SUPPORTING INFORMATION

Inventory of supporting information items

SUPPLEMENTAL MATERIALS AND METHODS

DNA and RNA array profiling in MB-series

DNA from 1001 tumors was hybridized to Affymetrix SNP 6.0 arrays to assay genomic copy number (Curtis *et al.*, manuscript in preparation). Following correction for allelic cross-talk, probe-level normalization and summarization as implemented in the aroma.affymetrix *crma* function (Bengtsson et al, 2009), log2 ratios were obtained by comparison against a pooled reference. Data were then segmented using the circular binary segmentation algorithm implemented in the DNAcopy Bioconductor package (Olshen et al, 2004), with thresholds adjusted to account for differences in cellularity.

Matched mRNA was hybridized to Illumina HT-12 BeadArrays for gene-expression analysis. Hybridized BeadArrays were processed using a custom R script, which performs quality assessment and adjustment for spatial artifacts with the BASH tool (Cairns et al, 2008), followed by bead-level summarization and outlier removal as implemented in the beadarray Bioconductor package (Dunning et al, 2007). Data were normalized using a variant of quantile normalization that accounts for the reliability of probes on the HT-12 array, wherein a target distribution was derived by removing all control probes or probes that did not meet strict annotation criteria. Samples were classified into the intrinsic subtypes based on the PAM50 gene list (Parker et al, 2009). Differential expression was performed on a subset of the normalized data that excluded any probes mapping to non-transcribed or repetitive regions. A probe-wise linear model was fitted to the data using the limma Bioconductor package (Smyth, 2004) with coefficients estimated for the contrasts of interest, with a coefficient for tissue-bank incorporated in the linear model. Here the primary contrast of interest was the comparison between Luminal B ZNF703 amplified or overexpressed *versus* neutral cases. After empirical Bayes' moderation of the variance, Benjamini-Hochberg adjusted p-values, log-ratios and log-odds were used to assess the evidence for differential expression.

Pathway-based analyses of MB-series tumor microarray expression data

Significantly altered pathways were identified in ZNF703-amplified versus neutral or ZNF703-upregulated versus neutral Luminal B tumors from the cohort of 1001

primary tumors using Ingenuity Pathways Analysis (Ingenuity® Systems, ww.ingenuity.com). Genes found to be differentially expressed ($p < 0.01$) with a fold change value of +/-1.5 in either of these contrasts were associated with canonical pathways in Ingenuity's Knowledge Base. The significance of the association between the data set and the canonical pathway was measured using Fisher's exact test. Network building tools from the Ingenuity knowledge base were used to identify associated molecules and relationships.

Immunohistochemistry (IHC)

Antigen retrieval from formalin-fixed paraffin-embedded single tumor or tissue microarray section was achieved by heating in Citrate buffer (pH 6.0) for twenty minutes. Staining was performed on a BondTM autostainer using polyclonal rabbit anti-ZNF703 (1:25) or anti-ERLIN2 (1:1000) antibodies (Atlas Antibodies, Sweden), and bound primary antibody was detected by a polymer-conjugated secondary antibody with DAB as a chromogen. The ZNF703 antibody was validated in werstern blots of MCF-7 cells with manipulation of ZNF703 expression, a panel of cell lines with known ZNF703 copy number and NIH-3T3 cells overexpressing human ZNF703. The ERLIN2 antibody has been extensively validated by the Human Protein Atlas (www.proteinatlas.org).

NIH 3T3 Transformation assay

hZNF703 and mKrasG12D4B cDNAs were cloned into pBabe-hygro (Addgene plasmid 1765) or pBabe-puro, respectively. Addgene plasmid was originated in the laboratory of Bob Weinberg.

Phoenix packaging cells and mouse NIH3T3 fibroblasts (ATCC) were cultured in DMEM (Invitrogen), supplemented with 10% fetal calf serum (Hyclone), 100 U/mL penicillin, and 100 µg/mL streptomycin (Invitrogen). Phoenix cells were plated 24 hr before transfection using the ProFection Mammalian Transfection System Calcium Phosphate (Promega). NIH3T3 cells were infected with retroviruses produced in the Phoenix packaging cells (24 and 48 hr after transfection) in the presence of 8 µg/ml polybrene (Sigma) and were selected with puromycin or hygromycin. Experiments were conducted within 2 weeks after infections and were performed using 2 independent NIH3T3 infected pools.

For the proliferation assay NIH3T3 cells (2.5×10^4 cells) were plated in triplicate in 12-well plates and counted every day for 4 days using a Z2 Coulter (Beckman). For the focus formation assay, 10^5 cells were plated in duplicate in 6-well plates and cultured until confluence. A week later, cells were washed twice with PBS and stained with GIEMSA (Sigma). Representative pictures of the foci were taken after GIEMSA staining at low and high magnification.

MCF-7 adherent cell culture

The ER-positive breast cancer cell line MCF-7 (Soule et al, 1973)from ATCC) was cultured in DMEM supplemented with 10% v/v Fetal Bovine Serum and penicillin/streptomycin (all from Gibco, Invitrogen) according to manufacturer's recommendations.

ZNF703 knock down

For knock-down experiments, sub-confluent cultures were transfected with either a si-ZNF703 pool tareting the 3'UTR of endogenous *ZNF703* (100 nM each siRNA1: GCUUGACCCUGCCGGGAUU & siRNA2: CUUACAAGCUGGGAAUAUA; Thermo Scientific Dharmacon RNAi Technologies) or a negative control non-targeting siRNA (100 nM, #1027281, Qiagen) in OptiMEM (Gibco, Invitrogen) using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. The transfection medium was replaced by culture medium after 6 hours, and after 48 hours protein, total RNA and chromatin were extracted from replicate samples.

Construction of lentiviral vector for ZNF703 overexpression

An 1804 bp fragment of the ZNF703 mRNA spanning the entire open reading frame and native stop codon was amplified from HCC-1500 cDNA using ZNF-for-Not1 and ZNF-rev-Not1 primers (see table below). The NotI-digested fragment was cloned into the NotI site of HIV-*ZsGreen* (Welm et al, 2008) and the construct verified by sequencing. The parental vector was used as a negative control. Cleared virus-containing supernatants from transfected HEK-293T cells were obtained as described in (Welm et al, 2008) and used directly without further concentration or titration. Sub-confluent MCF-7 cultures were infected twice on consecutive days in

the presence of 1 μg/ml polybrene (Sigma). Infected cells were sorted by GFP expression using FACS Aria SORP (BD Biosciences) to eliminate non-infected cells.

Protein extraction and Western blot analysis

For protein analysis native, transfected or infected MCF-7 cultures were lysed directly in standard Laemmli Protein Sample Buffer. Samples representing equivalent cell numbers were subjected to Western Blot analysis following standard protocols using polyclonal rabbit anti-ZNF703 (Atlas Antibodies; 1:500) or AC15 monoclonal mouse anti-beta-actin (Abcam; 1:10000) antibodies and HRP-conjugated rabbit-anti-goat or goat-anti-mouse (Dako) secondary antibodies, respectively, visualized by enhanced chemiluminescence using Amersham ECL Advanced Western Blotting Detection Kit (GE Healthcare).

Total RNA extraction, RT-QPCR and microarray analysis of MCF-7 cells

Total RNA was extracted using miRNeasy (Qiagen) following manufacturer's recommendations, its yield and purity monitored by spectroscopy at 260, 280 and 230 nm using the NanoDrop (Thermo Scientific) and its integrity prior to microarrays analysis verified by a Bioanalyzer trace (Agilent).

