Breast specific molecular clocks comprised of ELF5 expression and promoter methylation identify individuals susceptible to cancer initiation.

Masaru Miyano1#, Rosalyn W. Sayaman1,2#, Sundus F. Shalabi1,3, Parijat Senapati4, Jennifer C. Lopez1, Brittany Lynn Angarola5, Stefan Hinz1, Arrianna Zirbes1,3, Olga Anczukow5, Lisa D. Yee6, Mina S. Sedrak7,8, Martha R. Stampfer9, Victoria L. Seewaldt1, and Mark A. LaBarge*1,7,9,10

1Department of Population Sciences, Beckman Research Institute at City of Hope, Duarte, CA, USA
2Department of Laboratory Medicine, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA
3Irell and Manella Graduate School of Biological Sciences, City of Hope, Duarte, CA, USA
4Department of Diabetes Complications and Metabolism, Beckman Research Institute at City of Hope, Duarte, CA, USA
5The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
6Department of Surgery, City of Hope National Medical Center, Duarte, CA, USA
7Center for Cancer and Aging, City of Hope, Duarte, CA USA
8Department of Medical Oncology and Therapeutics Research, City of Hope National Medical Center, Duarte, CA USA
9Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA USA
10Center for Cancer Biomarkers, University of Bergen, Bergen, Norway

#equal contributions
* Corresponding author.

The authors declare no potential conflicts of interest.
Abstract
A robust breast cancer prevention strategy requires risk assessment biomarkers for early detection. We show that expression of ELF5, a transcription factor critical for normal mammary development, is downregulated in mammary luminal epithelia with age. DNA-methylation of the ELF5 promoter is negatively correlated with expression in an age-dependent manner. Both ELF5 methylation and gene expression were used to build biological clocks to estimate chronological ages of mammary epithelia. ELF5 clock-based estimates of biological age in luminal epithelia from average-risk women were within 3 years of chronological age. Biological ages of breast epithelia from BRCA1 or BRCA2 mutation carriers, who were high risk for developing breast cancer, suggested they were accelerated by two decades relative to chronological age. The ELF5 DNA methylation clock had better performance at predicting biological age in luminal epithelial cells as compared to two other epigenetic clocks based on whole tissues. We propose that the changes in ELF5 expression or ELF5-proximal DNA methylation in luminal epithelia are emergent properties of at-risk breast tissue and constitute breast-specific biological clocks.
Introduction

Breast cancer (BC) is the most common cancer among women in the U.S. There are many known risk factors for BC including aging, obesity, alcohol consumption, tobacco smoke, and family history(1,2). Women who are germline carriers of pathogenic variants in BRCA1 or BRCA2 have approximately 70% risk of a BC diagnosis by 80 years of age compared to only 10% risk of a BC diagnosis in the general population(3). 90% of women who receive a BC diagnosis have no known inherited mutation(4) or family history(5), demonstrating the prevalence of sporadic BC over inherited BC. The fact that more than 75% of women diagnosed with BC are over age 50 indicates that aging is the greatest risk factor for BC though we do not know the totality of molecular mechanisms underlying this relationship(6). There is considerable interest to identify factors that change with age to be used as biomarkers for estimating the risk of age-related diseases.

Age-specific DNA methylation patterns have been reported in a number of tissues(7-9). Horvath proposed that age-associated DNA methylation changes, so-called epigenetic clocks, can be used for estimation of biological age(10). Normal breast tissue adjacent to tumors showed poor correlation with Horvath’s original pan-tissue epigenetic clock composed of 353 CpG sites(10), suggesting that breast may have a unique aging progression compared to other tissues. Age-dependent DNA methylation changes were reported in whole breast tissue at gene regulatory elements, at sites that show further alteration in cancer tissues(11,12). We previously showed that DNA methylation and gene expression patterns in luminal epithelial cells shifted toward that of the myoepithelial lineage in an age-dependent manner(13,14). The luminal-specific transcription factor ELF5, E74-like factor 5, stood out as a potential breast-specific aging biomarker because it exhibited excellent dynamic range of gene expression between luminal epithelial cells collected from average-risk younger (<30y) and older (>55y) women(13).

We hypothesized that measurements of ELF5 expression or ELF5 promoter methylation could be used to estimate chronological age of normal, non-cancer breast tissue. We examined RNA sequencing and genome-wide DNA methylation in luminal epithelia from reduction mammoplasty tissue from women who are at Average-Risk for BC and showed that the expression of lineage-specific transcription factor ELF5 was downregulated with age and negatively correlated with DNA methylation on its promoter.
region. In contrast, luminal epithelia from prophylactic mastectomy tissue from women with verified germline mutations that classify them as high-risk for breast cancer, the acceleration of *ELF5* downregulation was detected in individuals with *BRCA1*, *BRCA2* and *PALB2* mutations. We propose that the changes in *ELF5* expression or *ELF5*-proximal DNA methylation are emergent properties of at-risk breast tissue and constitute breast-specific biological clocks that could be used to identify women at higher than Average-Risk of breast cancer.

**Materials and Methods**

**Breast tissue collection and HMEC culture**

Prophylactic mastectomy and contralateral to tumor breast tissues were collected at City of Hope (Duarte, CA) under approved Institutional Review Boards protocols which included written informed consent. Breast organoids from reduction mammoplasty and peripheral to tumor breast tissues were collected in Dr. Stampfer’s lab at Lawrence Berkeley National Laboratory (Berkeley, CA) with approved IRB protocols which included written informed consent for tissue collection and sample distribution. 4th passage HMEC were generated and maintained according to previously reported methods using M87A medium containing cholera toxin and oxytocin at 0.5 ng/ml and 0.1nM, respectively(15,16).

**Flow cytometry**

For dissociation of uncultured cells from organoids, organoids were digested with 0.5% trypsin/EDTA for 10min at 37°C with agitation. After trypsin treatment, organoids were disrupted by vigorous shaking for 30sec. Then, cells were passed through 40um cell strainer (BD Falcon). Dissociated uncultured cells from breast organoids and 4th passage HMEC were stained with anti-CD133-PE (Biolegend, clone7) and CD271 (Biolegend, clone ME20.4) by following standard flow cytometry protocol. Cells were sorted by AriaIII (Becton Dickinson).

**Gene expression analysis**

Total RNAs were isolated from FACS-enriched LEp and MEp with Quick-RNA Microprep kit (Zymo Research). For qPCR, cDNAs were synthesized with iScirpt reverse transcriptase (BioRad) according to manufacturer’s manual. Quantitative gene
expression analysis was performed by CFX384 real-time PCR (BioRad) with Universal
SYBRGreen supermix (BioRad). Data was normalized to RPS18 by relative standard
curve method. For RNA-seq, isolated RNAs were submitted to Integrative Genomic Core
at City of Hope (IGC at COH) for library preparation and sequencing. Primers are listed
in Table 1.

RT-PCR analysis for isoform detection
400 ng of RNA was reverse transcribed using Superscript III reverse transcriptase
(Invitrogen). Semi-quantitative PCR was used to amplify 10 ng cDNA with Phusion High-
Fidelity DNA polymerase (NEB) at 56°C for 22 cycles (GAPDH) and 31 cycles (ELF5)
with isoform specific primers listed in Table 1. PCR products were separated in 1.8%
tagorose gel stained with SYBR Safe (Invitrogen) and imaged using ChemiDoc MP
Imaging System (Bio-rad). PCR bands were quantified using ImageLab 6.0 software
(Bio-rad). Bands were authenticated by Sanger sequencing (Eton Bioscience). Primers
are listed in Table 1.

