Clinical Epigenetics

RESEARCH

Open Access

Aberrant epigenetic and transcriptional events associated with breast cancer risk

Natascia Marino[1,2*], Rana German[1], Ram Podicheti[3], Douglas B. Rusch[3], Pam Rockey[1], Jie Huang[3], George E. Sandusky[4], Constance J. Temm[4], Sandra Althouse[5], Kenneth P. Nephew[6], Harikrishna Nakshatri[7], Jun Liu[3], Ashley Vode[1], Sha Cao[5] and Anna Maria V. Storniolo[1,2]

Abstract

Background: Genome-wide association studies have identified several breast cancer susceptibility loci. However, biomarkers for risk assessment are still missing. Here, we investigated cancer-related molecular changes detected in tissues from women at high risk for breast cancer prior to disease manifestation. Disease-free breast tissue cores donated by healthy women ($N = 146$, median age = 39 years) were processed for both methylome (MethylCap) and transcriptome (Illumina's HiSeq4000) sequencing. Analysis of tissue microarray and primary breast epithelial cells was used to confirm gene expression dysregulation.

Results: Transcriptomic analysis identified 69 differentially expressed genes between women at high and those at average risk of breast cancer (Tyrer-Cuzick model) at FDR < 0.05 and fold change ≥ 2. Majority of the identified genes were involved in DNA damage checkpoint, cell cycle, and cell adhesion. Two genes, FAM83A and NEK2, were overexpressed in tissue sections (FDR < 0.01) and primary epithelial cells ($p < 0.05$) from high-risk breasts. Moreover, 1698 DNA methylation changes were identified in high-risk breast tissues (FDR < 0.05), partially overlapped with cancer-related signatures, and correlated with transcriptional changes ($p < 0.05$, $r \leq 0.5$). Finally, among the participants, 35 women donated breast biopsies at two time points, and age-related molecular alterations enhanced in high-risk subjects were identified.

Conclusions: Normal breast tissue from women at high risk of breast cancer bears molecular aberrations that may contribute to breast cancer susceptibility. This study is the first molecular characterization of the true normal breast tissues, and provides an opportunity to investigate molecular markers of breast cancer risk, which may lead to new preventive approaches.

Keywords: Cancer risk, Transcriptome, DNA methylation, Normal breast

Background

Genetic and epigenetic alterations in breast cancer (BC) have been widely investigated. However, when, during the carcinogenesis process, these events first emerge remains unknown. The identification of molecular aberrations associated with BC development can provide a conceptual framework for a deeper understanding of this complex disease.

Genome-wide association studies (GWAS) have detected more than 170 genomic loci harboring common variants associated with BC risk including modifier alleles with high (e.g., BRCA1, BRCA2, TP53, PTEN) to moderate penetrance (e.g., BRIP1, CHEK2, ATM, and PALB2) [1–4]. Nevertheless, many variants are located in noncoding or intergenic regions and their functional role in cancer transformation remains largely unknown. Recently, transcriptome-wide association

*Correspondence: marinon@iu.edu
[2] Department of Medicine, Hematology/Oncology Division, Indiana University School of Medicine, Indianapolis, IN 46202, USA
Full list of author information is available at the end of the article

studies were used to integrate GWAS and gene expression datasets and identified 154 genes whose predicted expression associated with the risk for BC [5–9]. However, these studies drew data from the Genotype-Tissue Expression (GTEx) project, and, because of the use of autopsy-derived normal breast tissues, the breast-specific transcriptome profilings may be questionable. The relative lack of molecular profiling of normal breast tissue from subjects who are disease-free makes the field challenging.

Many studies searching for cancer biomarkers have identified gene expression signatures, epigenetic signatures, loss of heterozygosity and allelic imbalance resulting from the development of malignancy [10]. Among the molecular processes linked with cancer, DNA methylation has a key role in early cancer development through a process known as epigenetic reprogramming [11], potentially leading to silencing and loss of expression of tumor suppressor genes [12], and genomic instability [13].

Here, we performed an integrated analysis of DNA methylation and transcriptome profiling of cancer-free breast tissues donated by healthy women at either average or high risk for BC. In addition to early epigenetic events, we identified two molecules, FAM83A and NEK2, overexpressed in high-risk breasts and, therefore, potential markers of BC susceptibility. Moreover, using a subcohort of repeated breast tissues donation by the same donors, we confirmed that the molecular changes identified in high-risk subjects are age-independent. These findings will lead to a deeper understanding of BC susceptibility and also provide the scientific community with the molecular profiling of the true normal breast tissue.

Results

Study cohort used to investigate molecular aberrations in association with breast cancer (BC) risk

To identify transcriptomic and epigenetic differences linked with BC risk, we analyzed cancer-free breast tissue cores donated by 146 healthy women (median age: 39 years), including 112 Caucasian, 24 African American, and 10 Asian subjects (Additional file 1: Table S1). Out of 146 participants, 117 were pre- and 29 post-menopausal women. Tyrer-Cuzick model was employed to estimate the lifetime risk of developing BC and allocated the subjects into either high- (score $\geq$ 20%, $N=68$) or average-risk group (score < 20%, $N=78$) (Fig. 1A, Table 1 and Additional file 1: Table S1).

Characterization of the transcriptome alterations in high-risk breast

We performed a transcriptome analysis of the fresh frozen disease-free breast tissue donated by all the participants. Differential expression analysis was performed using DESeq2. From a total of 22,344 genes, the differential expression analysis between high- and average-risk breasts revealed 1874 transcripts to be significant at 5% false discovery rate (FDR). Of these, 1798 transcripts also passed the cutoff of *t*-test *p*-value $\leq$ 0.05 (Additional file 1: Table S2). Sixty-nine genes, including 51 upregulated and 18 downregulated genes, were identified with a fold change $\geq$ 2 (Table 1). Because both groups consisted of non-malignant breast tissue, a limited number of differentially expressed genes was expected [14]. Canonical pathway analysis revealed enrichment in pathways related to kinetochore signaling ($p=1.3E-05$), DNA damage checkpoint ($p=0.0005$), granulocytes adhesion ($p=0.002$), and the IL17 pathway ($p=0.004$) (Fig. 1B, Additional file 1: Table S3). Our data further confirm the impact of dysregulated DNA damage in breast carcinogenesis, as previously described [15]. Molecular network analysis showed an enrichment in functional categories involved in cell cycle, DNA replication and repair (Fig. 1C, Additional file 1: Table S3). One of the major molecular networks regulating cell cycle is centered around AKT and the transcription factor FOXM1 [16].

Except for DCX, the transcriptional changes detected between high- and average-risk breasts listed in Table 2

(See figure on next page.)

Fig. 1 Transcriptome profiling of breast tissues from women at either high- or average risk of breast cancer. **A** Schematics of the study design. Cancer-free breast tissue cores ($N=146$) were divided in either high-risk or average-risk group according to the Tyrer-Cuzick breast cancer risk evaluation score (20% used as threashold). The tissues were processed and analyzed for whole transcriptome and methylome profiling and differentially expressed genes (DEG) and differentially methylated sites between high- and average-risk samples were identified. Thirty five women (10 high risk and 25 average risk) donated also a second biopsy (D2) allowing to detect age-dependent aberrations. **B** Pathway analysis of the transcripts differentially expressed (FDR < 0.05) between average and high- risk breasts. **C** Major molecular network of the differentially expressed transcripts between the two experimental groups. Genes upregulated in high-risk breasts are in red, while downregulated genes are in green. **D** FAM83A and NEK2 transcription level in breast tissues from women at either average- or high-risk of developing breast cancer. **E** Upper panel: Representative image of the immunofluorescence staining of primary breast epithelial cells with the epithelial marker, E-Cadherin (red), mesenchymal marker, Vimentin (green) as control, and nuclear staining, DAPI (blue). E-Cadherin and Vimentin staining of primary cells revealed that isolated primary cells are epithelial in nature. Lower panel: FAM83A and NEK2 expression in primary epithelial cells isolated from either average-risk ($n=4$) and high-risk breast ($n=3$). **F** FAM83A and NEK2 expression in primary and h-TERT immortalized isogenic breast epithelial cells ($n=7$) from the GSE108541 dataset. **G**) Representative images of immunohistochemical staining for FAM83A and NEK2 are shown at 20X magnification. Staining quantification is expressed as positivity and H-score. Data are shown as mean ± standard error. #FDR < 0.005, *$p<0.05$,**$p<0.001$, ***$p<0.0001$. *P*value is calculated using either unpaired nonparametric Mann–Whitney test or nonparametric Wilcoxon test