Random-primed cDNA was generated from 1 μg total RNA using either MultiScribe (Applied Biosystems) or SuperScript III (Invitrogen) according to manufacturers' recommendations. 0.5, 1 and 2 μg of a pooled sample were reverse-transcribed alongside the individual experimental samples to control for saturation of this step. A sample containing all reagents except the reverse transcriptase was included to account for genomic amplification or reagent contamination. Triplicate cDNA aliquots were subjected to real time PCR analysis (QPCR) on the ABI Prism 7900HT system (Applied Biosystems) strictly following the MIQE guidelines (Bustin et al, 2009). Each 10 ml reaction contained 1 ml cDNA, 1x SYBR Green Mastermix (Applied Biosystems) and 300 nM each forward and reverse gene-specific primers (see table below) designed to bind to different exons or over exon-exon boundaries. After an initial dissociation step of 95°C for 10 min, 40 cycles of amplification were carried out at 95°C for 15 sec and 60°C for 60 sec. The end product was assessed by a dissociation step using a ramp from 60-95°C, and all PCR products exhibited a single peak. Background readings from non reverse-transcribed RNAs and non-

template controls were at least 20-fold lower than experimental samples. Absolute expression values accounting for differences in amplification efficiency were calculated by automated software (SDS 2.3, Applied Biosystems) using linear regression of a standard titration curve included for each plate. Expression was normalized for each sample by dividing the relative expression of each gene by the geometric mean of the relative expression values of multiple internal reference genes.

Analysis of expression microarrays from lentivirus-infected and/or siRNA-transfected MCF-7 were performed in a manner similar to that described above, with the exception that data were quantile normalized. Genes with and FDR adjusted p-value < 0.05 were included in downstream geneset enrichment.

Chromatin Immunoprecipitation (ChIP)

ChIP experiments were performed as described previously (Schmidt, et al Methods, 2009) using anti-ZNF703 (sc-82451x), anti-p300 (sc-585) and anti-HDAC1 (sc-6299 sc-6298) from Santa Cruz Biotechnologies (CA, USA). Primers used for real-time PCR are listed in the table below. Statistical analysis was performed using two tailed paired T-tests. P-value cut-offs are indicated in the relevant figures.

ZNF703 expression in human mammary epithelium cells (HMECs)

The 184-hTERT-L9 immortalized normal mammary epithelial cell lines were clonally derived from the parental 184-hTERT cell line constructed in the Laboratory of Molecular Carcinogenesis at the NIEHS (Horikawa et al., 2002). These cells were cultured in a humidified incubator at 37 ºC with 2.5% CO2 in MEBM medium (Lonza) supplemented with MEGM Single Quots (Lonza), 5ug/ml transferrin (Sigma) and 10-5M isoproteranol (Sigma).

184-hTERT-L9 cells were plated at a density of 1.8x10^4 cells/cm2 in 6cm cell culture dishes 24 hours prior to transduction. Cells were infected with freshly thawed lentiviral particles at an estimated multiplicity of infection (MOI) of 5 with 8ug ml–1 polybrene in the standard 184-hTERT-L9 culture media. After 18 hours at 37 ºC, cells were washed three times in HBSS with 2% FBS and returned to standard media. Cells expressing ZsGreen were collected on a FACSDiva (Becton Dickinson) using gates that excluded 99.9% of events present in uninfected control cells, events

with very high forward and side light scatter profile, and dead cells identified with Dapi at 0.1 ug ml–1(Sigma). Zsgreen positive cells were expanded and subsequently subjected to cell cycle and ZNF703 expression analysis.

Cell cycle analysis was performed using the Click-iT EdU Alexa Fluor 647 Flow Cytometry Kit (Invitrogen) following manufacture's suggestions with 45 minutes of 10μm EdU labeling and saponin-based permeabilization. EdU incorporation and DNA content were assessed in triplicate on a FACSCalibur (Becton Dickinson) and data analysis was performed using FlowJo software (Tree Star).

ZNF703 expression in primary human mammary epithelium

Lentiviral production and titering. Lentiviral particles were generated by transient transfection of the HIV-*ZsGreen*-ZNF703 lentiviral vector or the HIV-*ZsGreen* empty vector control along with the pCMV R8.91 packaging plasmid and pMD2 VSV-G envelope virus obtained from the RNAi Consortium (Broad Institute, MA). Briefly, HEK 293T cells (ATCC) were plated at a density of 5x10^5 cells per cm^2 in DMEM (Stem Cell Technologies) supplemented with 10% FBS (Gibco) 24 hours prior to transfection. A ratio of 10:9:1 of vector to pCMV R8.91 and pMD2 VSV-G was used for co-transfection using TransIT-LTI transfection reagent (Mirus Bio) as per the manufacturer's recommendations. The culture medium was replaced 18 h post-transfection and viral supernatant was collected 24 h and 48 h later and passed through a 0.45 um pore-size cellulose-acetate filter. Pooled supernatants were concentrated 100-fold by ultracentrifugation at 100,000 xg for 105 min. Pellets were re-suspended by shaking in DMEM for at least 30 min and then stored in aliquots at -80°C until use. Viral titers were determined by flow cytometry in HEK 293T cells. For this, five-fold serial dilutions of viral stocks were prepared in HEK 293T growth media and added to the cells with 8 ug ml–1 polybrene (Santa Cruz Biotechnology). The transduction mix was removed from the cells 6 hours post-infection. The percentage of ZsGreen positive cells was determined 72 hours post-transduction on a FACSCalibur (Becton Dickinson) and used to calculate viral titers.

Human mammary epithelial cell preparation and separation. Discarded tissue was collected with informed consent from premenopausal women (ages 19-40) undergoing reduction mammoplasty surgery as approved by the University of British Columbia Research Ethics Board. Post-collection processing was performed as

previously described (Stingl et al, 2005). Briefly, tissue was transported from the operating room on ice and was minced with scalpels prior to dissociation in Ham's F12 and DMEM (1:1 vol/vol, F12 to DMEM, StemCell Technologies) supplemented with 2% wt/vol BSA (Fraction V; Gibco Laboratories), 300 U ml–1 collagenase (Sigma) and 100 U ml–1 hyaluronidase (Sigma) for 18 hours. Differential centrifugation at 80g for 4 min was used to collect an epithelial-cell rich pellet. This was cryopreserved in media containing 6% dimethylsolfoxide at –135°C until use. Single cell suspensions were subsequently prepared from freshly thawed pellets by sequentially treating the cells with 2.5mg ml–1 trypsin with 1mM EDTA (StemCell Technologies) and 5 mg ml–1 dispase (StemCell Technologies) with 100 mg ml–1 DNase1 (Sigma). Between treatments, cell were washed with cold HBSS (StemCell Technologies) supplemented with 2% FBS (Gibco). Cell suspensions were passed through a 40-um filter (BDBiosciences) to remove remaining cell aggregates. For the separation, cells were pre-blocked in HBSS supplemented with 2% FBS and 10% human serum (Jackson ImmunoResearch). An allophycocyanin-conjugated rat antibody to human CD49f (clone GOH3, R&D Systems) and a phycoerythrin-conjugated mouse antibody to human EpCAM (clone 9C4, Biolegend) were used for Basal and Luminal population discrimination as previously described (Eirew et al, 2008). Hematopoietic and endothelial cells were labeled with biotin-conjugated mouse antibodies to human CD45 (clone HI30, Biolegend) and human CD31 (cloneWM59, eBiosciences), respectively, followed by FITC–conjugated streptavidin (BD Biosciences). Either propidium iodide at 10 ug ml–1 or Dapi at 0.1 ug ml–1 (Sigma) was used for live/dead cell discrimination. Sorting was performed on a FACSDiva (Becton Dickinson) using gates that excluded 99.9% of events present in negatively stained control samples and events with very high forward and side light scatter profile.