DNA methylation
Genomic DNA purifications from FACS-enriched each lineage were performed with
Quick-gDNA Microprep kit (Zymo Research). Genomic DNA was digested with McrBC
(New England BioLabs) and EcoRI (New England BioLabs), or EcoRI only as a control.
DNA methylation was measured by real-time PCR using CFX384 (BioRad). Amount of
DNA methylation was normalized by internal primer control that targeted the DNA not
containing CG dinucleotide. DNA methylation by McrBC method shows % of cells with
methylated DNA. For high throughput DNA methylation analysis, purified gDNAs were
submitted for sample preparation to UCLA Neuroscience Genomics Core (UNGC) and
Integrative Genomic Core at COH for HumanMethylation450 and Human
MethylationEPIC Beadchip array, respectively. Primers are listed in Table 1.

Luminal and Myoepithelial RNA-Sequencing Data
RNA-sequencing pre-processing is fully described in (14). Briefly, raw counts from
FACS-sorted LEp and MEp were normalized and regularized log (rlog) transformed in
DESeq2 package. Rlog values were batch-adjusted using sva’s ComBat function with
the experimental design group as covariate in the model matrix. The experimental
design group was defined by the combination of the culture condition (organoid, 4th
passage), cell type (LEp, MEp) and age/risk status (Average risk young <30y, Average
risk old >55y, and Prophylactic mastectomy/Contralateral/Peripheral tissue to tumor
without or with germline mutation) of the samples.

Batch-adjusted rlog values were used for visualization of *ELF5* expression
values. The mean rlog gene expression was calculated for individuals with replicate
samples. For each of the 30,079 mapped transcripts, the mean rlog value from RM
samples from each lineage and age group in 4th passage HMECs (<30y LEp/MEp n=11,
>55y LEp/MEp n=8) and organoids (<30y LEp n=4, <30y MEp n=3, >55y LEp n=3, MEp
n=1) was calculated and a linear regression between organoid and HMEC mean
expression was plotted in (Figure 1).

During quality control assessment, genes with low counts were determined using
eedgeR’s filterbyExpr function with experimental design group and batch as covariates in
the design matrix and were removed. Regression of batch-adjusted rlog values between
cells isolated from organoid or 4th passage culture was performed in each of the
reduction mammoplasty LEp <30y, LEp >55y, MEp <30y, MEp >55y subsets, and genes
with absolute regression residuals ≥ 6 in either of the 4 subsets were considered
outliers and flagged for exclusion. Normalization factors were calculated for the filtered
raw count data using edgeR’s calcNormFactors function. Genes with changes in
lineage-specific expression (*adj. p-val* < 0.1) in young <30 LEp and MEp between
organoid and 4th passage culture were identified by differential expression analysis in the
subset of Average risk Reduction mammoplasty samples young <30y LEp and MEp
using limma’s voom function with subject IDs used to calculate duplicate correlation and
blocking, and design group and batch modeled in the design matrix. Genes with
discordant expression between organoid and 4th passage culture were subsequently
excluded from lineage-specific and age-specific differential expression and downstream
analysis of 4th passage data. A final set of 17,328 genes were analyzed for differential
expression.

Lineage-specific (young <30y LEp vs. <30y MEp) and age-dependent (young
<30y LEp vs. >55y LEp) differential expression analysis in the subset of 4th passage
Reduction mammoplasty LEps and MEps was conducted using limma’s voom function
as described(14). Subject IDs were used to calculate duplicate correlation and for
blocking, and design group and batch were modeled in the design matrix. Moderated statistics were computed using eBayes function in limma. P-values were adjusted for multiple-testing using Benjamini-Hochberg method. *ELF5* differential expression in <30y and >55y in 4th passage LEps and MEps were obtained from (14). Statistical comparison of lineage-specific (young <30y LEp vs. <30y MEp) and age-dependent (young <30y LEp vs. >55y LEp) *ELF5* expression in Reduction mammoplasty organoids was performed using a two-sided Welch t-test.

**Normal Primary Breast Tissue Gene Expression Data**

Normalized microarray expression data from normal primary breast tissue from 114 women (GSE102088, ≤30y n=35, >30y <55y n=68, ≥55y n=11) (17) were downloaded from the Gene Expression Omnibus (GEO) database using GEOquery package. *ELF5* expression in bulk tissue was obtained from this set. Statistical comparison of expression of *ELF5* was performed across 3 age groups: young <30y, middle aged >30y <55y, and old >55y using non-parametric Kruskal-Wallis test. Post-hoc analysis was performed using pair-wise Wilcoxon test comparing each age group, and p-values were adjusted for multiple-testing across groups using Benjamini-Hochberg method.

**TCGA RNA-Sequencing Data**

RNA-sequencing FPKM-UQ values from TCGA were downloaded using TCGAbiolinks package. Analysis was restricted to samples from women with annotated PAM50 breast cancer subtype (PAM50 LumA n=566, LumB n=207, Her2 n=82, Basal n=194, Normal n=40). *ELF5* log2(FPKM-UQ+1) expression values in breast cancer tissue across PAM50 subtypes were obtained from this set. Statistical comparison of *ELF5* gene expression were performed across the 5 PAM50 subtypes: LumA, LumB, Her2, Basal and Normal-like using non-parametric Kruskal-Wallis test. Post-hoc analysis was performed using pair-wise Wilcoxon test comparing each PAM50 subtype, and Wilcoxon p-values were adjusted for multiple-testing using Benjamini-Hochberg method.

**DNA Methylation at the ELF5 locus in LEp**
Illumina 450K array IDAT files pre-processing is fully described(14). Briefly, IDAT values were loaded and detection p-value filtering was conducted across the dataset in ChAMP package. BMIQ normalization was performed and DNA methylation m-values were calculated from beta values using lumi package. The experimental design group was defined by the combination of the cell type and age group.

For visualization, DNA methylation m-values were batch-adjusted using sva’s ComBat function with the experimental design group as covariate in the model matrix. Batch-adjusted m-values were converted to batch-adjusted beta-values in lumi, and were used to compare methylation levels of ELF5 CpG sites across age/risk status and cell types.

Differential methylation of CpG sites was conducted in limma using filtered non-batch-adjusted m-values. Array weights were calculated, subject IDs were used to calculate duplicate correlation and for blocking, and design group and batch were modeled in the design matrix. Moderated statistics were computed using eBayes function in limma with trend applied. P-values were adjusted for multiple-testing using Benjamini-Hochberg method.

Normal Primary Breast Tissue DNA Methylation Data

Illumina 450K array IDAT files from primary breast tissue were downloaded from two independent GEO datasets with 121 samples (GSE101961, ≤30y n=38, >30y <55y n=71, ≥55y n=12) (17) and 100 samples (GSE88883, ≤30y n=36, >30y <55y n=53, ≥55y n=11) (12), respectively. As with HMEC data, detection p-value filtering and BMIQ normalization were performed in each dataset. The experimental design group was defined by age groups (young <30y, middle aged >30y <55y, and old >55y). For visualization, DNA methylation m-values were calculated and batch-adjusted within each dataset using ComBat with the experimental design group and BMI group as covariates in the model matrix. Batch-adjusted m-values were converted to batch-adjusted beta-values, and were used to compare methylation levels of ELF5 CpG sites in normal breast tissue across age groups.

Statistical comparison of DNA methylation of ELF5 CpG sites was performed across 3 age groups: young <30y, middle aged >30y <55y, and old >55y using non-parametric Kruskal-Wallis test. KW p-values were adjusted for multiple-testing across CpGs using Benjamini-Hochberg method. Post-hoc analysis was performed using pair-
wise Wilcoxon test comparing each age group, and Wilcoxon p-values were adjusted for multiple-testing using Benjamini-Hochberg method.