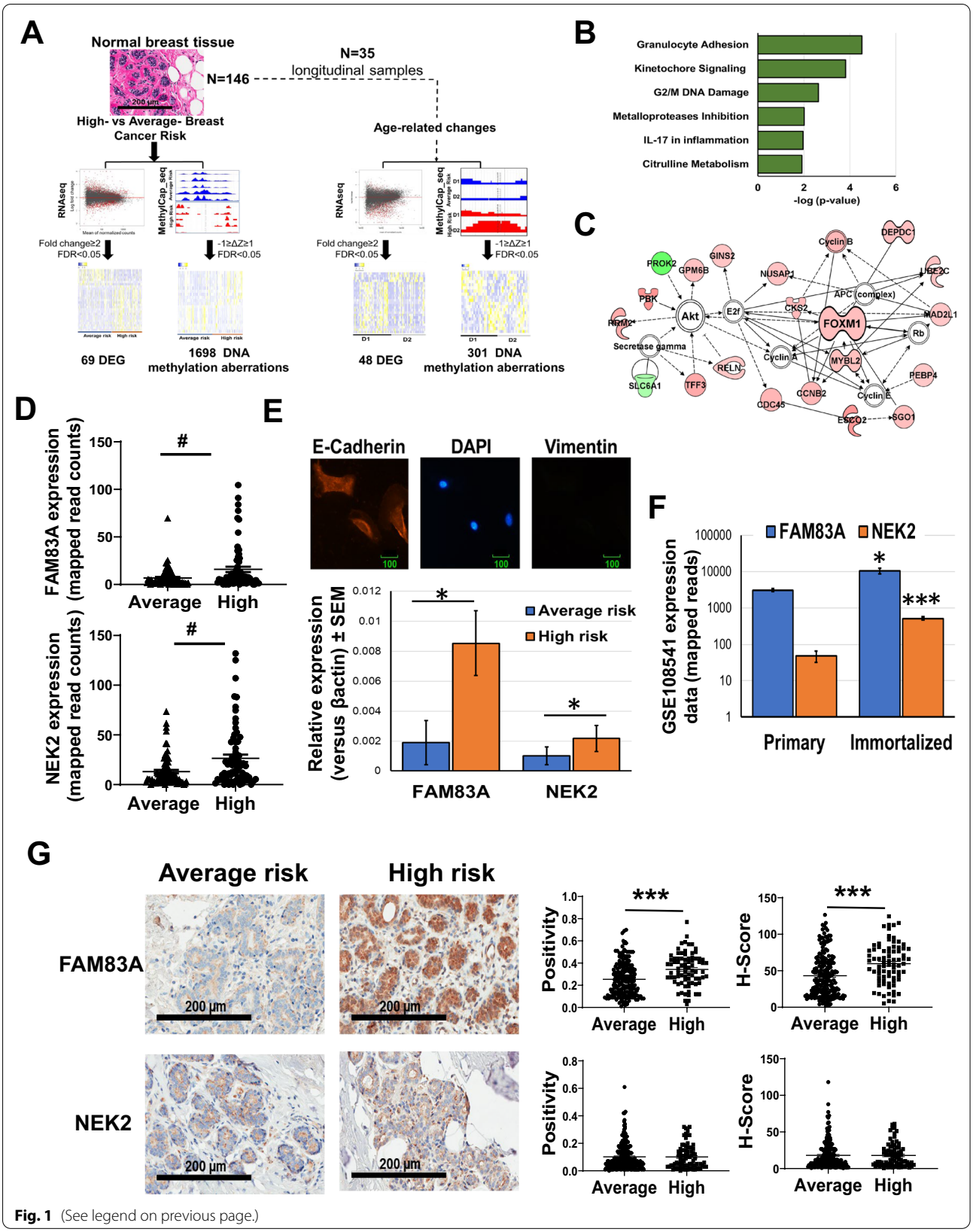

Fig. 1 (See legend on previous page.)

Table 1 Gene expression differences in high- versus average-risk breasts (FC > 2; FDR < 0.05)

Gene name	Description	log2fc[a]	FDR	% genetic alterations[b]	Tumor/ Normal expression (*p* value)[c]	Copy number variation (CNV)[d]		Oncoscore
						CNV = 2 (%), *p* value	CNV = − 2 (%), *p* value	
MEPE	Matrix extracellular phosphoglycoprotein	2.28	2E-02	0.7	0.02 (n.s.)	12 (0.6), *n.s*	2 (0.1), < 0.001	15.6
OPRPN	Opiorphin prepropeptide	2.10	3E-03	1.3	N.A	31 (1.4), *n.s*	0 (0)	N.A
CXCL13	C-X-C motif chemokine ligand 13	2.07	4E-03	1.3	6.6 (0,003)	26 (1.2), *n.s*	0 (0)	33.7
APELA	Apelin receptor early endogenous ligand	1.87	8E-04	0.3	N.A	N.A	0 (0)	N.A
CA6	Carbonic anhydrase 6	1.78	6E-04	0.8	0 (n.s.)	2 (0.1), *n.s*	3 (0.1), < 0.001	14.4
DIO2	Iodothyronine deiodinase 2	1.60	2E-03	0.6	1.94 (n.s.)	13 (0.6), *n.s*	0 (0)	7.7
FEZF2	FEZ family zinc finger 2	1.55	7E-03	0.7	0.04 (< 0.001)	3 (0.1), *n.s*	0 (0)	16.1
TNNT1	Troponin T1%2C slow skeletal type	1.52	9E-03	2.3	51.87 (n.s.)	36 (1.7), *n.s*	0 (0)	12.3
MMP3	Matrix metallopeptidase 3	1.43	2E-02	1.8	5.66 (< 0.001)	26 (1.2), *n.s*	1 (0.04), < 0.001	31.9
SERPINA12	Serpin family A member 12	1.42	2E-02	0.9	1.26 (< 0.001)	12 (0.6), *n.s*	1 (0), < 0.001	11.9
C8B	Complement C8 beta chain	1.42	3E-02	1.8	0.014 (n.s.)	37 (1.7), *n.s*	1 (0), < 0.001	7.3
KCNJ13	Potassium voltage-gated channel subfamily J member 13	1.41	3E-03	0.6	0.16 (0.03)	2 (0.1), *n.s*	1 (0), < 0.001	9.0
CXCL6	C-X-C motif chemokine ligand 6	1.37	5E-03	2.2	0.10 (0.04)	43 (2), *n.s*	0 (0), *n.s*	31.0
SLC12A1	Solute carrier family 12 member 1	1.33	1E-02	0.9	0.48 (< 0.001)	4 (0.2), *n.s*	1 (0), < 0.001	5.6
CYP24A1	Cytochrome P450 family 24 subfamily A member 1	1.33	3E-02	7.0	0.22 (n.s.)	164 (7.5), < 0.001	1 (0), n.s	30.2
ASB5	Ankyrin repeat and SOCS box containing 5	1.29	4E-03	1.3	0.01 (n.s.)	6 (0.3), n.s	5 (0.2), < 0.001	0.0
NPY2R	Neuropeptide Y receptor Y2	1.27	3E-02	1.0	0.003 (< 0.001)	10 (0.5), n.s	0 (0)	7.9
C2CD4A	C2 calcium dependent domain containing 4A	1.26	2E-02	0.6	0.9 (< 0.001)	12 (0.6), n.s	1 (0), < 0.001	11.2
GABRR1	gamma-aminobutyric acid type A receptor rho1 subunit	1.26	3E-02	1.1	1.03 (0.03)	13 (0.6), n.s	5 (0.2), < 0.001	8.7
KIAA1210	KIAA1210	1.25	7E-03	1.6	0.43 (n.s.)	18 (0.8), n.s	3 (0.1), < 0.001	0.0
MMP10	Matrix metallopeptidase 10	1.23	2E-02	1.6	7.07 (< 0.001)	26 (1.2), n.s	1 (0), < 0.001	38.2
FAM83A	Family with sequence similarity 83 member A	1.22	5E-03	16.0	1.23 (< 0.001)	503(23.1), < 0.001	0 (0)	74.5
LPO	Lactoperoxidase	1.21	1E-02	7.0	0.5 (< 0.001)	168 (7.7),2E-24	1 (0), n.s	11.5
CRISP2	Cysteine rich secretory protein 2	1.19	3E-02	1.5	0.06 (0.01)	31 (1.4), 3E-05	0 (0)	8.2
NMU	Neuromedin U	1.19	2E-02	0.8	3.6 (< 0.001)	18 (0.8), n.s	1 (0), < 0.001	41.6
MAGEB4	MAGE family member B4	1.18	9E-03	0.8	8.6 (0.004)	10 (0.5), n.s	2 (0.1), < 0.001	55.9

Table 1 (continued)