Transduction of mammary epithelial cells and *in vitro* mammary colony-forming cell assay. Sorted populations of primary human mammary epithelial cells were immediately infected with freshly thawed lentiviral particles at an estimated multiplicity of infection (MOI) of 3. Transduction was conducted in suspension with 8 ug ml–1 polybrene at a density of 1x10^5 cells in 100ul of Serum-Free 7 (Ham's F12 and DMEM supplemented with 0.1% BSA,10 ng ml–1 EGF (Sigma), 10 ng ml–1 cholera toxin (Sigma), 1 mg ml–1 insulin (Sigma), and 50 ug ml–1 GA-1000 (Lonza)).

After 18 hours at 37°C, cells were washed three times in HBSS with 2% FBS and re-suspended in 100ul of fresh Serum-Free 7. Cells were counted using a hemocytometer and equal numbers were plated in CFC assays performed as previously described (Raouf et al, 2008). For the assay, gridded 60 mm tissue culture dishes (Sarstead) were pre-coated with a 1:43 dilution of collagen solution in PBS (Stem Cell Technologies) for 1 h at 37°C and rinsed with additional PBS. Each dish was then seeded with 5000 mammary epithelial cells in 4 ml of Serum-Free 7 with 5% FBS and 1.6x10^5 freshly thawed NIH 3T3 cells, previously irradiated at 40Gy. After 8-10 days, dishes were fixed and stained with 0.8% w/vol methylene blue in methanol (Sigma). Colonies with 50 cells or more were visually scored under a dissecting microscope.

Comparison of colony formation counts of ZNF703 transfected Luminal and Basal fractions, adjusting for empty vector control colony formation counts, was performed using ANCOVA. The data for this experiment are shown in Table S4, and ANCOVA linear model fits plotted in Figure 6B. Fold change estimates and 95% confidence intervals within each fraction, comparing ZNF703 transfected counts to control counts, were estimated using linear model fits with intercept constrained to 0. Assumption of zero intercept was verified with ANCOVA model fit and by generating diagnostic plots. Adjustment for multiple comparisons for the fold change estimates was performed using the Bonferroni method.

Primers used throughout the study

Gene/Region	Usage	Sequence
TGFBRII upstream region FWD	ChIP-QPCR	CCTCCCTCTTTGTGGTTTGA
TGFBRII upstream region REV	ChIP-QPCR	CACGAACACCATCACTGTCC
TGFBRII promoter FWD	ChIP-QPCR	CATGATTGGCAGCTACGAGA
TGFBRII promoter REV	ChIP-QPCR	GGGGAAACAGGAAACTCCTC
PRKCE intron FWD	ChIP-QPCR	CAAGCCTATGGCACCATTTT
PRKCE intron REV	ChIP-QPCR	AATTCACGCCACTCCAAAAC
ZNF703 mRNA fwd	mRNA (RT-QPCR)	GATCAGGGTCCTGAAGATGC
ZNF703 mRNA rev	mRNA (RT-QPCR)	CCGAGTTGAGTTTGGAGGAG
ERLIN2 mRNA fwd	mRNA (RT-QPCR)	GCTCAGTTGGGAGCAGTTGT
ERLIN2 mRNA rev	mRNA (RT-QPCR)	GAAAGGGAGCATGAGATGGA
PROSC mRNA fwd	mRNA (RT-QPCR)	TTCATGCTGGAAACAGTGGA
PROSC mRNA rev	mRNA (RT-QPCR)	CCCAAAGCTTCCTATGGTCA
GPR124 mRNA fwd	mRNA (RT-QPCR)	CATCTCAGTGAATGCGAGGA
GPR124 mRNA rev	mRNA (RT-QPCR)	TGATGTGGAAGGACGACAGA

BRF2 mRNA fwd	mRNA (RT-QPCR)	ACATATTCCCGAAGCACAGG
BRF2 mRNA rev	mRNA (RT-QPCR)	CAGGTGATTAAGACGCAGCA
RAB11FIP1 mRNA fwd	mRNA (RT-QPCR)	GCAGGAAGACGCAGTGGTAT
RAB11FIP1 mRNA rev	mRNA (RT-QPCR)	GGTGTCTGACCCACTGTCCT
TGFBR2 mRNA fwd	mRNA (RT-QPCR)	TGTGTCGAAAGCATGAAGGA
TGFBR2 mRNA rev	mRNA (RT-QPCR)	GGTCCCAGCACTCAGTCAAC
PAI-1 mRNA fwd	mRNA (RT-QPCR)	TGATGGCTCAGACCAACAAG
PAI-1 mRNA rev	mRNA (RT-QPCR)	GGTCATGTTGCCTTTCCAGT
TGFBI mRNA fwd	mRNA (RT-QPCR)	AGCAGCCCTACCACTCTCAA
TGFBI mRNA rev	mRNA (RT-QPCR)	GACATTGCTGACCAGGGAGT
ZFP36L2 mRNA fwd	mRNA (RT-QPCR)	CCAGCATGTTGTTCAGGTTG
ZFP36L2 mRNA rev	mRNA (RT-QPCR)	TTCTGTCCGCCTTCTACGAT
DHRS2 mRNA fwd	mRNA (RT-QPCR)	CCTGCTGCTGAGCCAGTT
DHRS2 mRNA rev	mRNA (RT-QPCR)	AAGCTGCAATGGAAGAGACC
RALB mRNA fwd	mRNA (RT-QPCR)	GTGGAGACGTCAGCGAAGA
RALB mRNA rev	mRNA (RT-QPCR)	GCTGCTTTTCTTGCCATTCT
UGT2B11 mRNA fwd	mRNA (RT-QPCR)	GACCTGCTGAATGCACTGAA
UGT2B11 mRNA rev	mRNA (RT-QPCR)	ATCCAGGGGCTTTACTGGTT
GPER mRNA fwd	mRNA (RT-QPCR)	CTCCCTGCAAGCAGTCTTTC
GPER mRNA rev	mRNA (RT-QPCR)	TCAGCTTGTCCCTGAAGGT
CDKN2B mRNA fwd	mRNA (RT-QPCR)	GTGGACTTGGCCGAGGAG
CDKN2B mRNA rev	mRNA (RT-QPCR)	GGGTGGGAAATTGGGTAAGA
GAPDH mRNA fwd	mRNA (RT-QPCR)	GAGTCAACGGATTTGGTCGT
GAPDH mRNA rev	mRNA (RT-QPCR)	TTGATTTTGGAGGGATCTCG
SDHA mRNA fwd	mRNA (RT-QPCR)	CGGTCCATGACTCTGGAGAT
SDHA mRNA rev	mRNA (RT-QPCR)	AGGACCTGCCCCTTGTAGTT
UBC mRNA fwd	mRNA (RT-QPCR)	ATTTGGGTCGCGGTTCTTG
UBC mRNA rev	mRNA (RT-QPCR)	TGCCTTGACATTCTCGATGGT

SUPPLEMENTAL REFERENCES

Bengtsson H, Wirapati P, Speed TP (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics 25**:** 2149-2156

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL et al (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55**:** 611-622

Cairns JM, Dunning MJ, Ritchie ME, Russell R, Lynch AG (2008) BASH: a tool for managing BeadArray spatial artefacts. Bioinformatics 24**:** 2921-2922