TCGA DNA Methylation Data
Normalized and processed beta values from TCGA were downloaded using TCGAbiolinks package. Analysis was restricted to samples from women with annotated PAM50 breast cancer subtype and age at diagnosis (PAM50 LumA n=137, LumB n=140, Her2 n=43, Basal n=137, Normal n=34). DNA methylation level (beta values) of ELF5 CpG sites in breast cancer tissue across PAM50 subtypes were obtained from this set.

Statistical comparison of DNA methylation of ELF5 CpG sites were performed across the five PAM50 subtypes: LumA, LumB, Her2, Basal and Normal-like using non-parametric Kruskal-Wallis test. KW p-values were adjusted for multiple-testing across CpGs using Benjamini-Hochberg method. Post-hoc analysis was performed using pair-wise Wilcoxon test comparing each PAM50 subtype, and Wilcoxon p-values were adjusted for multiple-testing using Benjamini-Hochberg method.

DNA methylation in high risk LEp
DNA methylation measured using Infinium 450K Methylation and EPICMethylation BeadChips were analyzed using a custom R script. The arrays were read and normalized using the minfi package which return methylation m-values(18). 450k and EPIC array data were normalized by removal of batch effects using ComBat(19). Conversion of methylation beta-values from m-values were carried out using Lumi package. Methylation beta-values, which is an approximation of the percentage of methylation of a given CpG site, and is calculated as the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities).

ELF5 DNAm multiple regression
A multiple linear regression DNAm age predictor was generated based on the ELF5 DNA methylation beta values in average-risk LEps of 5 CpGs selected based on their high correlation to chronological age and anti-correlation to ELF5 expression (Age ~ cg04504043 + cg11875459 + cg21017775 + cg11343506 + cg22731981). This ELF5-
based multiple regression was used to predict the DNAm age of average-risk and high-risk LEps. A separate tissue-based multiple regression was generated based on the DNA methylation beta values of the five selected ELF5 CpGs in normal primary breast tissues from 3 public datasets: GSE101961 (n=121)(17), GSE88883 and GSE74214 (n=100 and n=18)(12). For LEp and primary tissue, ELF5-based DNAm age were then plotted against chronological age, and linear regression R2, R and p-values calculated.

Horvath Clock

DNAm age was calculated for average-risk LEp and for each of 3 publicly available datasets of normal primary breast tissue from the Infinium 450K platform: GSE101961 (n=121)(17), GSE88883 and GSE74214 (n=100 and n=18)(12) based on the 353 CpG Horvath pan-tissue clock(10) using the publicly available code (https://horvath.genetics.ucla.edu/html/dnamage/). Un-filtered and un-normalized DNAm beta values were inputed as required. Data was normalized using BMIQ normalization as provided in the code. DNAm age was computed for each sample. Horvath clock predicted DNAm ages were averaged in average-risk LEp with replicate samples. For LEp and primary tissue, Horvath DNAm age were then plotted against chronological age, and linear regression R2, R and p-values calculated. Error was calculated as the median absolute difference in biological and chronological age. Because the Horvath clock is not compatible with EPIC arrays, DNAm age of average-risk and high-risk LEp from the EPIC platform were not calculated.

MEAT DNAm clock

DNAm age was calculated for average-risk and high-risk LEp and for each of 3 publicly available datasets of normal primary breast tissue from either the Infinium 450K or EPIC platform: GSE101961 (n=121)(17), GSE88883 and GSE74214 (n=100 and n=18)(12) using the Bioconductor MEAT package (http://www.bioconductor.org/packages/release/bioc/html/MEAT.html). Un-filtered and un-normalized DNAm beta values were inputed as required. Data was normalized using BMIQ normalization as provided in the code. DNAm age was computed for each sample. For LEp and primary tissue, MEAT DNAm age were then plotted against chronological age, and linear regression R2, R and p-values calculated. Error was calculated as the median absolute difference in biological and chronological age.
For LEp samples, the MEAT pipeline was adapted to allow for samples distributed across 450K and EPIC arrays to first be combined and ComBat batch-adjusted post-normalization and prior to DNAm age calculation. This was done to closely match the methodology we used in pre-processing and combining the LEp data from the 2 platforms. Briefly, 450K and EPIC dataset were independently cleaned and calibrated using the MEAT pipeline. MEAT generated normalized beta values were then converted to m-values. Each data set was then ComBat batch-adjusted based on the chip IDs. Batch adjusted m-values from 450K and EPIC arrays were then merged based on common probes across platforms. The merged dataset was then further ComBat batch-adjusted based on platform type. Finally, the merged batch-adjusted m-values were converted back to beta values and prepared for DNAm age prediction in MEAT. MEAT clock predicted DNAm ages were averaged in Average-Risk LEp with replicate samples.

**Ethical guidelines**

The patient studies were conducted in accordance with Declaration of Helsinki.

**Results**

**ELF5 is downregulated as a function of age**

To address whether age-associated ELF5 expression in human breast can be used as an aging biomarker, we utilized primary human mammary epithelial cells (HMEC) derived from breast tissues of women who range in age from 16-72 years (Supplementary Table 1)(15,16,20). Next generation single-end RNA sequencing (RNA-seq) was applied to examine transcriptomes in FACS-enriched CD133+/CD271-luminal epithelial cells (LEp) and CD133-/CD271+ myoepithelial cells (MEp) (n=43). Linear models of transcriptomes in LEp and MEp from reduction mammoplasty HMEC at 4th passage (n=11 <30y, n=8 >55y) and uncultured primary epithelial organoids (n=4 <30y, n=3 >55y LEp; n=3 <30y, n=1 >55y MEp) showed high correlation in both age groups (r = 0.96 p < 2.2e-16 in LEp in <30y, r = 0.93 p<2.2e-16 in MEp in <30y, r = 0.97, p < 2.2e-16 in LEp in >55y, r = 0.96, p < 2.2e-16 in MEp in >55y) (Fig. 1A). The magnitude of ELF5 expression in both MEp and LEp were highly correlated between cultured HMEC and organoids (labeled in each plot). ELF5 expression is LEp-specific and higher expression was detected in LEp from women who were <30y compared to LEp from women >55y in both HMEC and organoids (Fig. 1B). These data indicated that
HMEC cultured in the M87A media (16) maintained lineage-specific and age-dependent transcription profiles that are consistent with in vivo. ELF5 expression level as a function of age showed negative correlation in LEp (n=21, r = -0.76, p < 7.2x10^-5) (Fig 1C), but not in MEp (n=21, r=0.089, p=0.7) (Supplementary Figure S1A). ELF5 shows tissue-specific isoform expression (21) so we next examined the possibility of age-dependent isoform expression in LEp. Expression levels of four ELF5 transcript variants in HMEC were examined using paired-end RNA-seq (n=3 <30y, n=3 >55y). ELF5 transcript variant 2 (Accession # NM_001422) had 34- to 62-fold higher expression than other isoforms (p <0.001) and it was downregulated with age (p <0.01) (Supplementary Figure S1B and Fig. 1D). This isoform was expressed in a lineage-specific manner independent of age. We confirmed predominant expression of transcript variant 2 and its age-dependence using semi-quantitative RT-PCR method (p <0.05) (Supplementary Figure S1B-C). Transcript variant 3 expression is lower, but the expression also appeared to be age-dependent. Taken together, ELF5 expression is strongly correlated with age in LEp from both cultured primary HMEC and uncultured organoids.