Gene name	Description	log2fc[a]	FDR	% genetic alterations[b]	Tumor/ Normal expression (*p* value)[c]	Copy number variation (CNV)[d]		Oncoscore
						CNV = 2 (%), *p* value	CNV = − 2 (%), *p* value	
MAG	Myelin associated glycoprotein	1.17	4E-02	2.3	5.3 (< 0.001)	42 (1.9), 0.001	0 (0)	13.2
DAPL1	Death associated protein like 1	1.17	5E-03	0.7	0.09 (n.s.)	10 (0.5), n.s	0 (0)	14.0
PRSS51	Serine protease 51	1.16	2E-02	1.6	N.A	0 (0)	0 (0)	N.A
PBK	PDZ binding kinase	1.14	4E-03	3.0	15.7 (< 0.001)	20 (0.9), n.s	15(0.7), < 0.001	28.3
KRT77	Keratin 77	1.13	4E-02	0.8	0.04 (n.s.)	12 (0.6), n.s	0 (0)	0.0
CALML3	Calmodulin like 3	1.12	3E-02	4.0	0.15 (n.s.)	108 (5), < 0.001	0 (0)	37.7
ACBD7	Acyl-CoA binding domain containing 7	1.12	3E-03	2.3	1.13 (0.002)	78 (3.6), < 0.001	0 (0)	0.0
UNC5D	Unc-5 netrin receptor D	1.11	2E-02	8.0	0.001 (n.s.)	152 (7), n.s	6 (0.3), < 0.001	44.8
ESCO2	Establishment of sister chromatid cohesion N-acetyltransferase 2	1.11	2E-03	3.0	8.02 (< 0.001)	20 (0.9), n.s	14(0.6), < 0.001	25.1
BARX1	BARX homeobox 1	1.09	4E-02	5.0	1.54 (9E-08)	9 (0.4), n.s	1 (0), < 0.001	22.3
CTXND1	Cortexin domain containing 1	1.09	3E-02	0.0	N.A	0 (0)	0 (0)	N.A
SYT13	Synaptotagmin 13	1.08	4E-03	1.3	4.6 (< 0.001)	36 (1.7), < 0.001	1 (0), n.s	38.8
PRAME	Preferentially expressed antigen in melanoma	1.06	2E-02	1.2	1.8 (< 0.001)	21 (1), n.s	1 (0), < 0.001	82.6
SLC39A12	Solute carrier family 39 member 12	1.05	4E-03	2.4	0.18 (n.s.)	72 (3.3), < 0.001	1 (0), n.s	12.0
IGHV2-26	Immunoglobulin heavy variable2-26	1.04	4E-02	0.1	N.A	0 (0)	0 (0)	N.A
APLN	Apelin	1.04	7E-04	0.6	0.93 (n.s.)	16 (0.7), n.s	2 (0.1), < 0.001	13.8
IGHV3-30	Immunoglobulin heavy variable3-30	1.04	2E-02	0.1	N.A	0 (0)	0 (0)	48.0
LPAR3	Lysophosphatidic acid receptor 3	1.04	8E-03	0.9	0.28 (n.s.)	13 (0.6), n.s	0 (0)	12.9
ECEL1	Endothelin converting enzyme like1	1.03	2E-02	0.8	0.9 (n.s.)	1 (0), n.s	1 (0), < 0.001	N.A
DCX	Doublecortin	1.03	6E-03	0.5	0.1 (0.02)	13 (0.6), n.s	2 (0.1), < 0.001	8.7
NEK2	NIMA related kinase 2	1.02	7E-03	12.0	25.78 (< 0.001)	473 (21.8), < 0.001	0 (0)	61.4
CWH43	Cell wall biogenesis 43 C-terminal homolog	1.02	3E-02	1.0	0.5 (< 0.001)	6 (0.3), n.s	0 (0)	12.9
PRSS21	Serine protease 21	1.01	3E-02	5.0	0.2 (n.s.)	154 (7.1),5E-102	0 (0)	46.3
FOXI3	Forkhead box I3	1.01	2E-02	0.3	0.01 (< 0.001)	10 (0.5), n.s	0 (0)	8.5
FCER2	Fc fragment of IgE receptor II	-0.98	1E-03	1.3	0.07 (0.04)	11 (0.5), n.s	2 (0.1), < 0.001	17.1
DACH2	Dachshund family transcription factor 2	− 1.01	1E-02	0.8	0.3 (n.s.)	17 (0.8), n.s	9 (0.4), < 0.001	25.3
LILRB5	Leukocyte immunoglobulin like receptor B5	− 1.02	8E-04	2.1	0.15 (< 0.001)	39 (1.8), n.s	0 (0)	0.0
SBK3	SH3 domain binding kinase family member 3	− 1.03	7E-03	2.3	N.A	48 (2.2), n.s	0 (0)	0.0
TRDN	Triadin	− 1.03	3E-02	2.3	0.02 (n.s.)	41 (1.9), n.s	1 (0), < 0.001	1.0

Table 1 (continued)

Gene name	Description	log2fc[a]	FDR	% genetic alterations[b]	Tumor/ Normal expression (*p* value)[c]	Copy number variation (CNV)[d]		Oncoscore
						CNV = 2 (%), *p* value	CNV = − 2 (%), *p* value	
NXF3	nuclear RNA export factor 3	− 1.04	3E-03	0.6	0.9 (n.s.)	8 (0.4), n.s	4 (0.2), < 0.001	32.2
LILRA6	leukocyte immunoglobulin like receptor A6	− 1.05	2E-03	2.1	1 (n.s.)	39 (1.8), n.s	1 (0), n.s	0
SYNDIG1L	synapse differentiation inducing 1 like	− 1.07	9E-03	0.5	N.A	8 (0.4), n.s	1 (0), < 0.001	0
ARPP21	cAMP regulated phosphoprotein 21	− 1.13	2E-02	1.1	0.42 (n.s.)	11 (0.5), n.s	1 (0), < 0.001	24.04
SLC22A12	solute carrier family 22 member 12	− 1.13	2E-02	1.1	0.9 (< 0.001)	20 (0.9), n.s	0 (0)	8.9
CCL24	C–C motif chemokine ligand 24	− 1.17	1E-02	0.7	0.98 (< 0.001)	21 (1), n.s	0 (0)	16.2
TPSD1	tryptase delta 1	− 1.17	2E-02	5.0	0.55 (0.04)	170 (7.8), < 0.001	0 (0)	0
PROK2	prokineticin 2	− 1.19	2E-02	0.7	0.24 (0.01)	5 (0.2), n.s	1 (0), 0.001	18.8
HBG2	hemoglobin subunit gamma 2	− 1.59	4E-02	1.0	0.2 (n.s.)	19 (0.9), n.s	0 (0)	11.3
FGF8	fibroblast growth factor 8	− 1.68	3E-04	0.3	0.88 (0.005)	2 (0.1), n.s	1 (0), < 0.001	14.3
SULT1C2	sulfotransferase family1C member2	− 1.74	2E-03	0.5	1.6 (0.02)	9 (0.4), n.s	0 (0)	21.8
MS4A6E	membrane spanning 4-domains A6E	− 2.24	4E-02	0.9	N.A	23 (1.1),n.s	0 (0)	0

N.A., not available; a, log fold change; b, breast cancer data from cBioportal; c: from UALCAN portal; d: data retrieved from the METABRIC, number of samples with either CNV = 2 or − 2 (%), *p* value

Table 2 The 20 most differentially methylated regions between the high- and average-risk breast tissues

Genomic Locus	Overlapping Gene Feature	Gene Name	Description	ΔZ[#]	FDR
Chr8:120,669,501	Intron	SNTB1	syntrophin beta 1	2.4	7E-53
Chr18:6,803,751	Intron	ARHGAP28	Rho GTPase activating protein 28	2.0	3E-35
Chr6:12,944,751	Intron	PHACTR1	phosphatase and actin regulator 1	1.9	1E-31
Chr21:30,743,501	promoter	KRTAP21-4P	keratin associated protein 21–4 2C pseudogene	1.9	6E-31
Chr2:115,663,751	Intron	DPP10	dipeptidyl peptidase like 10	1.8	2E-30
Chr4:87,111,001	Intron	AFF1	AF4/FMR2 family member 1	1.8	2E-29
Chr3:33,638,251	Intron	CLASP2	cytoplasmic linker associated protein 2	1.8	2E-29
Chr14:106,498,501	Intron	LINC01881	long intergenic non-protein coding RNA1881	1.8	2E-29
Chr6:129,416,001	Intron	LAMA2	laminin subunit alpha 2	1.8	4E-29
Chr14:31,750,251	Intron	NUBPL	nucleotide binding protein like	1.8	2E-28
Chr8:37,842,001	Coding	ADGRA2	adhesion G protein-coupled receptor A2	− 1.2	1E-12
Chr1:44,724,501	Coding	C1orf228	chromosome 1 open reading frame 228	− 1.2	9E-13
ChrX:46,575,001	Coding	CHST7	carbohydrate sulfotransferase 7	− 1.2	5E-13
Chr14:104,729,501	Coding	ADSSL1	adenylosuccinate synthase like 1	− 1.3	1E-15
Chr2:202,774,251	Intron/promoter	ICA1L	islet cell autoantigen 1 like	− 1.3	1E-15
Chr1:155,190,001	Coding	MUC1	mucin 1, 2C cell surface associated	− 1.4	2E-17
Chr20:3,751,751	Coding	HSPA12B	heat shock protein family A (Hsp70) member 12B	− 1.4	8E-18
Chr11:58,141,001	Intron	OR9Q1	olfactory receptor family 9 subfamily Q member 1	− 1.6	4E-22
Chr1:45,803,751	Coding	MAST2	microtubule associated serine/threonine kinase 2	− 1.6	3E-23
Chr7:636,001	Intron	PRKAR1B	protein kinase cAMP-dependent type I regulatory subunit beta	− 1.7	1E-24

[#] High- versus average-risk value

are independent of both racial background and menopausal status of the tissue donors (Additional file 1: Table S4 and Additional file 2: Fig. S1).