Dunning MJ, Smith ML, Ritchie ME, Tavare S (2007) beadarray: R classes and methods for Illumina bead-based data. Bioinformatics 23**:** 2183-2184

Eirew P, Stingl J, Raouf A, Turashvili G, Aparicio S, Emerman JT, Eaves CJ (2008) A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. Nat Med 14**:** 1384-1389

Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5**:** 557-572

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27**:** 1160-1167

Raouf A, Zhao Y, To K, Stingl J, Delaney A, Barbara M, Iscove N, Jones S, McKinney S, Emerman J et al (2008) Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell Stem Cell 3**:** 109-118

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3**:** Article3

Soule HD, Vazguez J, Long A, Albert S, Brennan M (1973) A human cell line from a pleural effusion derived from a breast carcinoma. J Natl Cancer Inst 51**:** 1409-1416

Stingl J, Emerman JT, Eaves CJ (2005) Enzymatic dissociation and culture of normal human mammary tissue to detect progenitor activity. Methods Mol Biol 290**:** 249-263

Welm BE, Dijkgraaf GJ, Bledau AS, Welm AL, Werb Z (2008) Lentiviral transduction of mammary stem cells for analysis of gene function during development and cancer. Cell Stem Cell 2**:** 90-102

Supplemental Figures and Tables

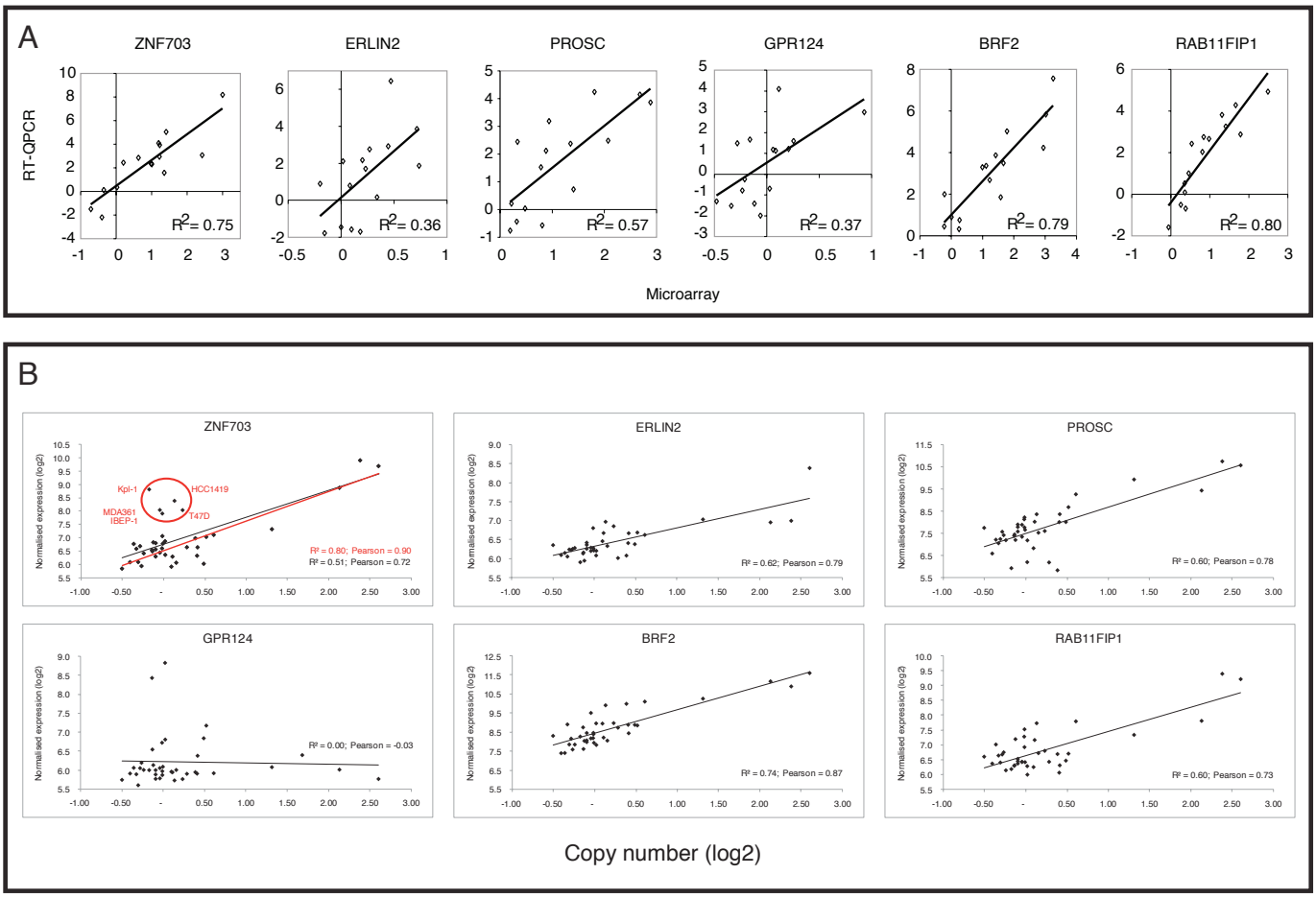

Figure S1: Validation of expression data by RT-QPCR and correlation between copy number and gene expression in breast cancer cell lines.

A: Correlation of RT-QPCR validation of microarray expression data for the six genes within the telomeric 8p12 amplicon.

B: Correlation of expression and gene copy number of the telomeric 8p12 amplicon in a panel of 39 breast cancer cell lines. Normalized values of gene expression (measured by Illumina expression arrays) were plotted against corresponding normalized gene copy number (measured by 28k-array) for a cohort of 39 cell lines. Linear regression R2 values and Pearson correlation coefficients for the entire dataset are indicated in black for each gene. A group of 5 cell lines with copy-number-independent upregulation of ZNF703 is circled and named (MDA361; MDA-MB-361), and the coefficients calculated omitting these outliers are indicated (all in red).