DNA methylation of the ELF5 promoter is negatively correlated with age-dependent ELF5 expression

ELF5 expression is regulated by DNA methylation on its promoter during mammary gland and embryonic development in mice (22,23). RT-qPCR for ELF5 transcript and a qPCR-based DNA methylation assay using a DNA methylation-dependent endonuclease, McrBC, were used to measure ELF5 expression and DNA methylation in the ELF5 promoter region. In LEp, methylation levels were positively correlated with age (Fig. 2A, black circles, r = 0.8824, p < 1E-4) and RNA levels were negatively correlated (Fig.2A, light blue circles, r=-0.635, p <0.05). In MEp, DNA methylation at the ELF5 promoter was consistently high regardless of age and little transcript was detectable (Fig. 2B). The McrBC-qPCR method is not suitable for identifying specific single CpG dinucleotides. To identify age-associated differential methylation sites with single nucleotide resolution, we utilized Illumina 450K DNA methylation arrays (n=4 <30y, n=4 >55y). The ELF5 promoter and gene body regions were covered by 21 sequence specific probes. Analyses of beta-values (ratio of the methylated probe intensity to overall intensity) showed 20 out of 21 loci were differentially methylated with age in LEp (BH adj.p-val < 0.05, with 17 out of 20 showing
higher significance, BH adj.p-val < 0.001) (Fig. 2C). Correlation of age and DNA methylation levels in each probe showed that probe cg19658620, cg04504043, cg02882375, cg22731981 and cg11875459 had high Pearson’s correlation coefficients (>0.918) and the errors (median absolute differences of chronological age and biological age by ELF5 expression(1)) were 4.4 to 5.6 years (Fig. 2D and Supplementary Table S2). DNA methylation and expression from the ELF5 locus show striking lineage specificity in LEp versus MEp, and in LEp the age-dependent decrease in ELF5 expression may be caused by increased promoter DNA methylation.

Detection of age-related ELF5 changes in bulk breast tissue

We examined DNA methylation and gene expression in publicly available datasets that were derived from whole breast tissue with the intent of extending our understanding of the robustness of this phenomenon. Utilizing the publicly available Illumina 450K DNA methylation array data sets (GSE101961, n=121; and GSE88883, n=100) from discarded reduction mammoplasty, we detected age-associated differential DNA methylation at the ELF5 locus (Fig.3A)(12,17). DNA methylation beta-values from younger (≤30y) women were lower than those from middle age (>30y, <55y) and older (≥55y) women. The baseline DNA methylation beta-values in the younger samples from the ensemble tissue data sets were higher than those from FACS-enriched LEp suggesting the presence of more background signal from more abundant cell types. The decrease in beta-value dynamic range at CpGs in the ELF5 locus from whole tissue can be seen more clearly when viewed as box plots of the individual probes (Fig. 3B), as compared to the range between age-groups in FACS-enriched LEp (e.g. Fig. 2). Whereas the beta-values for the CpG island annotated cg11875459 in FACS enriched LEp covered about 70% of the total range from younger to older, in bulk tissues we only detected changes of <20% of the total possible range with age.

We next asked whether age-dependent ELF5 RNA expression was detectable in bulk tissues from a publicly available data set. The data set, GSE102088 (n=114), was prepared from reduction mammoplasty and is composed of 33, 70, and 11 individuals who were younger (≤30y), middle age (>30y, <55y), or older (≥55y), respectively(17). Decreasing ELF5 expression as a function of age was confirmed even in the whole tissue (Kruskal-Wallis p < 0.001) with significance of post-hoc pair-wise p-values < 0.05 and < 0.001 for younger vs middle age and for middle age vs older, respectively (Fig.
Linear regression showed a weaker correlation between ELF5 expression and age compared to FACS-enriched LEp (Fig. 3D, \( r = -0.3840, p < 1E-4 \)). Robust regression, shown by a red line, revealed 12.3% outliers (14 out of 114) (Fig. 3D, marked by red closed circles). The age-related changes in ELF5 expression and DNA methylation beta-values in whole breast tissue is such that the publicly available data parallel our findings from FACS enriched LEp.

CpG sites of the ELF5 gene locus were not detected as age-associated CpGs in previous reports (12,17). To compare with our findings, we examined three previously identified top age-associated CpGs from bulk tissue in our LEp dataset (Supplementary Figure S2). The two CpG sites (cg07303143 and cg06458239) also showed age-dependent increase of DNA methylation (Supplementary Figure S2A) though the correlations were weaker than that with ELF5 CpG in LEp, with MEp also contributing to the signal (Figure 2C and D). However, cg01271695 on the PXDN gene locus showed poor correlation with age and DNA methylation in LEp in contrast to MEp (Supplementary Figure S2B). The CpG cg01271695 has been identified to be negatively correlated to PXDN expression (12). We neither see age-dependent expression in MEp nor LEp. This difference could be due to the contribution of other cell types in the whole tissue context. We used FACS-enriched lineage specific epithelial cells from HMEC that removed other cell types such as fibroblast and adipose cells in contrast to whole breast tissue preparations used in other studies. This illustrates the need for lineage-specific analysis to elucidate the age-dependent molecular changes that regulate cell type-specific function and that lead to cancer-predisposing dysregulation.

**Relationship between age-related DNA methylation and luminal breast cancers.**

Luminal A and B breast cancer subtype incidence increases with age and comprise 80% of age-related breast cancers (24). Downregulation of ELF5 in PAM50 Luminal subtypes compared to that in normal breast was reported (25) and we provide a confirmatory reanalysis from more than 1000 women with the data generated by the TCGA Research Network in Fig. 4A. We next examined DNA methylation at the ELF5 locus from more than 450 women across BC subtypes in TCGA. DNA hypermethylation was more common in Luminal A, Luminal B, HER2, and normal PAM50 subtypes as compared to Basal (Fig. 4B). DNA methylation levels in the ELF5 gene body regions.
showed more DNA methylation than those in the promoter region. Accordingly, beta values tended to be lowest in Basal BC (Fig. 4C). The higher DNA methylation states in luminal subtype BCs were not age-dependent. This suggests that age-dependent DNA methylation states at the ELF5 locus could be a priming event for luminal subtype BCs, but that once cancer has set in the relationship between age and ELF5 expression or methylation ceases to exist.

**ELF5 down-regulation is accelerated in women with germline pathogenic BRCA1 and BRCA2 mutations**

We next determined whether age-dependent changes in ELF5 expression in LEp followed the same pattern in women at an average level of cancer risk and in women with clinically verified pathogenic germline mutations that significantly increases their lifetime risk for BC. We enriched LEp by FACS from normal tissue of high-risk (HR) individuals who were verified to have germline mutations in BRCA1, BRCA2, or PALB2, and who ranged in age between 31-55 years (sample details in Supplementary Table 1). ELF5 expression in high-risk LEp (n=12) was decreased compared to younger (<30y) average-risk (AR) women (n=21) (p < 0.001) and the magnitude of the decrease in expression was similar to that of older women (>55y) (Fig. 5A). Neither BRCA1, BRCA2, nor PALB2 showed age-dependent differential expression in LEp from Average-Risk women (Supplementary Fig. 3A, B, and C, respectively). Nor did BRCA1, BRCA2, or PALB2 expression show any correlation with ELF5 expression in LEp from average-risk women (Supplementary Fig. 3D, E, and F, respectively). These data suggest that ELF5 is not a direct target of BRCA1, BRCA2 or PLAB2, and vice versa.