METABRIC, cBioportal, UALCAN and Oncoscore databases were interrogated to determine the cancer-related relevance of the 69 differentially expressed genes. Among them, FAM83A and NEK2 showed overexpression in BC ($p > 0.001$), high genetic alteration frequency (> 10%), high gene amplification rate, and an Oncoscore > 50, and therefore were selected for further investigation (Table 1 and Additional file 2: Fig. S2) [17–19]. The expression of FAM83A and NEK2 in the breasts of high- and average-risk women is shown in Fig. 1D. We detected a 4.5-fold increase in FAM83A and 2.2-fold increase in NEK2 expression in primary epithelial cells isolated from the breast of high-risk women when compared with cells isolated from breast tissue of average-risk women (Fig. 1E). Overexpression of both targets was detected also in a dataset of hTERT-immortalized epithelial cells as compared with the isogenic primary cells [20] (Fig. 1F). Moreover, immunostaining of a breast tissue microarray showed a 1.4 fold increase in FAM83A protein levels in the breast tissues from women at high risk of BC as compared with the breast tissues from subjects at average risk ($p < 0.0001$, Fig. 1G, Additional file 2: Fig. S3A and Additional file 1: Table S5). FAM83A overexpression in normal breast tissues was associated with parity ($p < 0.001$), tobacco use ($p = 0.01$), and family history of BC ($p = 0.02$) (Additional file 1: Table S6). On the contrary, NEK2 staining showed no difference in protein levels between the two groups (Fig. 1G). No difference in Ki67, estrogen receptor alpha (ERα), FOXA1, and GATA3 staining between high- and average-risk breasts was observed (Additional file 2: Fig. S3B and Additional file 1: Table S5). This data shows that FAM83A expression changes are specific to breasts of women at high risk of developing BC.

Genome-wide DNA methylation analysis reveals 1698 aberrant DNA methylation sites in normal breast tissue of high-risk women

With the goal of identifying alterations in regulatory regions leading to BC susceptibility, we performed a methylome analysis using the MethylCap-seq approach. Differential analysis of the methylated regions detected in the breasts from average-risk women and those from women at high risk of cancer revealed a wide chromosomal distribution of the epigenetic alterations (Fig. 2A). DNA methylation changes with a $\Delta Z \geq 1$ (hypermethylated) or ≤ -1 (hypomethylated) were selected. We identified 1698 regions methylated that differentiate the breast tissue of high-risk women from that of women at average risk (FDR ≤ 5%), mapping to 1115 unique genes (Additional file 1: Table S7). Neither FAM83A or NEK2 genomic loci were found among the regions affected by BC risk-related DNA aberrations, suggesting that an alternative process than DNA methylation may regulate their expression. The twenty most hypermethyated and hypomethylated regions are shown in Fig. 2B and Table 2. Interindividual variability in DNA methylation can be observed within each experimental group. Among the detected DNA methylation changes, 98.9% consisted of hypermethylated loci ($p = 9 \times 10^{-8}$; Fig. 2C). More than 90% of hypermethylated loci localized in regulatory regions including the promoter, untranslated region, and introns, whereas only 41% of hypomethylated loci localized in these regions (Fig. 2C). Hypomethylated regions were found predominantly in the gene body (59%), a phenomenon that has been linked with the activation in cancers of aberrant intragenic promoters that are normally silenced [21, 22].

Pathway analysis revealed the involvement of cell adhesion (aka synaptogenesis, $p = 1.2E\text{-}06$), ErbB ($p = 3.7E\text{-}04$) and protein kinase A ($p = 4.8E\text{-}04$) signaling pathways (Fig. 2D, Additional file 1: Table S8). Notably, one of the molecular networks showed ESR2 as the central molecule (Fig. 2E). Although ESR2 expression decreased in high-risk breasts (fold change = 0.82), the intronic ESR2 hypermethylation showed no inverse correlation with ESR2 expression ($r = -0.03$, $p = 0.4$; Additional file 2: Fig. S4A-C). One of the hypomethylated genes, MUC1 ($\Delta Z = 1.4$, FDR = 1.6E-17) is reported to be aberrantly overexpressed in over 90% of breast tumors [23, 24] (Additional file 2: Fig. S4D). However, no significant difference in MUC1 expression was observed between high- and average-risk breasts (Additional file 2: Fig. S4E). In the analyzed cohort, DNA methylation was not highly affected by either racial background or menopausal status of the tissue donors (FDR > 0.05; Additional file 1: Table S9). Finally, we found overlap between DNA methylation changes in high-risk breasts and breast cancer-related DNA methylation signatures such as those identified by Saghafinia et al.(4%, 25/666, [25]), Chen et al.(6%, 10/174, [26]), de Almeida et al.(9%, 25/285, [27]), and Xu et al. (9%, 37/414, [28]) (Additional file 1: Table S10).

DNA methylation and gene expression changes in high-risk breast show a weak correlation

To identify potential epigenetically regulated genes linked with BC risk, we performed a Pearson's correlation test on paired DNA methylation and gene expression data (Fig. 3). Among the 69 differentially expressed genes in Table 1, the expression level of eight genes was associated with aberrant intronic DNA methylation, including six genes showing a direct correlation (APELA,