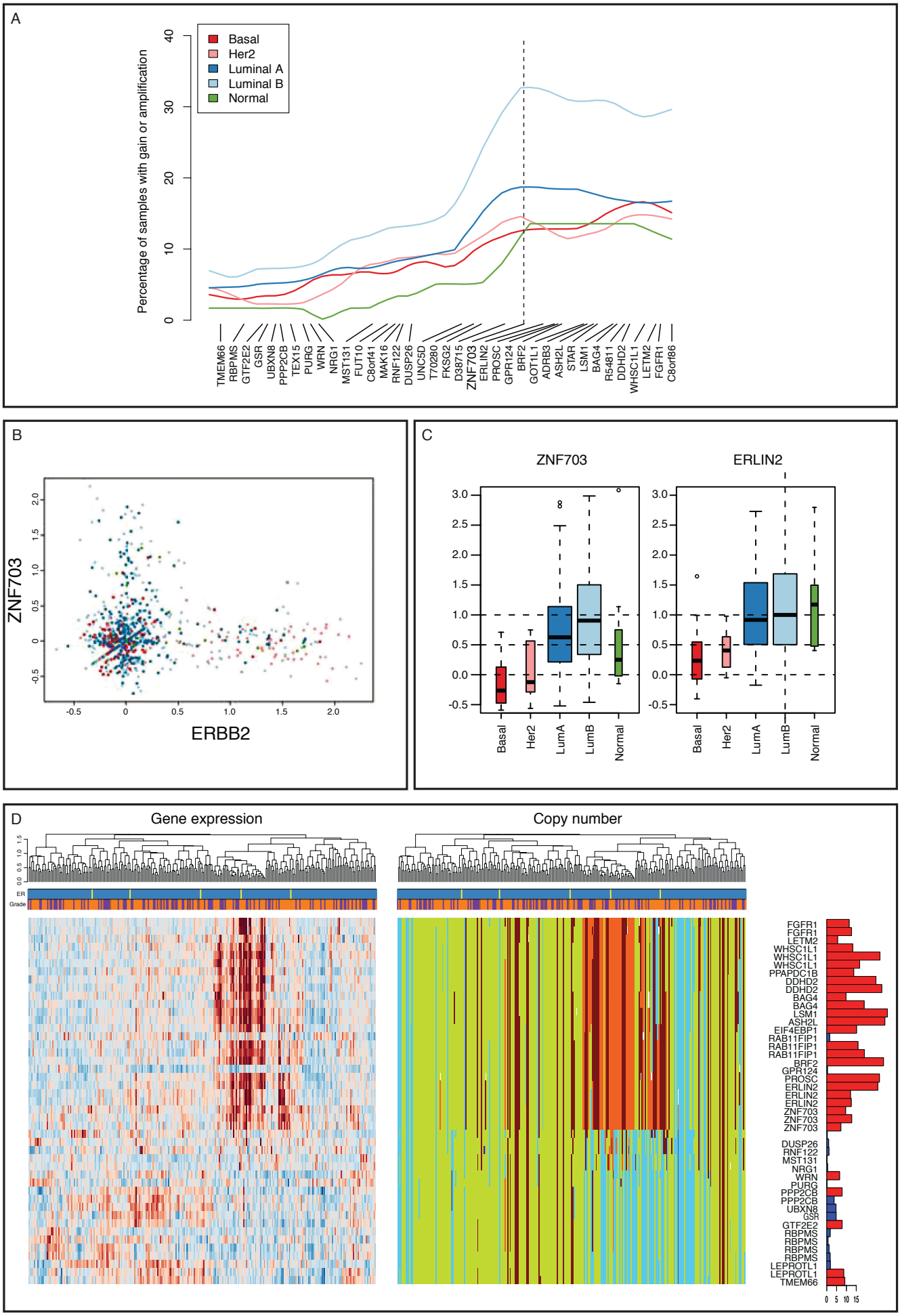
A
Basal
Her2
Luminal A
Luminal B
Normal
Percentage of samples with gain or amplification
0
10
20
30
40
TMEM66
RBPMS
GTF2E2
GSR
UBXN8
PPP2CB
TEX15
PURG
WRN
NRG1
MST131
FUT10
C8orf41
MAK16
RNF122
DUSP26
UNC5D
T70280
FKSG2
D38715
ZNF703
ERLIN2
PROSC
GPR124
BRF2
GOT1L1
ADRB3
ASH2L
STAR
LSM1
BAG4
R54811
DDHD2
WHSC1L1
LETM2
FGFR1
C8orf86
B
ZNF703
ERBB2
C
ZNF703
ERLIN2
Basal
Her2
LumA
LumB
Normal
D
Gene expression
Copy number
ER
Grade
FGFR1
FGFR1
LETM2
WHSC1L1
WHSC1L1
WHSC1L1
PPAPDC1B
DDHD2
DDHD2
BAG4
BAG4
LSM1
ASH2L
EIF4EBP1
RAB11FIP1
RAB11FIP1
RAB11FIP1
BRF2
GPR124
PROSC
ERLIN2
ERLIN2
ERLIN2
ZNF703
ZNF703
ZNF703
DUSP26
RNF122
MST131
NRG1
WRN
PURG
PPP2CB
PPP2CB
UBXN8
GSR
GTF2E2
RBPMS
RBPMS
RBPMS
RBPMS
RBPMS
LEPROTL1
LEPROTL1
TMEM66
0 5 10 15

Figure S2:

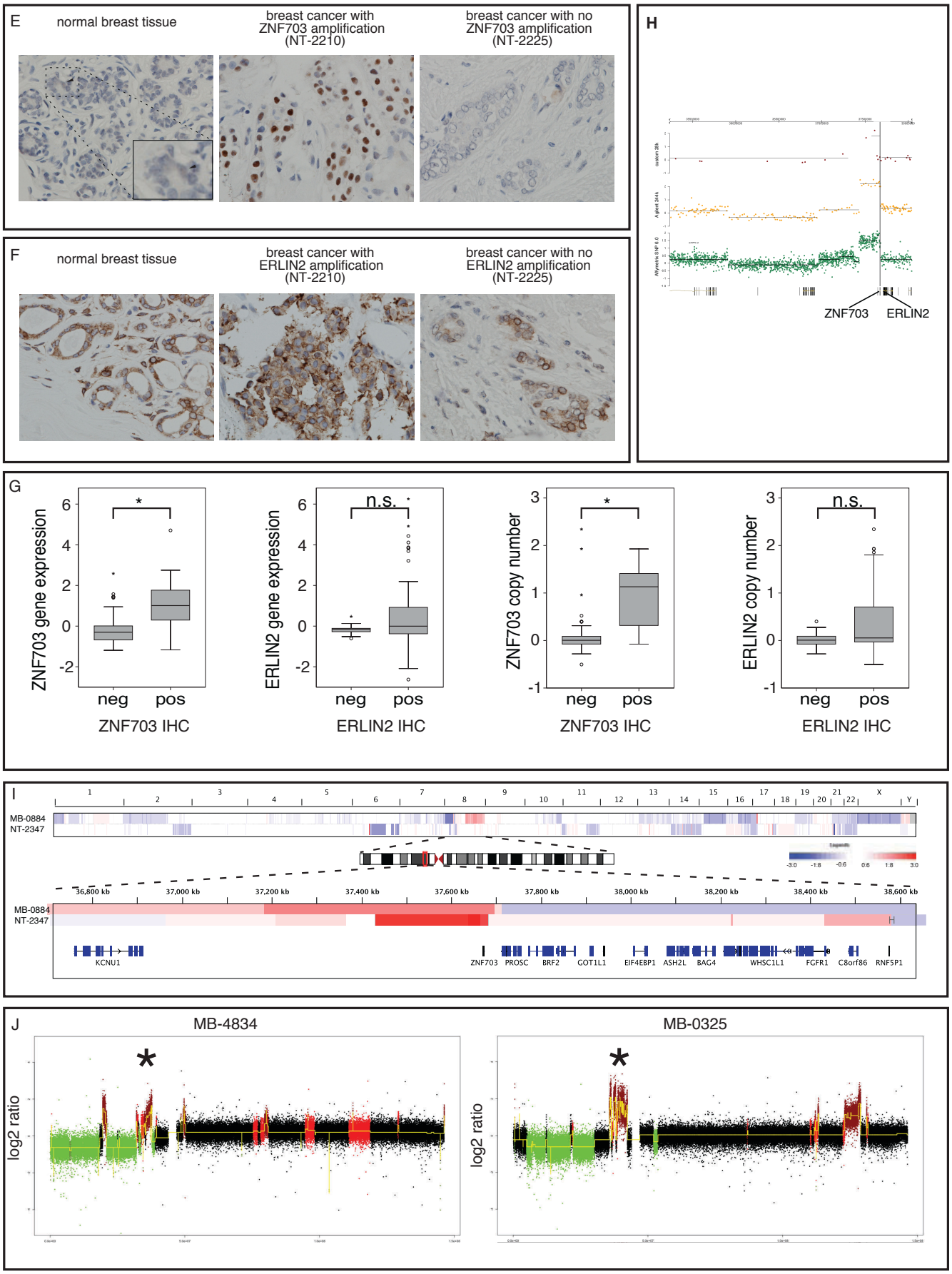
E
normal breast tissue
breast cancer with ZNF703 amplification (NT-2210)
breast cancer with no ZNF703 amplification (NT-2225)
F
normal breast tissue
breast cancer with ERLIN2 amplification (NT-2210)
breast cancer with no ERLIN2 amplification (NT-2225)
H
ZNF703
ERLIN2
G
ZNF703 gene expression
ERLIN2 gene expression
ZNF703 copy number
ERLIN2 copy number
*
n.s.
neg
pos
ZNF703 IHC
ERLIN2 IHC
I
MB-0884
NT-2347
KCNU1
ZNF703
PROSC
BRF2
GOT1L1
EIF4EBP1
ASH2L
BAG4
WHSC1L1
FGFR1
C8orf86
RNF5P1
J
MB-4834
MB-0325
log2 ratio