We then asked whether normal epithelia from otherwise average-risk individuals diagnosed with BC, but without a known pathogenic germline mutation, showed a decline of ELF5 expression. Those samples, from eight discarded contralateral and two normal peripheral-to-tumor tissues, were examined from women who ranged in age from 25-72y (n=10). There was no difference in ELF5 expression between LEp from tissue that was contralateral or peripheral to tumor and age-matched average-risk tissues, and overall the trends of ELF5 expression tracked with age (Fig. 5A). We established a regression model with the average-risk LEp that showed ELF5 expression is negatively correlated with age ($r = -0.76$, $p = 7.2e^{-5}$)(Fig 5B). Data from high-risk LEps were then fit to the average-risk regression and 92% (11 out of 12) high-risk LEp showed
lower \textit{ELF5} expression than would have been predicted at the same chronological age based on our model (Fig. 5B). Based on \textit{ELF5} expression levels the biological ages of high-risk LEp would be predicted to be 11 to 42 years older than their chronological ages (Supplementary Table S3A). These results suggest that age-dependent decrease in \textit{ELF5} is accelerated in epithelia from women with germline mutations that predispose them to BC.

Next, we examined DNA methylation of the \textit{ELF5} region in high-risk LEp (n=12) with the Illumina EPIC platform (the updated version of 450K). Principal component analysis (PCA) showed that DNA methylation levels around the \textit{ELF5} locus in high-risk LEp were closer to those from average-risk older LEp and separated from those from average-risk younger LEp (Fig. 5C). We then built a multiple regression predictor of biological age based on the methylation levels of five \textit{ELF5} probes in average-risk women. These five probes were selected based on their high correlation with age and anti-correlation with expression in LEp from average-risk (Fig. 2D). The predicted biological ages of average-risk and high-risk women are shown in (Supplementary Table S3B) and are plotted against chronological age (r=0.99, slope m=0.97) (Fig. 5D). Predicted age of 64\% (nine out of 14) of the high-risk samples based on the multiple regression model shifted above the upper 95\% confidence interval (CI) suggesting advanced biological age relative to chronological age. Individual linear regression models of DNA methylation level of each of the 5 probes vs. chronological age likewise show the majority of high-risk LEp are shifted above the 95\% CI (Supplementary Fig. 4A). However, one or two high-risk samples in each probe were located below lower 95\% CI that indicated younger biological age. For each sample, we calculated the absolute difference between chronological age and the predicted biological age from the average-risk multiple regression model. There is a significantly greater error in high-risk LEp (error=15.9y) compared to average-risk (error=2.9y) (Fig. 5E), which suggests increased variance in methylation at these sites is a property associated with cancer risk.

Because a majority of the high-risk samples did show significantly decreased \textit{ELF5} expression, we examined the correlation of matched \textit{ELF5} expression and DNA methylation in the high-risk LEp (n=7) as a function of the five chosen probes (Supplementary Fig. 4B). These regression models indicated that 71–86\% of High-risk women tend to be outside of the 95\% CI for any given regression compared to 0–25\% of
average-risk. This suggests that the high-risk LEp may have decoupled the relationship between *ELF5* expression and methylation. Taken together, quantification of *ELF5* expression and DNA methylation can be used as biological clocks for human mammary epithelia. Moreover, biological age estimates relative to chronological ages of high-risk LEp showed increased variance relative to LEp from average-risk, further suggesting dysregulation of the LEp lineage in high-risk women. We hypothesize these *ELF5*-based clocks may be used to estimate breast cancer risk in a manner that is independent of the specific underlying monogenic risk factor.

Lastly, we compared how established DNAm clocks developed from bulk tissue predict biological age of isolated cell lineages. We applied the Horvath pan-tissue clock (10) to the average-risk LEp from the 450K array (the Horvath clock is not compatible with EPIC arrays). Likewise, we adapted the MEAT DNAm clock (26) with ComBat batch-adjustment post calibration to the average-risk and high-risk LEp data from the combined 450K and EPIC arrays. We found both clocks calculated DNAm ages of LEp samples with strong correlation to chronological age (*r*=0.96, *r*=0.92 respectively), but with a slope much less than 1 (*m*=0.65 and *m*=0.51, respectively) leading to high errors in predicting biological age in average-risk LEp samples (error=8.3y, error=9.4y, respectively) compared to the *ELF5*-based clock (*r*=0.99, *m*=0.97, error=2.9y) (Supplementary Fig.5A-5B). Based on the linear regression of chronological vs. biological age generated from average-risk LEp, the MEAT clock likewise predicts older biological ages in 8 of 14 high-risk LEp (error=7.2y) (Supplementary Fig.5B). The high error in age estimation suggests that tissue-based clocks are inadequate in capturing the age-dependent changes in the luminal lineage, and are susceptible to changes in composition of the breast. Indeed, differential age acceleration in LEp may partially explain why the Horvath clock was shown to be poorly calibrated in breast tissue with high error of 8.9y in normal tissue (cor *r*=0.73) and error of 13y in normal adjacent to tumor (cor *r*=0.87) (10). We validated this finding using 3 publicly available datasets of normal primary breast tissue from the Infinium 450K platform: GSE101961 (*n*=121)(17), GSE88883 and GSE74214 (*n*=100 and *n*=18)(12). Predicted biological ages of normal breast tissue recapitulated previous results showing high error despite strong correlation with chronological age (*r*=0.88, error=12y, and *r*=0.78, error=7.3y respectively) with the Horvath clock leading to an overestimate of biological age particularly in young women (Supplementary Fig.5C-5D). Moreover, multiple regression performed on the normal
whole breast tissue based on DNA methylation of the five selected ELF5 probes shows ELF5-based predicted biological age have weak correlation to chronological age ($r=0.33$, error=9.5y) (Supplementary Fig.5E) likely due to contamination by other cell types that mute the ELF5 signal from LEp. Together, these results indicate that biological age in breast tissue is complicated by many other factors and a lineage-specific clock is needed to investigate the effects of age acceleration on individual cell types that may contribute differentially to aging-associated cancer initiation.

Discussion

The overwhelming majority of breast cancers (BCs) are not attributable to monogenic germline risk factors, and may instead be attributed to polygenic, epigenetic, and/or environmental risk factors. There is a need to develop reliable tools for breast cancer risk assessment that will bolster early detection and prevention efforts. Here we show that ELF5 is expressed in an age-dependent manner in normal whole breast tissue and in LEp, and that the changes in expression are due to changes in promoter methylation that also are age-dependent. In luminal subtype BCs, ELF5 expression is exceedingly low concomitant with high levels of promoter methylation. Using average-risk epithelial cells to establish a model that relates chronological age to ELF5 expression or methylation, we found that LEp from women with high-risk germline mutations in BRCA1 and/or BRCA2, or PALB2 show an accelerated aging phenotype based on ELF5 expression and DNA methylation in the promoter proximal region. We conclude from these data that ELF5 expression and DNA methylation of its promoter constitute breast-specific aging biomarkers, or molecular clocks. We speculate that deviations from a model of chronological age based on ELF5 expression or promoter methylation that is established from phenotypically normal average-risk breast will reveal a tissue’s biological age. We hypothesize that significant deviation of biological age from chronological age is a property of tissue that is more susceptible to cancer initiation.