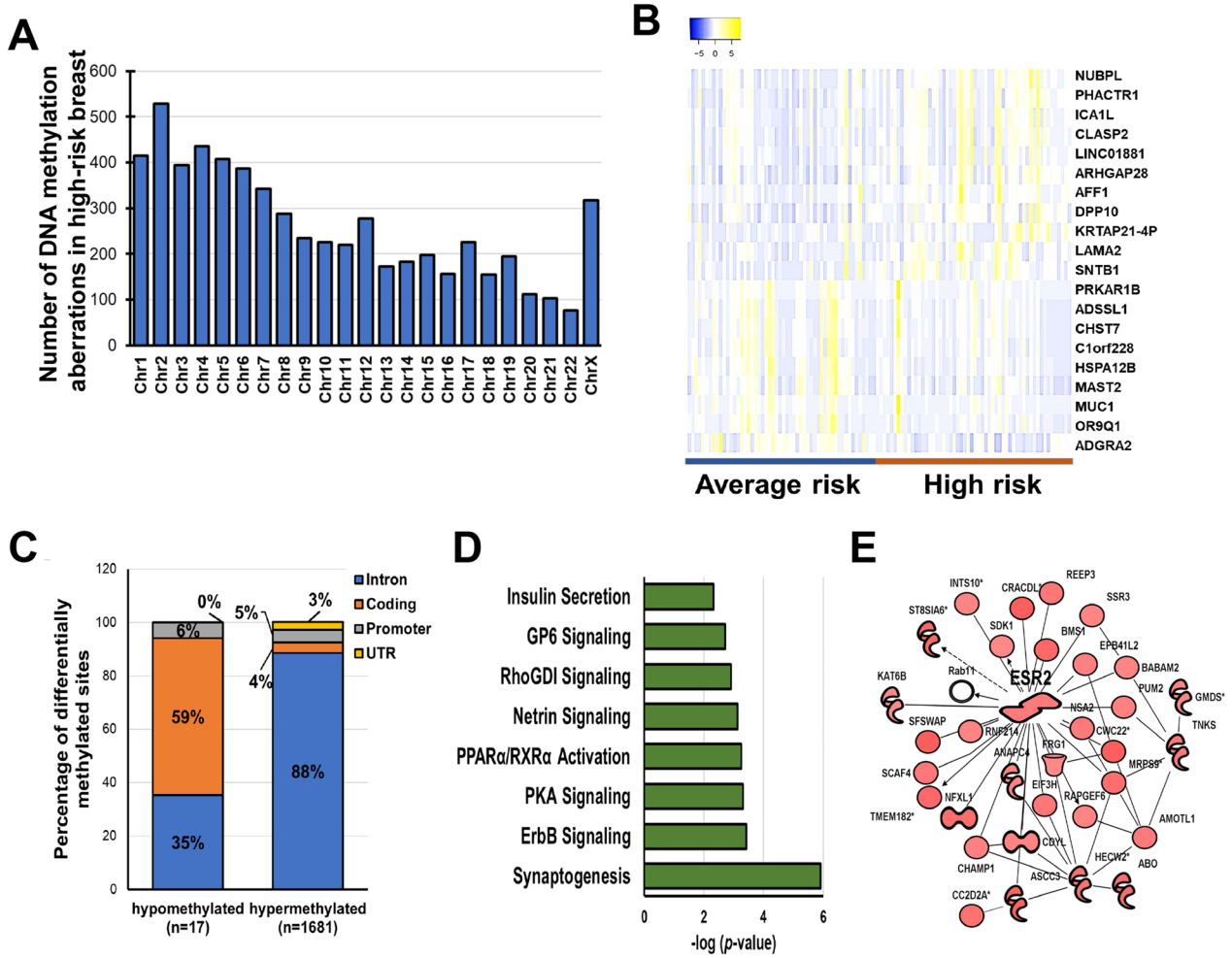

Fig. 2 Methylome profiling of breast tissues from women at either high- or average risk of breast cancer. **A** Chromosomal distribution of the DNA methylation aberrations observed in high-risk versus average-risk group. **B** Heatmap of the 20 highest differentially methylated regions in high-risk breasts as compared with average-risk breasts at FDR < 0.05. The overlapping gene name is indicated on the left. **C** Genomic localization (intron, coding, promoter or UTR) of the DNA methylation aberrations including regions either hypo- or hyper-methylated in high-risk versus average-risk breasts. Data are shown as percentage of each genomic localizaton versus the total number of sites. **D** Pathway analysis of the genes affected by DNA methylation aberrations (FDR < 0.05) in high-risk breasts as compared with breast from women at average risk for breast cancer. **E** One of the molecular networks including the genes affected by DNA hypermethylation

DIO2, FEZF2, LPAR3, UNC5D, and PRSS51) and two genes (PROK2 and SULT1C2) with a negative correlation (Fig. 3A). Furthermore, among the DNA methylation changes in Table 2, only the intronic hypermethylation of PHACTR1 ($\Delta Z = 1.88$, FDR = 1.0E-31) was negatively correlated with PHACTR1 expression (fold change = 0.77, FDR = 0.006, $r = -0.21$) (Fig. 3B). Overall, the correlations identified were weak (r: -0.2,-0.5), suggesting that other regulatory events (chromatin modifications, gene amplification, nucleotide variants), rather than DNA methylation aberrations, may be the determinants of the transcriptomic changes observed in the high-risk breasts as compared with average-risk breasts.

Age-related molecular changes in cancer-free breast tissues in relation with cancer risk

Age is the strongest demographic risk factor for most human malignancies, including BC. Age-related transcriptome and DNA methylation aberrations were investigated on breast tissues cores donated by 35 women at two separate time points (Additional file 1: Table S11). Differential expression analysis (FDR < 0.05) between the two donation time points revealed the dysregulation of 205 genes involved in LXR/RXR activation ($p = 7E-04$), immune response ($p = 2E-03$), and senescence ($p = 7E-03$) (Additional file 1: Tables S12 and S13). Among 25 age-related transcriptomic changes

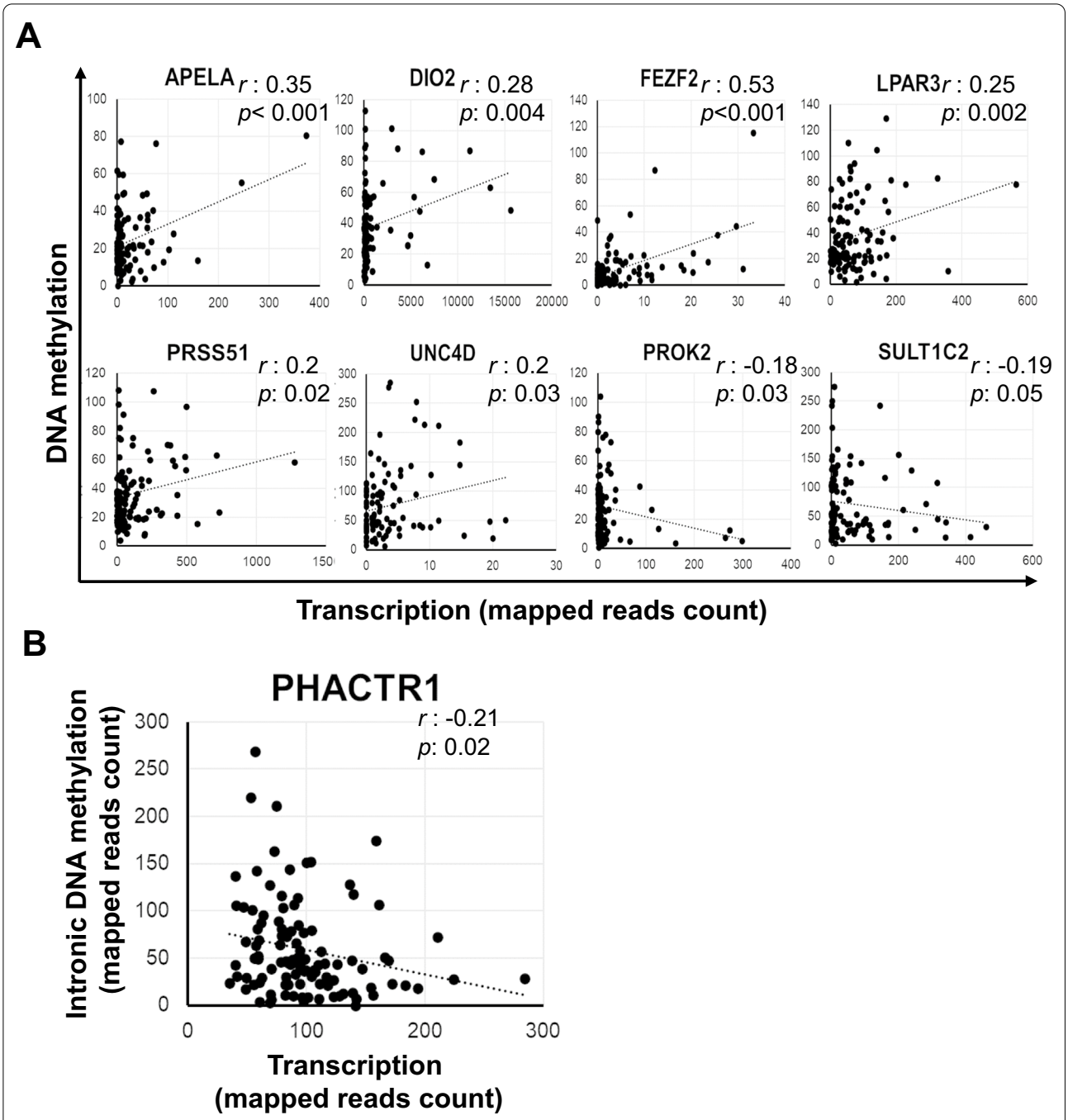

Fig. 3 Correlation between degree of DNA methylation and gene expression. **A** Pearson's correlation analysis between DNA methylation value and expression of the genes found differentially expressed between high- and average-risk breasts. **B** Pearson's correlation analysis of the DNA methylation and expression of PHACTR1, hypermethylated in the breasts of high-risk women. *r* is the correlation coefficient and *p* is *p*value

with a fold change (fc) ≥ 2 and FDR < 0.05 seven genes showed the highest expression level and included two upregulated genes, CETP (fc=2.4;FDR = 0.04) and HP (fc = 2.3; FDR = 0.03), and five downregulated genes, SLC5A1 (fc = 0.4; FDR = 0.03), SLCO1A2 (fc = 0.4; FDR = 0.03), GRIA4 (fc = 0.4; FDR = 0.01), IL22RA2 (fc = 0.4; FDR = 0.01), and CHRM1 (fc = 0.4; FDR = 0.03) (Additional file 2: Fig. S5). Furthermore, age-dependent dysregulation of the following five genes was enhanced in breast tissues from high-risk women:

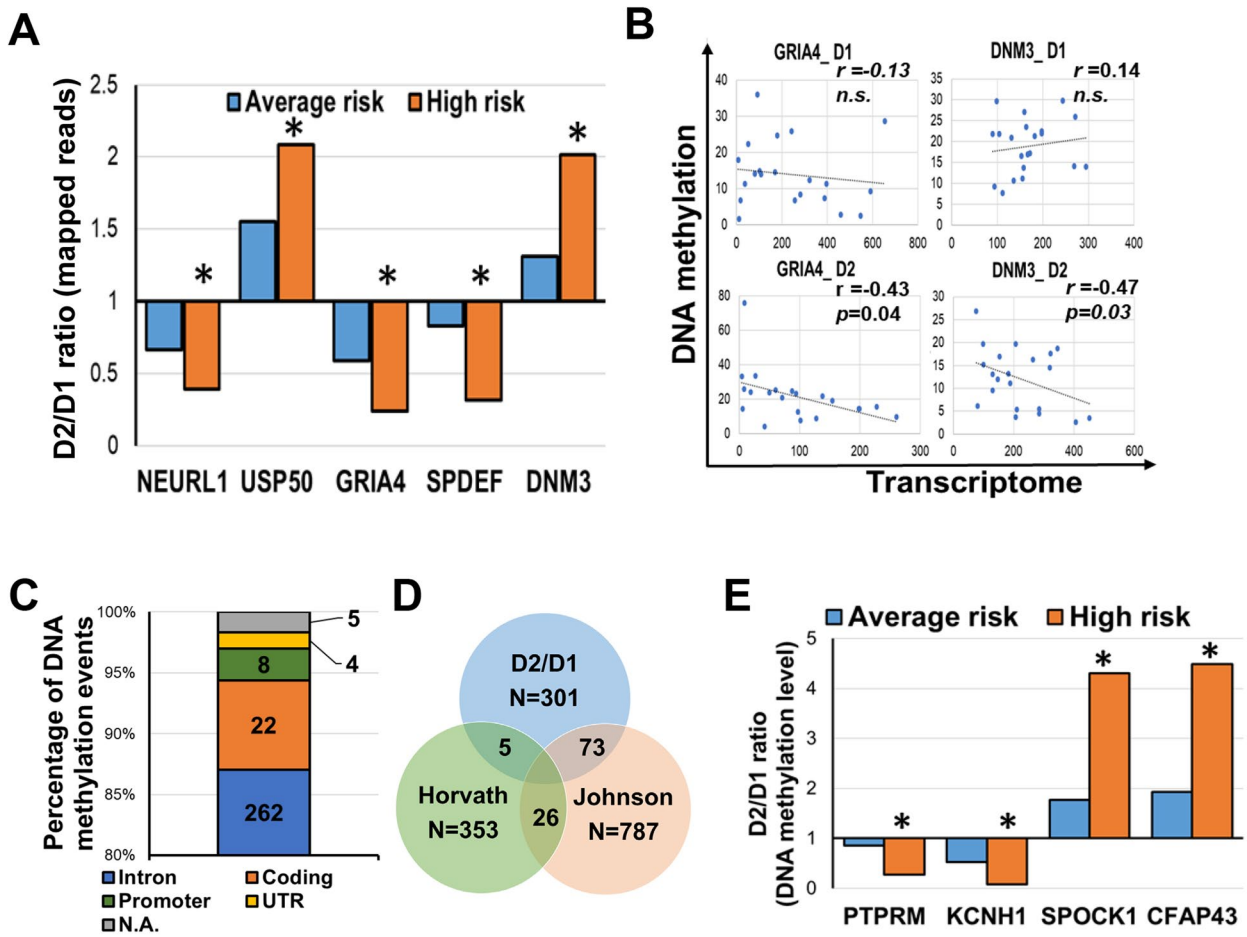

Fig. 4 Age-related transcriptome and DNA methylation changes in healthy breast tissues. **A** Differentially expressed genes between the first (D1) and second (D2) donation time point in the breast tissues from average (blue bars) and high- (orange bars) risk women. Ratio between D2 and D1 is shown. **B** Pearson's correlation test between DNA methylation and transcription of GRIA4 and DNM3 in average- and high-risk breasts at the two time points, D1 and D2, **C** Number of genomic locations (intron, coding regions, promoter, UTR) of the age-related DNA methylation events. N.A.: not available. **D** Venn diagram of the DNA methylation changes associated with age comparing our data set (D2/D1) with Horvath' epigenetic clock (353 CpGs) or Johnson's age-associated loci (787 CpGs) **E** Differentially methylated regions between the first (D1) and second (D2) donation time point in the breast tissues from average (blue bars) and high- (orange bars) risk women. Ratio between D2 and D1 is shown. *$p < 0.05$; **$p < 0.001$

NEURL1, USP50, GRIA4, SPDEF, and DNM3 (Fig. 4A). Notably, the expression of GRIA4 ($r = -0.43$, $p = 0.04$) and DNM3 ($r = -0.47$, $p = 0.03$) showed a negative correlation with their DNA methylation pattern, thus suggesting a potential epigenetic regulation for these two molecules (Fig. 4B). Neither FAM83A or NEK2 were found among the genes affected by age-dependent transcriptomic changes.

Age-dependent DNA methylation aberrations affected 301 loci corresponding to 280 unique transcripts (Additional file 1: Table S14). As previously reported [29], age-related DNA methylation alterations consisted predominantly of hypermethylation events (85.4%) and affected the intronic regions (Fig. 4C). DNA methylation measurements were previously used to develop epigenetic biomarkers of aging, otherwise known as "DNA methylation age" or the "epigenetic clock" [30, 31]. We observed a limited overlap between the 301 DNA methylation aberrations and the epigenetic clocks described by Horvath et al.(1.4%, [31]), whereas 73 genes associated with the differentially methylated bins in our dataset overlapped with age-associated DNA methylation alterations reported by Johnson et al.(24.2%, [29]) (Fig. 4D). Finally, we identified age-related DNA methylation aberrations enhanced in high-risk breasts, localized on

four genes: PTPRM, SPOCK1, KCNH1, and CFAP43. ($p < 0.001$, Fig. 4E and Additional file 1: Table S14).

Discussion

This study aimed to define the distinct features of cancer-free breast tissues from women at high risk for breast cancer (BC) and, thus, identify molecular markers that could potentially screen for women susceptible to cancer. We conducted transcriptome and methylome analyses using breast tissue cores donated by healthy women. The participants were divided into two cohorts based on their risk of developing breast cancer, according to the Tyrer-Cuzick lifetime risk assessment score: high-risk ($\geq 20\%$) and average-risk ($< 20\%$) [32]. Among the genes upregulated in high-risk breast, we identified two promising markers of BC susceptibility, FAM83A and NEK2. Furthermore, when investigating DNA methylation aberrations in high-risk breasts, we detected 4–10% overlap with cancer-related signatures.

Our transciptomic analysis of high- and average-risk breasts revealed significant changes in the expression of 69 genes (FDR < 0.05). Pathway analysis suggested the activation of cell cycle and cell adhesion in the high-risk breasts. Furthermore, one of the molecular networks including the differentially expressed genes revealed the involvement of FOXM1 signaling. FOXM1 itself is upregulated 1.6 fold in high-risk breasts ($p = 0.001$). The transcription factor FOXM1 regulates the transcription of cell-cycle genes essential for transit from the G1/S phase into the G2/M phase, such as cyclin A2, JNK1, ATF2, and CDC25A phosphatase as well as genes critical for chromosome segregation and cytokinesis [33]. FOXM1 is overexpressed and plays a critical role in tumorigenesis, metastasis, and drug resistance in a broad range of human cancer types, such as lung, gastric, and breast cancers [16]. Compounds targeting FOXM1 expression or activity are under investigation [16]. Our results suggest that the transcriptional dysregulation detected in high-risk breasts may be driven by FOXM1.

Two genes, FAM83A and NEK2, both upregulated in high-risk breast, showed a high Oncoscore (75.5 and 61.4, respectively), and have been reported amplified in BC. FAM83A is the smallest member of the eight-member FAM83 family of proteins, that share a conserved amino-terminal Domain of Unknown Function (DUF1669 domain) [34]. It was identified based on its transforming potential [35–37]. FAM83A upregulation has been detected in multiple human tumor types, including breast, lung, pancreatic and cervical cancer [37–44]. Lee et al. [45, 46] revealed the ability of FAM83A to confer resistance to epidermal growth factor receptor-tyrosine kinase inhibitors (EGFR-TKIs) through interactions with c-RAF and PI3K p85 in BC. The authors also showed that BC patients with high FAM83A expression had a worse prognosis. FAM83A depletion suppressed proliferation and invasiveness in vitro as well as tumor growth in vivo [36]. Based on the aforementioned studies, FAM83A is considered a candidate oncogene. Our findings suggest that FAM83A may be one of the first molecules dysregulated in cancer transformation and thus a marker of BC susceptibility. The functional role of FAM83A in BC initiation is currently being investigated by our team. Moreover, our DNA methylation data, in agreement with previous literature, suggest that FAM83A overexpression is mainly driven by genomic amplification rather than epigenetic regulation [47, 48]. Additional studies such as dual color fluorescence in situ hybridization and deep whole genome sequencing of DNA from breast tissues of high-risk women are required to support this hypothesis.

The NIMA-related kinase 2 (NEK2) protein belongs to a family of serine/threonine kinases, which have a role in mitotic progression by initiating the separation of centrosomes [49]. NEK2 overexpression was previously reported in BC as result of gene amplification [47, 50]. NEK2 depletion blocks cell cycle progression and tumor cell growth, making it an ideal therapeutic target [51]. Notably, FOXM1 is reported to both bind NEK2 promoter and interact with NEK2 [52, 53]. Our study further suggests a role of NEK2 dysregulation in breast carcinogenesis. However, we did not observe changes in NEK2 protein levels in breast tissues of high-risk women, suggesting a disconnect between mRNA and protein levels, which is not uncommon, due to a more complex regulatory pathway. Our observations indicate that, while increased NEK2 mRNA expression may be indicative of BC risk, post-transcriptional events may bring NEK2 to its basal protein level. NEK2 may have a more critical functional role in a late phase of BC development. Further investigation of the role of NEK2 in breast carcinogenesis is needed.