Figure S2:

Figure S2: Characterization of the 8p12 telomeric amplification core

A: Percentage of samples with gain or amplification of ZNF703 (MB-series). For each Illumina probe mapping to the 8p12 cytoband, the percentage of samples belonging to each intrinsic subtype that had copy-number gain/amplification is indicated. Probes are ordered according to genomic location (not to scale) and a smoothed-curve produced for each subtype. Selected gene symbols are indicated.

B: Scatterplot comparing ZNF703/ERLIN2 segmented means versus ERBB2 segmented means. Each dot represents one tumor sample color coded according to PAM50-subtype (see A for legend).

C: Box plots depicting association between copy-number states and expression levels for ZNF703 and ERLIN2. For each subtype the box represents the distribution of expression in samples that have copy-number amplification of the gene relative to the median expression of copy-number neutral samples.

D: Paired heatmaps of the 8p12 expression data (left) and copy-number states (right) for samples in the MB-series of 1001 primary tumors that were classified as Luminal B according to PAM50. The ordering of rows and columns is identical in both heatmaps. Columns were ordered by clustering the samples based on the expression profiles of 8p12 genes. Rows are ordered according to genomic coordinates of the microarray probe. The barplot indicates the –log10 p-value from the Wilcoxon test for association between copy-number and expression of each probe (red- significant; blue- not significant).

E: Examples of ZNF703 immunohistochemistry (IHC) staining in normal breast tissue (left) and ZNF703-amplified and copy-neutral cases from the NT series (middle and right, respectively). Arrowhead: exemplar ZNF703-weakly positive nucleus in normal breast epithelium, also shown in a digitally magnified insert.

F: Examples of ERLIN2 immunohistochemistry (IHC) staining in normal breast tissue (left) and ERLIN2-amplified and copy-neutral cases from the NT series (middle and right, respectively).

G: Box plots of gene expression and copy number vs ZNF703 and ERLIN2 IHC status (negative or positive). IHC staining was conducted on tissue-microarrays representing 125 primary breast carcinomas from both NT- and MB-series for which copy-number and gene-expression data were available. IHC staining was scored according to an Allred semi-quantitative system (Allred score = intensity score + proportion score. Intensity score: 0 = none, 1 = weak, 2 = moderate, 3 = strong; Proportion score: 0 = none, 1 = <1%, 2 = 1-10%, 3 = 11-33%, 4 = 34-66%, 5 = >66%) and cases were divided into 'negative' (neg; Allred <5) and 'positive' (pos; Allred score ≥5). Significance was assessed by a Mann-Whitney U test and a p-value <0.01 was considered significant (*); n.s. non significant.

H: Probe-level data in sample NT-2347, which exhibits the narrowest ZNF703 amplicon, profiled on several aCGH platforms. Segmented means are overlaid for the 28k-array, Agilent 244k array, and Affymetrix SNP 6.0 platform.

I: Whole-genome heatmap overview (top) and 8p12 heatmap (bottom) of the two tumors with a ZNF703 single-gene amplicon (using IGV viewer).

J: Probe-level data (log2) for chromosome 8 in two tumors with a ZNF703 8p12 amplicon. Red: gain or amplification; green: loss or deletion. Asterisks were placed over examples of areas with high complex arm aberration indices.

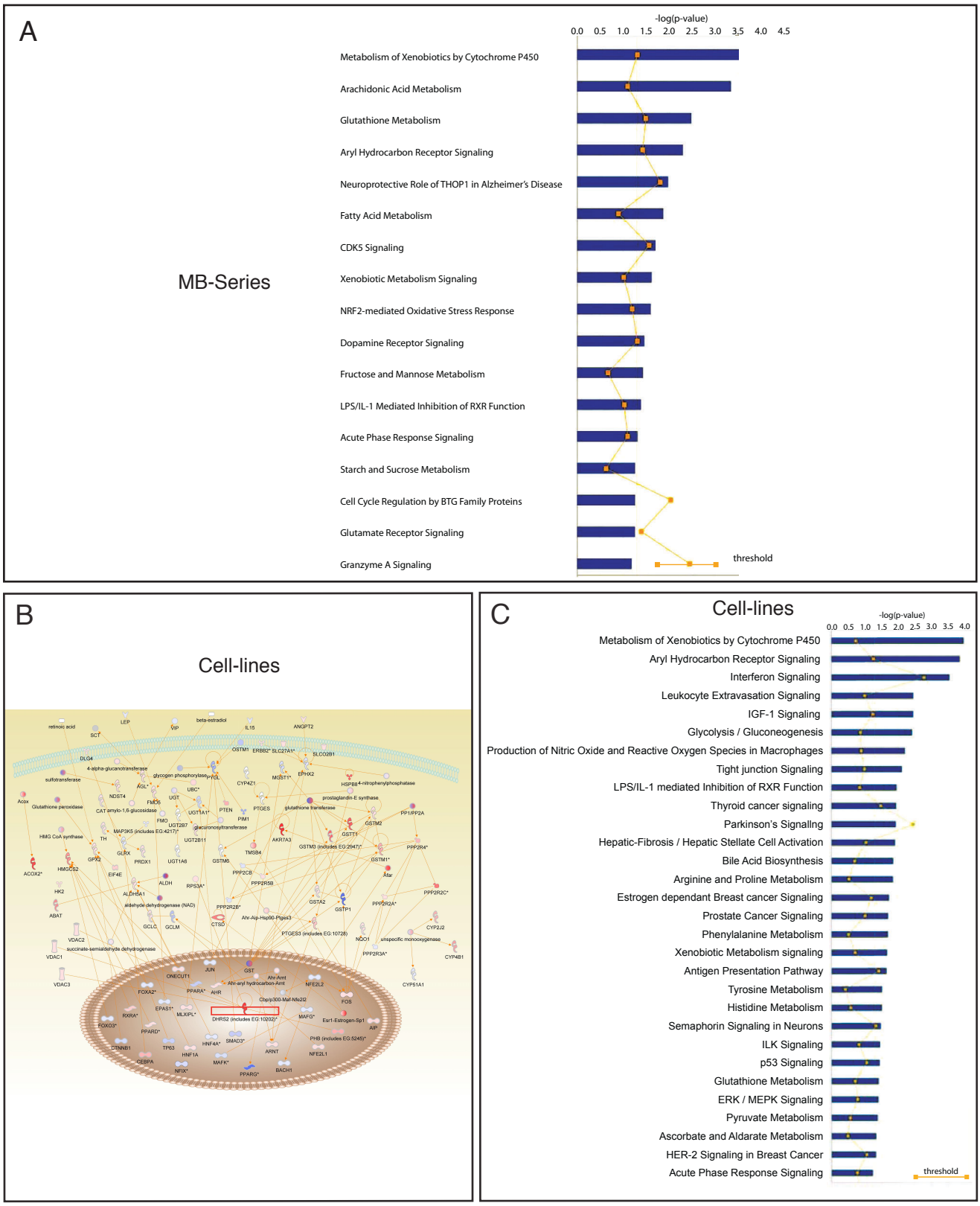

Figure S3: Differentially expressed genes in ZNF703-upregulated vs neutral tumors and cell lines are enriched for similar pathways.

