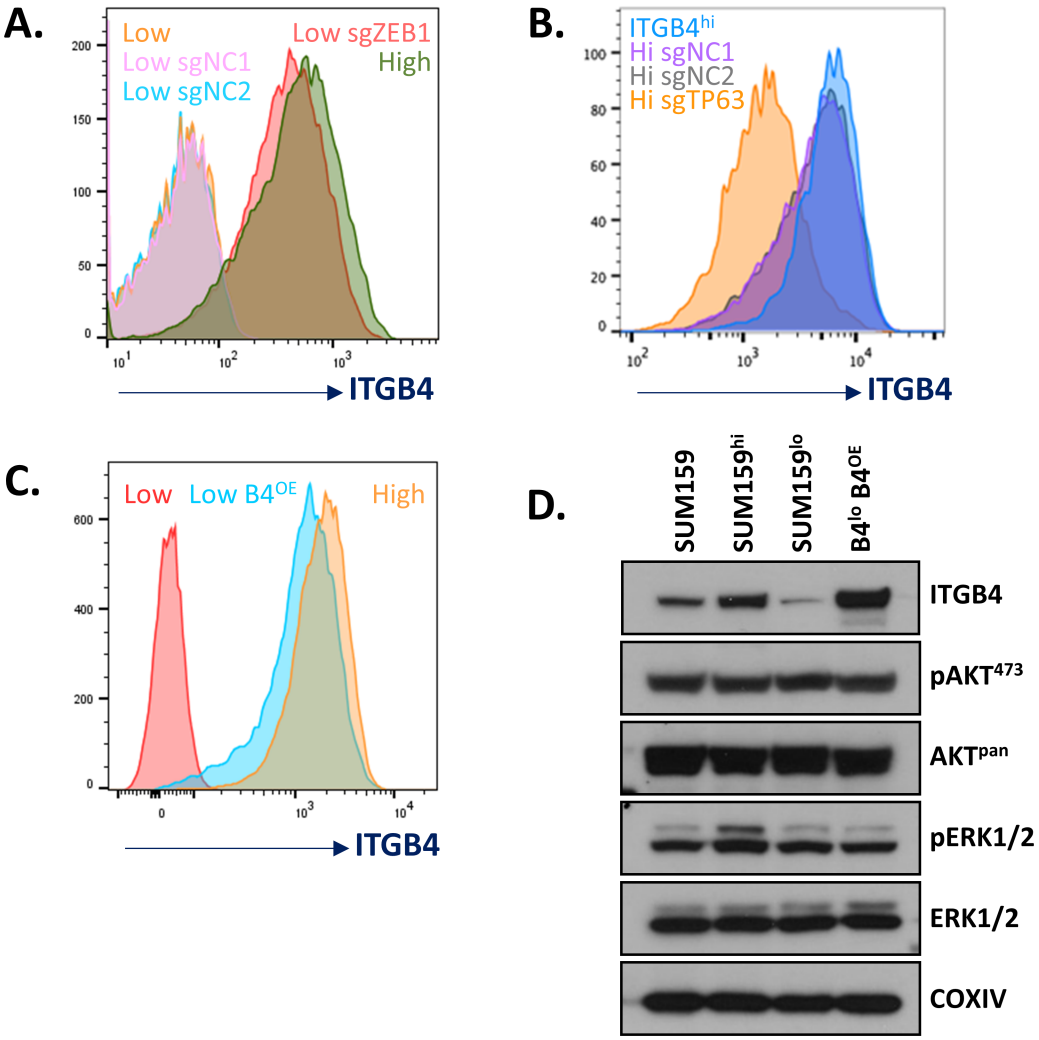
Supplemental Figure S7. Reduction of ITGB4 cell surface abundance as a result of HRAS-dependent induction of EMT and characteristics of the NAMEC populations used for tumor initiation assays.



Supplemental Figure S8. Characteristics and gene expression profiles of ITGB4^{hi} and ITGB4^{lo} MDA-MB-231 cells.



Supplemental Figure S9. Clinical correlations between ITGB4 mRNA expression and patient relapse- and progression-free survival in triple-negative subtype breast cancer, lung adenocarcinoma, stage 4 ovarian cancer, and gastric cancer.



Supplemental Figure S10. FACS and western blot analyses of SUM159 ITGB4^{hi} and ITGB4^{lo} cells that were transduced with CRISPR Cas9-sg or constitutive ITGB4 expression constructs.