We observed DNA methylation changes in high-risk breasts, consisting mostly of hypermethylation (98.8%) in the intronic regions (88%). Previous studies reported aberrant hypermethylation in normal breast tissue adjacent to the tumor [54]. Hypermethylation in specific gene promoters is indeed linked to carcinogenesis through transcriptional silencing of tumor suppressor genes or regulatory regions within the genome, leading to dysregulation of cell growth, cancer initiation and progression [55–57]. We identified a 4–10% overlap between methylome aberrations in high-risk breasts and previously reported cancer-related signatures [25–28]. The limited overlap may be linked to the different technical approaches (Methyl-capture vs Infinium HumanMethylation450 array) but may also suggest that most of

the cancer–related epigenetic marks are newly acquired during cancer initiation rather than being imprinted into the genome. Moreover, neither FAM83A or NEK2 was found among the genes affected by DNA methylation, suggesting that a different regulatory process may control their transcription. Although the expression of epigenetic modifiers such as DNMTs remain unaffected, we detected the upregulation in high-risk breasts of HASPIN (fc = 1.7; FDR < 0.005), a serine/threonine kinase involved into the phosphorylation of the histone H3 during mitosis [58], suggesting that other genetic and epigenetic mechanisms rather than DNA methylation may drive the transcriptomic aberrations detected in high-risk breasts.

Age is the strongest demographic risk factor for most human malignancies, including BC [59]. The limited size of our age-related cohort ($N = 35$) prevented us from subdividing the subjects by age at tissue donation. Nevertheless, we identified age-related transcriptomic aberrations enhanced in high-risk breasts including GRIA4 and DNM3, which resulted as potentially epigenetically regulated. In terms of DNA methylation aberrations, we found a limited overlap between the age-related DNA methylation changes from our cohort and the epigenetic clock from Hovarth et al. [31] (Additional file 1: Table S14). However, a 24.2% overlap of our dataset with age-related DNA methylation aberrations described by Johnson et al. [29] was detected. The limited overlap is probably due to the different platform used for DNA methylation detection (Infinium Human Methylation 450 array vs Methyl-Cap-seq) and the type of analysis (epithelium-specific deconvolution vs whole tissue) [29, 31]. Notably, we identified specific age-related DNA methylation changes, located on PTPRM, KCNH1, SPOCK1, CFAP43 gene region, enhanced in the high- versus average-risk breasts.

This study harbors some limitations: the relatively small sample size prevented us from investigating in details cancer-related variables such as racial background. The selection of normal breast tissue cores with high content in epithelial compartment limited the number of available samples (Additional file 2: Fig. S6). Outcome data for the women at high risk for BC is not available at this time; however, this cohort is under an ongoing annual medical follow up. Because of the faster processing time and smaller cost, we performed whole tissue analysis instead of the more epithelium-specific laser microdissection or single-cell analysis. This limits the compartment specificity of the data but generates a more comprehensive view of the molecular features of the entire breast tissue core. Further deconvolution analysis may overcome this limitation [60, 61].

Conclusions

The present study reveals transcriptomic and epigenetic aberrations linked with BC risk and, thus, provides an avenue for deciphering the functional relevance of genes involved in BC development. We defined a panel of 1698 methylated regions that could be used to predict BC risk. Moreover, among the transcriptional targets here identified, FAM83A showed an increase in both mRNA and protein expression in the breast of women at high BC risk, and therefore may represent a novel tissue biomarker of BC risk.

Methods

Study cohorts

Breast specimens were obtained from the Susan G. Komen Tissue Bank at the IU Simon Comprehensive Cancer Center (KTB) and donated voluntarily upon informed consent by healthy women. Subjects were recruited under a protocol approved by the Indiana University Institutional Review Board (IRB protocols number 1011003097 and 1607623663). Subject demographics and breast cancer (BC) risk factors were collected using a questionnaire administered by the KTB and summarized in Additional file 1: Table S1, S5 and S11. Breast tissue cores are collected by using a needle biopsy as previously described [14]. The study cohort consisted of two groups: 1) For the transcriptome and methylome analyses, 146 women (median age: 39 years) were selected based on the lack of clinical and histological breast abnormalities and high content in breast epithelial compartment (cellularity > 40%). Germline mutation status of the subjects was obtained from KTB. Data were retrieved from the LifeOmic's Precision Health Cloud platform (https://lifeomic.com/products/precision-health-cloud/). Nine established breast cancer–predisposition genes (BRCA1, BRCA2, PALB2, ATM, CHECK2, BARD1, RAD51C, RAD51D, CDH1) were evaluated for variants identified as "pathogenic" or "likely pathogenic" in the ClinVar database (https://preview.ncbi.nlm.nih.gov/clinvar/) (Additional file 1: Table S1) [2, 3].

Thirty-five of these 146 women, including 10 at high risk and 25 at average risk for BC, donated their breast tissue at two time points at intervals from 1–10 years (mean: 3.2) between the tissue donations (Fig. 1A and Additional file 1: Table S11). 2) In a second analysis, paraffin-embedded breast tissue blocks related to 395 healthy women were obtained from the KTB and used to generate tissue microarrays. The cohort included 287 Caucasian, 66 African American, 49 Asian, with age ranging from 18 to 61 (Additional file 1: Table S5).

Breast cancer risk assessment

Lifetime risk of developing BC was estimated by using the Tyrer-Cuzick risk score (IBISv8) [32] and a threshold of 20% to separate high- (≥ 20%) from average-risk (< 20%) individuals. The Tyrer-Cuzick model was selected over the other risk estimation tools for its accuracy and inclusion of young subjects [62].

Tissue processing and nucleic acid extraction

To limit stromal contamination, only breast tissue samples abundant in epithelial compartment (cellularity > 40%) were selected and processed. Total DNA and RNA were isolated from fresh frozen breast tissue biopsies (80-150 mg) using AllPrep DNA/RNA/miRNA kit (Qiagen). Tissues were homogenized by using 2 ml prefilled tubes containing 3 mm zirconium beads (Benchmark Scientific, cat.# D1032-30), 350 µl Lysis Buffer and 2-Mercaptoethanol, and BeadBug 6 homogenizer (Benchmark Scientific) in a cold room at the following conditions: 4000 rpm for 45 s was repeated twice with 90 s rest time. The concentration and quality of total RNA and DNA samples were first assessed using Agilent 2100 Bioanalyzer. A RIN (RNA Integrity Number) and DIN (DNA integrity number) of six or higher is required to pass the quality control.

Whole transcriptome analysis

cDNA library was prepared using the TruSeq Stranded Total RNA Kit (Illumina) and sequenced using Illumina HiSeq4000. Data included 146 paired-end fastq sequence libraries (raw read length: 38 × 2). Reads were adapter trimmed and quality filtered using Trimmomatic ver. 0.38 (http://www.usadellab.org/cms/?page=trimmomatic) setting the cutoff threshold for average base quality score at 20 over a window of 3 bases. Reads shorter than 20 bases post-trimming were excluded. About 94% of the reads have both the mates passing the quality filters. Cleaned reads mapped to Human genome reference sequence GRCh38.p12 with gencode v.28 annotation, using STAR version STAR_2.5.2b [63]. Only samples with about 99% of the cleaned reads aligned to the genome reference. Differential expression analysis was performed using DESeq2 ver. 1.12.3 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html). Counts table containing mapped read counts for each gene, to be used as input for DESeq2 was generated using featureCounts tool of subread package (https://doi.org/10.1093/bioinformatics/btt656). Alternatively, we ran *t*-tests comparing the normalized read counts for the set of replicates from High risk samples to those for the set of replicates from Average risk samples. The normalized read counts were obtained from the DESeq2 run described above. The *p*values from the *t*-test were corrected for multiple testing using Benjamini–Hochberg method.