Pathway-enrichment analysis was carried out as in main Fig.3A-B.

A: Barplot with enrichment for each of the network components in primary tumors, where the strength of the association is represented by the –log10(p-value).

B: Graphical representation of the key enzymes and molecules involved in canonical lipid metabolism and detoxification pathways enriched for amongst ZNF703-upregulated cell lines. Dark red represents molecules whose expression levels are up-regulated greater than 1.5 fold, whereas dark blue represents molecules whose expression levels are down-regulated greater than 1.5 fold. All genes/molecules are significantly differentially expressed with FDR-adjusted p-values < 0.01. Arrows represent biological relationships between molecules and the various shapes represent the functional class of the gene product.

C: Barplot with enrichment for each of the network components in cell lines, where the strength of the association is represented by the –log10(p-value).

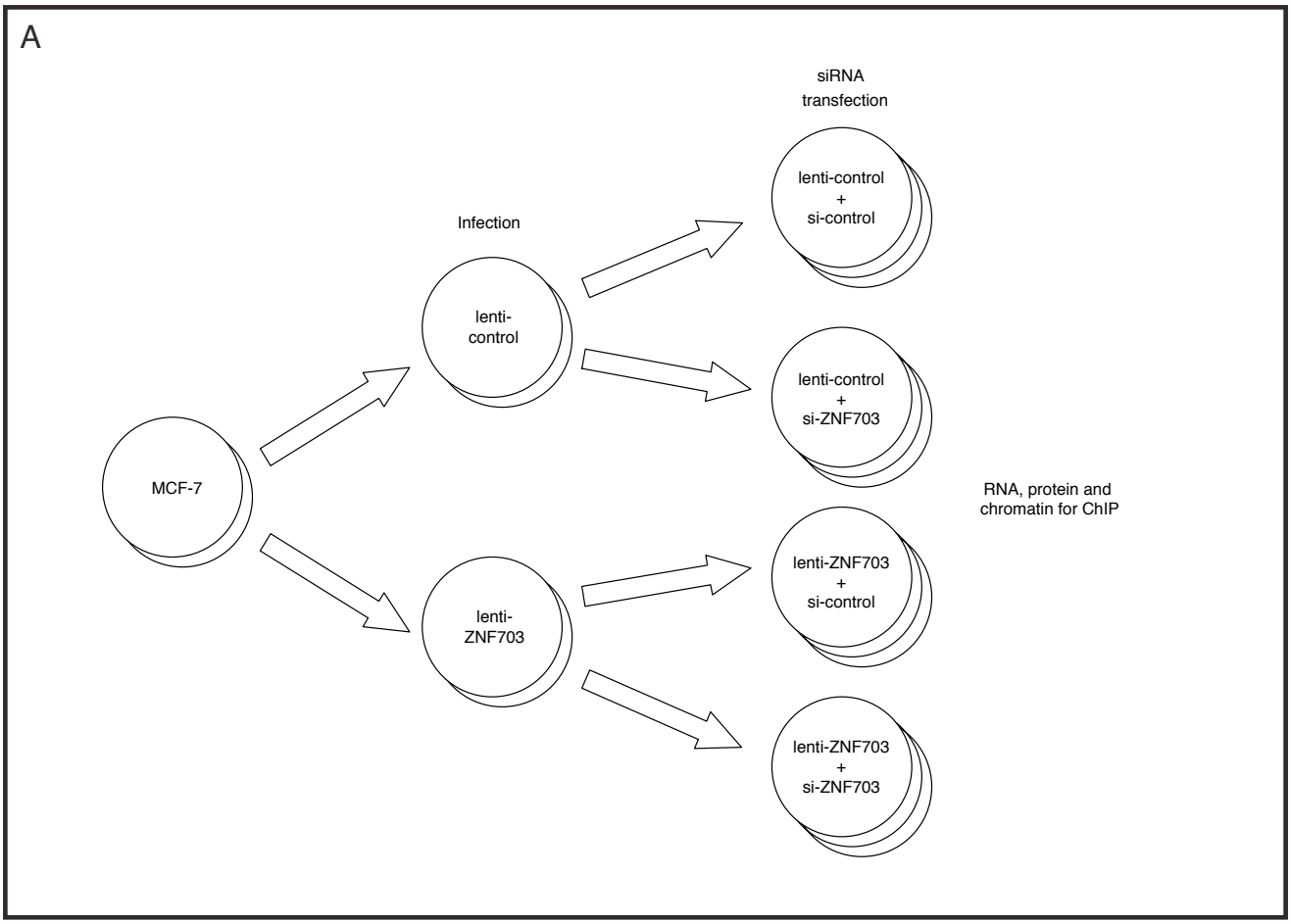

Figure S4: Schematic representation of the experimental design for manipulation of ZNF703 expression in MCF-7 cells

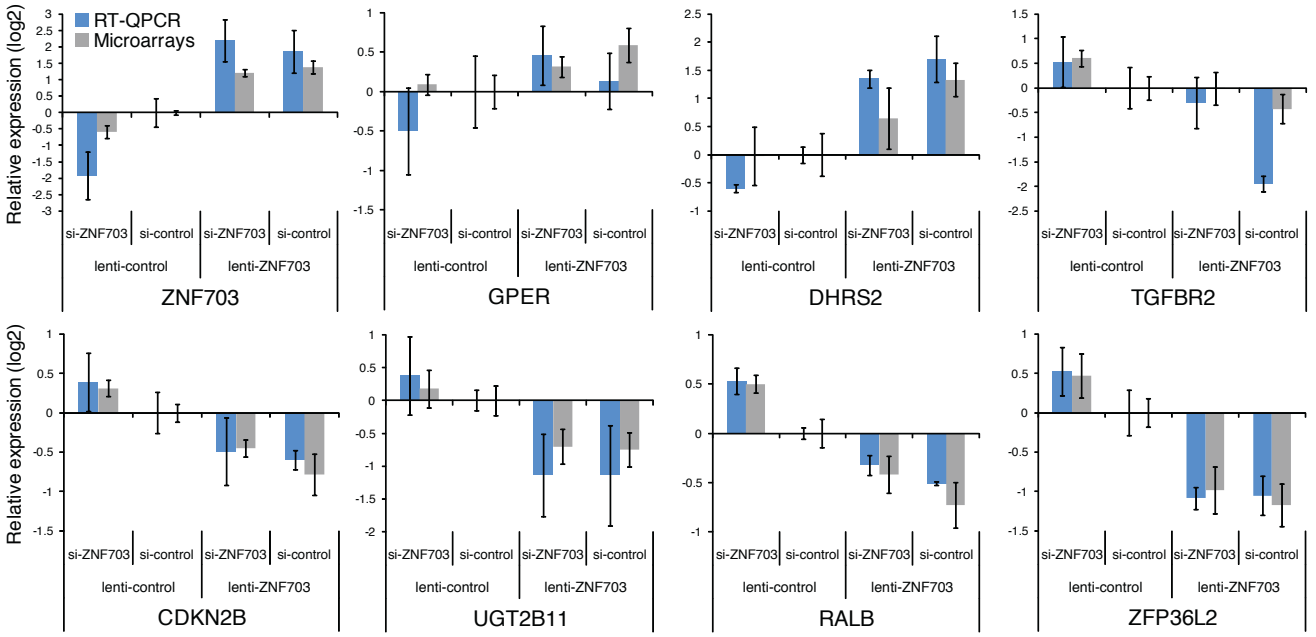

Figure S5: Expression of selected genes as assessed by microarray and validated by RT-QPCR, relative to lenti-control/si-control.
Bar plots represent averages for relative expression (RT-QPCR: log2 transformed normalised values; microarrays: normalised log2 values) with error bars representing standard deviations of biological replicates.