A number of molecular markers have been proposed for breast cancer risk assessment. Concentration of IGF-1 or sex hormones such as estradiol and testosterone in blood are associated with breast cancer risk(27,28). The relationship between risk and IGF-1 concentration is detected only in premenopausal women and there is no standardized method to measure blood levels of IGF-1. The risk assessment by concentration of sex hormones is suitable for postmenopausal women, but not for all
ages. Shortening of telomere length with age has been reported(29), and cancers have short telomeres, but a strong case cannot be made that age-dependent shortening occurs in normal breast epithelia(30). Telomerase activity in normal cells is lower than that in cancer cells, but it would be challenging to detect quantitatively with limited sample quantity. Thus, telomere length and telomerase activity are not good candidates for BC risk assessment. The relationship between DNA methylation and aging has been well discussed(7,8,31). Epigenetic clocks (DNA methylation age) have been proposed to estimate biological age by Horvath and others(9,10,32-34). The pan-tissue clock uses 353 CpG sites to estimate biological age in a number of tissues and showed an association between estimated biological age with a number of age-related diseases; however, the pan-tissue clock was poorly calibrated in predicting biological age of breast tissue leading to high errors (median absolute difference between biological chronological age) despite high correlation values (10). We found a similar magnitude of error between biological and chronological age despite high correlation values when we applied both the Horvath and MEAT clocks (another established bulk tissue-based DNAm age predictor) to 239 normal primary breast tissue from 3 publicly available datasets(12,17). This suggests that biological age in breast tissue is complicated by other factors including changing breast composition and hormone profiles. Breast-specific epigenetic clocks were developed that predicted accelerated biological age in luminal breast cancers relative to normal tissue(11,35). Aging changes overall breast tissue composition such that adipose increases, connective tissue decreases, and proportions of MEp decrease relative to an increase of LEp and epithelial progenitors in the epithelia(20). An example of the effect of these compositional changes was embodied by our observation that the dynamic range of \textit{ELF5} gene expression and methylation changes with age was as much as five-fold greater in purified LEp versus whole breast tissue. Indeed, the tissue-based clocks were not suited to predict biological age of LEp or to capture age-dependent changes in the luminal lineage. Lineage-specific age acceleration and its impact on cancer initiation may therefore be better measured by clocks developed specifically for a given cell type. Further studies will help to establish whether the estimated biological ages by the ELF5 clock is due to changes in DNA methylation or tissue components with age.

Random periareolar fine needle aspiration (RPFNA) or ductal lavage (DL) are minimally invasive clinical techniques for obtaining small amounts of breast tissue to
assess breast cancer risk in asymptomatic women who are suspected of being at increased risk due to germline inherited risk factors or family history(36). Materials obtained by either method can capture mixtures of epithelial, fibroblast, endothelial, adipose and immune cells (although DL is highly enriched for sloughed off epithelial cells). We confirmed age-associated ELF5 expression and DNA methylation was measurable even from normal whole breast tissue. This is important, because it seems likely that specimens collected by RPFNA and DL will be amenable to ELF5 analysis, thus helping to pave the way for testing in a translational setting. From publicly available data derived from whole tissue we observed that 12% (14 in 114) of women showed significantly lower ELF5 expression compared to what would have been predicted by chronological age - BC incidence in the U.S. is ~13%. We were unable to obtain follow-up information on the patients embodied in the GSE102088 dataset, but it is tempting to speculate that it may be possible to identify the minority of women who are at highest risk (independently of known genetic risk factors) based on regulation of ELF5 in breast.

Tools are needed to identify and predict vulnerable subgroups at risk for developing BC. Measures of biological age by the ELF5 clock represents a potential way to assess emergent properties of the aging process that are linked to increased BC risk. The ability to advance knowledge of the effects of aging on cancer has been constrained by a paucity of an agreed upon biological measure or a combination of biological measures of aging that are sensitive or specific enough to accurately assess the physiological aging process(37-40). Biomarkers that are currently available are unable to precisely measure an older individual’s physiological or functional age. However, in the future, biomarkers such as the ELF5 clock, in combination with clinical aging assessments like the geriatric assessment(41-44) could enhance our understanding of how these factors interact and contribute to cancer risk. Although measures of biological aging are still in their infancy, they demonstrate translational promise to better understand biological aging in humans, and more research is needed to develop and validate these tools.

Downregulation of ELF5 in non-human primate mammary glands occurs from the adult luteal to the postmenopausal stage(45), suggesting that downregulation of ELF5 is conserved as an aging mechanism. Accelerated ELF5 downregulation may be a sign of accelerated aging in breast tissue indicating greater than average susceptibility to cancer initiation. ELF5 expression is lower in luminal A and B BC subtypes than normal
tissue(25) and ELF5 downregulation was reported in all stages of cancer progression including atypical ductal hyperplasia, ductal carcinoma in situ, and invasive ductal carcinoma(46). We also observed decreased ELF5 levels and sites with increased DNA methylation in the ELF5 locus in luminal A and B BC subtypes, which are strongly age-associated cancers. We do not know whether ELF5 downregulation with age is a cause or a consequence of breast tissue aging. ELF5 is a transcription factor and regulates stem/progenitor cell fate decisions in mammary gland development(47). Dysregulation of ELF5 affects the expression of downstream target genes in epithelial cells including ESR1 and FOXA1(13,25). In triple negative breast cancer (TNBC), ELF5 seems to work as tumor suppressor to inhibit epithelial-mesenchymal transition by suppressing SNAI2(48) and loss of ELF5 enhances IFNGR1 expression that leads to tumor growth and metastasis(49). Thus, decline of ELF5 expression affects transcriptome networks that are associated with BC progression and subsequently could impact cellular function in a way that increases susceptibility to cancer initiation though we do not yet know the mechanism.

The ELF5 expression clock was generally consistent in identifying high-risk LEp by their accelerated decrease in expression relative to the model predicted by average-risk LEp. In contrast, the direction of change in methylation at any of the five CpG in high-risk samples was not consistent. Any given high-risk sample showed a change consistent with accelerated aging in only a subset of the five CpG islands that we evaluated for the clock (either based on the CpG-specific linear models or the multiple regression combining these CpGs). Indeed, the characteristic that distinguishes high-risk and average-risk samples according to the ELF5 methylation clock is increased variation in the methylation levels across the five selected CpG sites in the high-risk samples. Because control of ELF5 transcription is a composite outcome of the states of multiple promoter-proximal CpG sites it may not be surprising that changes in transcription were mainly in one direction. Comparing the expression and methylation clocks in their ability to distinguish the high-risk samples, we identified 92% of high-risk samples with the gene expression model and 64.3% of high-risk with the gene methylation model.

Age-related gene expression in breast is reported to be influenced by hormone changes(50). ELF5 was downregulated in the high-risk tissues regardless of age in our data set, thus even in women who were young and premenopausal. This suggests that
decreased ELF5 expression and changes in DNA methylation are more specifically related to aging and tissue level changes that presage cancer susceptibility, in a manner that may not be dependent on hormone changes. Reported age-related genes such as ESR1 or FOXA1 are direct target genes of ELF5 (25). We have reported that aging phenotypes in LEp are imposed by age-dependent cellular microenvironments(13). Therefore, estrogen signaling alone seems insufficient to explain the alteration of gene expression in breast as a function of age. We do not exclude the possibilities that BRCA1, BRCA2 and PALB2 directly regulate ELF5 expression though our data suggested no correlation. Both BRCA1 and ELF5 gene products reportedly repress ERα activity through different mechanisms(25,51). Accumulation of progenitors with more basal- or MEp-like features seems to be a phenotype in common amongst breast tissue of BRCA1-mutant (BRCA1mut) carriers, ELF5-deficient mouse mammary glands, and breast tissue from older women(20,47,52). Thus, BRCA1mut carriers may share a mechanism also present in ELF5-downregulated breast tissue from older women.