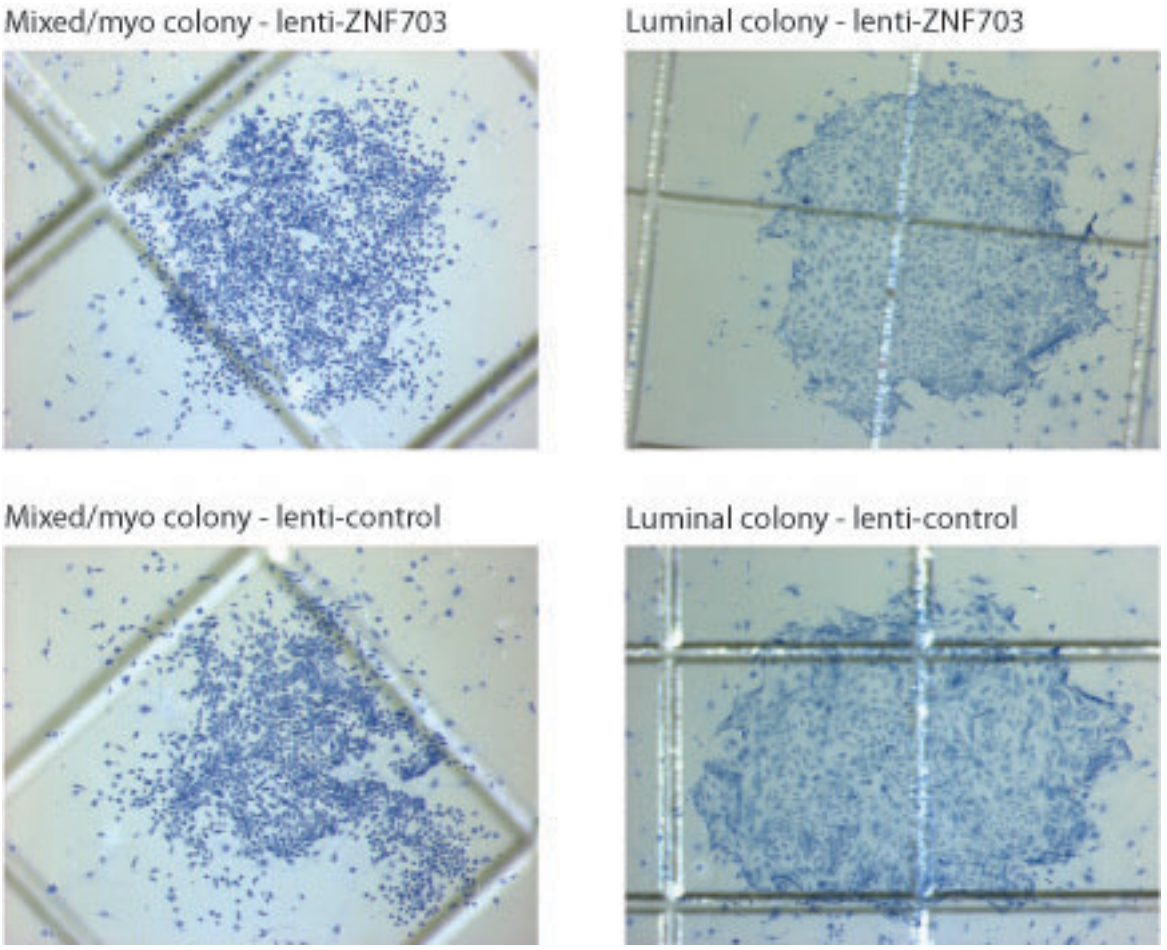

Figure S6: Representative images of mixed/basal and luminal colonies in lenti-ZNF703 and lenti-control infected normal mammary epithelial cells.

Table S1A (TableS1a_MB_ZNFZ03_AMPorOvExp.txt file): Differentially expressed genes for the Luminal B ZNF703 amplified or over-expressed versus neutral contrast in the MB-series.

Table S1B (TableS1b_Celline_ZNF703_AMPorOvExp.txt): Differentially expressed genes for the ZNF703 amplified or over-expressed versus neutral contrast in cell lines.

All Illumina probes are listed, those with an adjusted p value less that 0.05 were deemed differentially expressed.

Table S2 (TableS2.txt): Differentially expressed genes for the MCF7 ZNF703 lentiviral over-expression or siRNA knockdown assays.

All Illumina probes are listed, those with an adjusted pvalue less that 0.05 were deemed differentially expressed.

Gene	MCF7 *ZNF703* siRNA			ERpositive HER2 negative Tumours	
	logFC	p-value	adj. P-value	p-value	adj. P-value
TGFBR2	0.55	<0.001	0.998	<0.001	<0.001
ADM	0.87	<0.001	0.998	<0.001	<0.001
RALB	0.47	<0.001	0.998	<0.001	<0.001
BMP1	0.67	<0.001	0.998	<0.001	<0.001
LOC643446	-0.50	0.001	0.998	<0.001	<0.001
RIN2	0.56	0.001	0.998	<0.001	<0.001
UNKL	-0.47	0.001	0.998	<0.001	0.021
PTHLH	0.48	0.001	0.998	<0.001	0.002
CA12	-0.45	0.001	0.998	<0.001	<0.001
PDPK1	-0.40	0.001	0.998	<0.001	<0.001

Table S3: Genes affected by *ZNF703* knockdown in MCF7 cells are also differentially expressed in tumours with high *vs* normal ZNF703 expression.

10 out of 46 genes identified by the 50 most differentially expressed probes following *ZNF703* knockdown in MCF7 cells were differentially expressed in ER positive / HER2 negative tumors with high vs normal ZNF703 expression. Tumors were divided into ZNF703 high vs normal using a threshold of 0.5 for the z-score transformed *ZNF703* gene expression value (ERposHER2neg: 200 high vs 554 normal). Significance was calculated by performing a one-sided Mann-Whitney test for the expected directionality of differential expression. P-values were adjusted for multiple testing using the Bonferroni method.

		Empty Vector		ZNF703	
Patient Sample	Sorted Fraction	Luminal	Mixed/Myo	Luminal	Mixed/Myo
16-07	Luminal-enriched	33	0	164	2
33-08	Luminal-enriched	2	0	6	0
74-08	Luminal-enriched	6	0	29	1
51-09	Luminal-enriched	61	1	80	3
64-07	Luminal-enriched	85	0	138	11
71-07	Luminal-enriched	89	4	238	9
16-07	Basal-enriched	0	84	0	32
33-08	Basal-enriched	0	115	0	5
74-08	Basal-enriched	2	80	1	39
51-09	Basal-enriched	0	57	0	24
64-07	Basal-enriched	0	203	1	90
71-07	Basal-enriched	2	179	0	76

Table S4: Primary human mammary epithelium colony forming assay data
Distribution of morphologically discriminated colony types formed from fractionated human mammary epithelial cells infected with lenti-ZNF703 or lenti-control viruses.