BRCA1mut carriers and younger women tend to develop TNBC(24,53), whereas aging is mainly related with luminal subtype breast cancers. BRCA2mut carriers tend to develop luminal B subtypes. PALB2mut carriers show higher rates for both luminal and TNBC subtypes(54) and this is perhaps due to the features of PALB2 that interact with BRCA1 and BRCA2(55). ELF5 expression and ELF5 DNA methylation clocks can be used to detect accelerated aging in both average-risk and high-risk women. However, ELF5 expression tends to be higher and DNA methylation is lower in TNBC compared to luminal subtypes BCs. This suggests that the regulation of ELF5 gene in normal and cancer cells does not have the same mechanism, or that the difference between subtype may reflect cell of origin. Re-activation of ELF5 expression by an unknown mechanism has been reported after tumors acquired drug resistance to tamoxifen (25,56), again underscoring that ELF5 regulation is likely quite different in the normal or cancer contexts.

The cell-of-origin of TNBC from BRCA1mut carriers are thought to be cKIT+ luminal progenitors(52). Luminal progenitors also may be cell-of-origin of luminal subtype BCs(57). It has been reported that BRCA1mut carriers increase SNAI2 expression that inhibits luminal lineage commitment(58). BRCA1 does not directly regulate SNAI2 transcription, however, BRCA1 seems to contribute to stabilization of the SNAI2 gene product, the SLUG protein. Our data showed that ELF5 expression was
downregulated in $BRCA1^{\text{mut}}$ carriers, and Chakrabarti et al showed that $ELF5$ suppressed SLUG expression\cite{48}. This might be a potential mechanism relating to the susceptibility to develop TNBC in $BRCA1^{\text{mut}}$ carriers. However, a recent study indicates that the intrinsic states of each epithelial lineage have dominant contribution to tumor subtype\cite{59}, suggesting age-associated luminal subtype breast cancers originate from more mature LEp. Signaling pathways, epigenetic regulation such as microRNAs and/or dysregulated genes influence the determination of distinct BC subtypes\cite{60}. Indeed, age and the way cells bypass stress-associated stasis independently leads to distinct subtypes during the transformation from normal to immortal cells\cite{61}. ELF5 is a key regulator for mature LEp and luminal progenitors. Therefore, decline of $ELF5$ expression may impact maintenance of LEp lineage fidelity and the resulting age-dependent cellular microenvironment may be more conducive to cancer initiation.
Acknowledgements

We would like to acknowledge the contributions of our patient advocates, Susan Samson and Sandy Preto. Research reported in this publication included work performed in the Analytical Cytometry, Integrative Genomics and Bioinformatics, and Biomarker Analysis and Validation Cores supported by the National Cancer Institute of the National Institutes of Health under grant number P30CA033572. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The results shown and referenced herein are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

We are grateful for the generous support received from: Phi Beta Psi Sorority, California (MMiyano). The National Institutes of Health: NCI, Cancer Metabolism Training Program Postdoctoral Fellowship T32CA221709 (RSayaman), U01CA244109 and R01CA237602 (MLaBarge); NIBIB R01EB024989 (MLaBarge); NIA R33AG059206 (MLaBarge); the Department of Defense/Army Breast Cancer Era of Hope Scholar Award BC141351 and Expansion Award BC181737, Conrad N. Hilton Foundation, and City of Hope Center for Cancer and Aging (MLaBarge). NIH grants P30CA034196 (OAnczukow) and T32AG062409A (BAngarola)

References


30. Sputova K GJ, Pelissier FA, Chang E, Stampfer MR, Labarge MA. Aging phenotypes in cultured normal human mammary epithelial cells are correlated with decreased telomerase activity independent of telomere length. Genome Integr 2013


40. Margolick JB, Ferrucci L. Accelerating aging research: how can we measure the rate of biologic aging? Experimental gerontology 2015;64:78-80
42. Soto-Perez-de-Celis E, Li D, Yuan Y, Lau YM, Hurria A. Functional versus chronological age: geriatric assessments to guide decision making in older patients. The Lancet Oncology 2018;19:e305-e16
28


qRT-PCR

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPS18</td>
<td>5'-GGGCCGGGCAAATAG-3'</td>
</tr>
<tr>
<td>RPS18</td>
<td>5'-CCGCCCTTGTGAGGT-3'</td>
</tr>
<tr>
<td>ELF5</td>
<td>5'-TAGGGAACAGGAAATTTCGGG-3'</td>
</tr>
<tr>
<td>ELF5</td>
<td>5'-GTACACTAACCTCGGTCAACC-3'</td>
</tr>
</tbody>
</table>

RT-PCR for isoform detection

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF5_ex1A_F1</td>
<td>5'-CTTGCCTTGAAAGCCTCCTC-3'</td>
</tr>
<tr>
<td>ELF5_ex5_R1</td>
<td>5'-ACAGTCTGTATTTTGATGCCA-3'</td>
</tr>
<tr>
<td>GAPDH_F1</td>
<td>5'-AAGGTGAAGGTCGGAGTCAACCGG-3'</td>
</tr>
<tr>
<td>GAPDH_R1</td>
<td>5'-CCACTTGATTTTGGAGGATCT-3'</td>
</tr>
</tbody>
</table>

McrBC DNA methylation detection

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMP3</td>
<td>5'-TGTAATTCCCCCTCTGG-3'</td>
</tr>
<tr>
<td>TIMP3</td>
<td>5'-GTTGGCCCTTTCAGCAAGTTC-3'</td>
</tr>
<tr>
<td>ELF5</td>
<td>5'-GCCGTGCAGTGAAATAAGAC-3'</td>
</tr>
<tr>
<td>ELF5</td>
<td>5'-CACACTGTATGTCAACGTCATC-3'</td>
</tr>
</tbody>
</table>

Table 1. Description of ssDNA oligo primers

Figure legends

Figure 1. ELF5 downregulation with age in both breast tissue organoids and HMEC. (A) Linear regression plots of transcriptomes between organoids and HMEC in <30y and >55y LEp and MEp. X- and Y- axis show mean rlog values across subjects for 30,079 mapped transcripts from organoids (n=4 <30y, n=3 >55y LEp; n=3 <30y, n=1 >55y MEp) and 4th passage HMECs (n=11 <30y, n=8 >55y LEp/MEp) respectively for each lineage and age group. ELF5 is indicated in each lineage and age group. Transcriptomes between organogans and HMEC are highly correlated in both lineages (r = 0.96 p < 2.2e-16 in LEp in <30y, r = 0.93 p < 2.2e-16 in MEp in <30y, r = 0.97, p < 2.2e-16 in LEp in >55y, r = 0.96, p < 2.2e-16 in MEp in >55y). (B) Preserved in vivo lineage-specific (<30y LEp vs. <30y MEp) and age-dependent (<30y LEp vs. >55y LEp) ELF5 expression in HMEC. Y-axis indicates regularized log (rlog) gene expression in box plots. ELF5 is LEp-specific and the expression is downregulated with age in both organoids and HMEC. Differential expression (limma) adj. p-values are shown for 4th
Figure 2. DNA methylation in *ELF5* promoter region is negatively correlated to lineage-specific and age-dependent *ELF5* expression. Anti-correlation of DNA methylation to lineage-specific and age-dependent *ELF5* expression in LEp (A) and MEp (B). Left y-axis, and right y-axis show percent of cells with methylated DNA and *ELF5* gene expression, respectively. Black and light blue circles show DNA methylation and gene expression plots, respectively. DNA methylation was measured by McrBC digestion followed by qPCR. Gene expression was measured by qPCR and data was normalized by RPS18 expression. Significances of differential gene expression and DNA methylation changes with age in LEp are < 0.001 (Mann-Whitney test). Not significant in MEp. n=14 (n=7 in each <30y and >55y age group) in both A and B. (C) DNA methylation in *ELF5* region using Infinium 450K DNA methylation array (n=4 <30y, n=4 >55y). Differential DNA methylation states in each probe site are shown by beta-values of DNA methylation for <30y (light green) and >55y (dark green). Values 0.0 and 1.0 of Beta-values indicate hypo- and hyper-DNA methylation, respectively. Chromosomal mapping of each CpG site is shown by solid lines below *ELF5* map from USCS genome browser. Promoter region in two major isoforms of *ELF5* are marked by red and light blue lines, respectively. Significance levels of differential methylation analysis (limma) adj. p-values are denoted by asterisks: Benjamin-Hochberg, BH-adj-p-val: *< 0.05, **< 0.01, *** < 0.001, **** < 0.0001. (D) Linear regression of 450k DNA methylation array
probes in *ELF5* gene locus. Pearson’s correlation coefficient (*r*), *p* value and error of five probes are summarized in table. Errors were calculated by median of absolute differences between biological age and chronological age. Corresponding genomic loci of the probes are shown by red dots in C.

**Figure 3.** Age-dependent *ELF5* expression and DNA methylation are detectable with bulk breast tissues. Age-dependent differential DNA methylation analysis in *ELF5* gene locus with publicly available data sets of 450K DNA methylation array, GSE101961 (n=121, ≤30y n=38, >30y <55y n=71, ≥55y n=12) and GSE88883 (n=100, ≤30y n=36, >30y <55y n=53, ≥55y n=11), from reduction mammoplasty tissues are shown in (A) as line graph and (B) as box plot. DNA methylations in *ELF5* gene locus were increased with age in both data sets. Chromosomal mapping of each CpG site is shown by solid lines below *ELF5* map from USCS genome browser. Y-axis: DNA methylation beta values. Kruskal-Wallis test adj. *p*-value significance annotated in (A), pair-wise post-hoc Wilcoxon test adj. *p*-value significance annotated in (B). Significance of differential DNA methylation change with age are denoted by asterisks: Benjamin-Hochberg, BH-adj-*p*-val: *< 0.05, ** < 0.01, *** < 0.001, **** < 0.0001. (C) Age-dependent *ELF5* expression with publicly available data set, GSE102088, from reduction mammoplasty tissues (n = 114, ≤30y n=35, >30y <55y n=68, ≥55y n=11). Kruskal-Wallis test *p*-value annotated, pair-wise post-hoc, pair-wise post-hoc Wilcoxon test adj. *p*-value significance annotated: * * p < 0.05, ** * p < 0.01, *** * p < 0.001. (D) Correlation plots with the same data set in C. n=114, *r* = -0.3840, *p* < 1e-4. Linear regression fit and 95% confidence intervals are shown in black solid and dashed lines, respectively. Robust regression is shown in the red line and outliers shown by the red circle are detected (14 out of 114). GEO data set accession numbers are shown in either bottom or left in each graph.

**Figure 4.** Breast cancer molecular subtype specific *ELF5* expression and DNA methylation. *ELF5* expression (A) and DNA methylation status (B) of *ELF5* gene locus with TCGA dataset. *ELF* expression was lower in LumA, LumB and Her2+ cancer subtypes, whereas the expression was higher in Basal-like cancer comparing to that in Normal. DNA methylation states in *ELF5* gene locus of each subtypes were negatively
correlated to the *ELF5* expression. Breast cancer molecular subtypes are shown according to PAM50 classification. Color scales based on DNA methylation beta-values are shown below the heatmap. Age is shown in the left of heatmap. (C) Differential DNA methylation in *ELF5* gene with PAM50 subtype. DNA methylation beta-values in each probe from every subtypes are superimposed. Chromosomal mapping of each CpG site is shown. Y-axis shows DNA methylation beta-values. Significance of differential DNA methylation change with age are denoted by asterisks: Benjamin-Hochberg, BH-adj-p-val: *< 0.05, **< 0.01, *** < 0.001. LumA n=566, LumB n=207, Her2 n=82, Basal n=194, Normal n=40 for *ELF5* expression data set. LumA n=137, LumB n=140, Her2 n=43, Basal n=137, Normal n=34 for DNA methylation analysis.

**Figure 5. Accelerated decline of *ELF5* expression and increased DNA methylation in high-risk women.** (A) Grouped dot plot showing *ELF5* expression in LEp from average-risk and high-risk women. Contralateral/Peripheral are LEp derived from normal tissue that was contralateral or peripheral to a tumor. Y-axis shows rlog *ELF5* expression. P-values by one-way ANOVA with Tukey’s multiple comparison test were shown above on the dot plots. n=43 (Average Risk=21, black dots; Contralateral/Peripheral=10, gray dots; High Risk=12, red and blue dots) (B) Linear regression of *ELF5* gene expression and chronological age. Solid line and dashed lines show regression fit and 95% CI, respectively. Mutation types in HR are indicated by colors indicated on the graph. n=43 (Average Risk=21, Contralateral/Peripheral=10, and High Risk=12) (C) Dot-plot of the first two principal components based on the DNA methylation states of 21 probes in the *ELF5* gene locus. Each dot corresponds to an individual woman. Young or Old from AR and HR women are shown by the colors indicated on the graph. n=22 (Average Risk=8 and High Risk=14). (D) Predicted biological age of AR and HR based on a multiple regression generated with AR women based on the DNA methylation levels of 5 *ELF5* CpGs: cg19658620, cg04504043, cg02882375, cg22731981 and cg11875459. Predicted biological age is plotted against chronological age. Linear regression line with standard error is shown for AR women. Each point corresponds to an individual woman. n=22 (Average Risk=8 and High Risk=14). Tissue and Mutation types are labeled by different shapes and colors, respectively, as indicated on the graph. R square, correlation coefficient, p
value and regression equation are shown in the graph. Diagonal with slope=1 shown for reference. (E) Comparison of errors between Average Risk and High Risk women in predicted DNA methylation age. Errors (absolute difference between biological and chronological age) from the multiple regression based on the 5 ELF5 probes are plotted. 

$P$-value is calculated by two-sided t-test.
Fig. 1

A

<table>
<thead>
<tr>
<th></th>
<th>LEp &lt;30y</th>
<th>MEp &lt;30y</th>
<th>LEp &gt;55y</th>
<th>MEp &gt;55y</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.96, R² = 0.93, p&lt;2.2x10⁻¹⁶</td>
<td>0.93, R² = 0.85, p&lt;2.2x10⁻¹⁶</td>
<td>0.97, R² = 0.94, p&lt;2.2x10⁻¹⁶</td>
<td>0.91, R² = 0.91, p&lt;2.2x10⁻¹⁶</td>
</tr>
</tbody>
</table>

B

- HMECs
  - ELF5 Log Expression
  - p = 2.6x10⁻¹⁵
  - p = 0.0005

- Organoids
  - ELF5 Log Expression
  - p = 0.0006
  - p = 0.18

C

- ELF5 Log Expression vs Age
  - <30y
  - ≥30y and <55y
  - ≥55y
  - r = -0.76, p = 7.2e-05

D

- Gene Expression (FPKM)
  - p < 0.01
Breast specific molecular clocks comprised of ELF5 expression and promoter methylation identify individuals susceptible to cancer initiation

Masaru Miyano, Rosalyn W Sayaman, Sundus F Shalabi, et al.

Cancer Prev Res Published OnlineFirst June 17, 2021.