DNA methylation analysis

Library was generated by using MethylCap Library Kit (Diagenode, Denville NJ, US) according to the manufacturer's protocols followed by single-end 75-bp sequencing on Illumina Nextseq4000. Internal controls and duplicate samples were used to account for any batch effect and technical artifact. The data comprises of 146 paired end read libraries in FASTQ format. These libraries represent replicates for two samples—High risk (68 libraries) and Average risk (78 libraries). The libraries were sequenced across multiple runs and the combined read counts for each library were generated**.** Reads were adapter trimmed and quality filtered using Trimmomatic 0.38 (http://www.usadellab.org/cms/?page=trimmomatic) with the cutoff threshold for average base quality score set at 20 over a window of 3 bases. Reads shorter than 20 bases post-trimming were excluded. Approximately, 96% of the sequenced reads passed the quality filters to be considered "cleaned" reads. This quality control reduced the number of samples to 57 high- and 55 average-risk. Cleaned reads were mapped to Human genome reference GRCh38.p12 using BWA ver. 0.7.15 [64]. Insert sequences were imputed from the concordantly mapped read pair alignments. More than 95% of the cleaned read pairs were concordantly mapped. A previously described differential methylation analysis using either Zratio or ΔZ [65, 66] was applied to the current methylcapture dataset with a slight improvisation on the validation of the significance of differential methylation. For any given local bin of a given width on the genome, the method compares across samples, variation in deduplicated insert coverage distribution quantified as the bin's z-score with respect to a larger genome region containing the bin. For this analysis, we used local non-overlapping bins with a fixed width of 250 bp with their z-scores computed relative to 25 KB regions. Z-score is the number of standard deviations by which the bin coverage varies from the larger region's mean coverage. A significant difference in Z-scores, calculated as either as ΔZ or Zratio between the samples would indicate potential differential methylation for that bin, as previously described [67]. The analysis identified 159,438 bins, each 250 bp wide, to be potentially differentially methylated between High risk and Average risk samples with z-ratios or ΔZ significant at 5% FDR and *p*-values from t-test ≤ 0.05. Based on positional overlap, these bins were annotated using annotation from gencode v28.

Data analysis

Ingenuity Pathways Analysis (IPA, Qiagen, Redwood City, CA) was used for canonical pathway and molecular network analyses [68]. Publicly available transcriptomic data from primary and immortalized breast epithelial cells were obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108541) [20]. Analysis of The Cancer Genome Atlas (TCGA) was performed by interrogating both cBioPortal (https://www.cbioportal.org/) and UALCAN (http://ualcan.path.uab.edu/) databases [69]. Copy number variations (CNV) analysis was obtained from the interrogation of the Molecular Taxonomy of Breast Cancer International Consortium, METABRIC [17, 18]. Oncoscore was used to rank genes according to their association with cancer, based on the available scientific literature (http://www.galseq.com/next-generation-sequencing/oncoscore-software/; accessed on 3/31/2021) [19].

Primary breast epithelial cells and immunofluorescence

Primary breast epithelial cells were generated from cryopreserved breast tissue cores obtained from the KTB as previously described [14, 20]. Immunofluorescence staining was performed as previously described [14]. Briefly, 5000 cells were cultures overnight into each well of an 8 well-chamber slide (BD Biosciences, San Jose, CA) and fixed with acetone: methanol (1:1) at -20 °C for 10 min. After washing and blocking (PBS1X, 5% normal goat serum, 0.1%TritonX-100) steps, cells were incubated with primary either rabbit anti-vimentin (Cell Signaling, D21H3, 1:100) or mouse anti-E-cadherin antibody (Cell Signaling, 14472, 1:50) overnight. Upon three washes with PBS, cells were incubated with secondary antibodies (goat anti- mouse Alexa Fluor 568 or goat anti-rabbit Alexa Fluor 488; Thermo Fisher Scientific, 1:500) for 1 h at room temperature. After three washes with PBS, the coverslide was mounted using DAKO fluorescent mounting medium (S3023 Agilent, Santa Clara, CA) and the staining was visualized using a fluorescent microscope (Eclipse TS100, Nikon Instruments inc, Melville, NY).

Quantitative real time polymerase chain reaction (qPCR)

Total RNA was extracted from cells using AllPrep DNA/RNA/miRNA kit (Qiagen). Reverse transcription was performed using SuperScript™ IV VILO™ Master Mix (Invitrogen cat#: 11756050) according to the manufacturer's instructions. qPCR was performed using the TaqMan™ Universal PCR Master Mix (Applied Biosystems, cat# 4304437) and the following TaqMan Gene Expression Assays (Applied Biosystems/Thermo Fisher Scientific, Grand Island, NY): ACTB (Hs99999903_m1), FAM83A (Hs04994801_m1), and NEK2 (Hs00601227_m1). qPCR reactions were run on a StepOne Plus Real-Time PCR System (Applied Biosystems/Thermo Fisher Scientific), and data analyzed using the StepOne Software v2.3 (Applied Biosystems). Relative quantification was calculated with reference to ACTB and analyzed using the comparative C_T method. qPCR experiments were performed in triplicate.

Tissue microarray (TMA) immunohistochemistry (IHC) analysis

Normal breast tissues microarrays from 683 women were generated from paraffin-embedded blocks obtained from the KTB at the Tissue procurement & Distribution core of the IU Simon Comprehensive Cancer Center. Due to loss of material during TMA construction and processing, only 58% ($n = 395$) of these tissue biopsies were interpretable. TMA was analyzed with the following antibodies FAM83A (Protein Tech 20618-1-AP, 1:100), NEK2 (MyBioSource MBS9607934, 1:100), Ki67 (DAKO IR 626, ready-to-use), estrogen receptor alpha (ERα) (clone:EP1, DAKO IR 084, ready-to-use), FOXA1 (Santa Cruz Biotechnology sc-6553, 1:100), and GATA3 (Santa Cruz Biotechnology sc-268, 1:50) [70]. IHC was performed in a Clinical Laboratory Improvement Amendments (CLIA)-certified histopathology laboratory and evaluated by 3 pathologists in a blinded manner. Quantitative measurements generating positivity and H-score were obtained via the automated Aperio Imaging system using an FDA-approved algorithm [71].

Statistical analysis

Comparisons between groups were done using either Student's *t*-test or nonparametric Mann–Whitney test on GraphPad Prism 9. Difference between groups is considered significant at p-values < 0.05. Pearson's correlation analysis was performed to determine the strength and direction of the linear relationship between DNA methylation and transcription for given targets. Only correlations with a $p < 0.05$ are shown. For transcriptome and methylome data, differential analysis was performed using DESeq2 and the previously described Z-score method [65, 66], respectively. P-values < 0.05 are considered significanct and are corrected for multiple testing using the Benjamini–Hochberg False Discovery Rate (FDR) algorithm. For the tissue microarrays analysis nonparametric Wilcoxon rank-sum tests were used for unpaired analyses, as positivity and H-scores were not normally distributed, whereas nonparametric Wilcoxon signed-rank tests were used for paired analyses. The statistical software SAS version 9.4 (SAS Institute Inc., Cary, NC) was used to complete the statistical analyses with $p < 0.05$ considered significant. Baseline

demographic characteristics were summarized as median (range) for continuous variables and number and percentage for categorical variables. Comparisons between groups were done using Chi-square tests (or Fisher's Exact test, where appropriate) for categorical variables, or Wilcoxon test for continuous variables.

Abbreviations

BC: Breast cancer; KTB: Susan G. Komen Tissue bank at IU Simon Comprehensive cancer center; IHC: Immunohistochemistry; IPA: Ingenuity pathway analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-022-01239-1.

Additional file 1. It includes subjects demographics and raw data in form of tables.

Additional file 2. It includes additional data related to the main findings shown in the main figures.

Acknowledgements

Samples from the Susan G. Komen Tissue Bank at the IU Simon Comprehensive Cancer Center were used in this study. We thank the tissue donors, whose help and participation made this work possible. We thank the IU Simon Comprehensive Cancer Center for the use of the Tissue Procurement & Distribution Core, which provided the generation of the tissue microarray. We thank the Immunohistochemistry Core for the immunostaining and analysis of the tissue microarrays and the Center for Genomics and Bioinformatics Core at IU Bloomington for the –omics analyses.

Authors' contributions

N.M. conceived the idea and designed the experiments, analyzed, and interpreted the data, and was major contributor in writing the manuscript. N.M. and R.G. performed the experiments and generated the data. A.V. assisted with tissue processing. R.P, D.R., J.H., J.W., G.S., and S.A. analyzed the data. C.T. performed the immunostaining of the tissue microarray. P.R. provided the specimens used in this study. K.N. and H.N. provided input on the manuscript. A.M.V.S. participated in hypothesis generation, funded the work, and contributed to the manuscript preparation. All the authors read and approved the manuscript.

Funding

The work including sample processing, data collection and analysis was funded by the Breast Cancer Research Foundation (BCRF-18–155 and BCRF-19–155 to A.M.S) and the Hero Foundation. The Susan G. Komen Tissue Bank at the IU Simon Comprehensive Cancer Center (IUSCCC) is supported by the IUSCCC, Susan G. Komen, and the Vera Bradley Foundation for Breast Cancer Research.

Availability of data and materials

Methylome and Transcriptome data are available in Gene Expression Omnibus (GEO) with GSE164694 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164694) which includes the Sub Series GSE164640 (MeCap dataset) and GSE164641 (RNA-seq dataset). Transcriptomic data of primary and immortalized breast epithelial cells from Dr. Nakshatri's team [20] were obtained from GEO with accession number GSE108541 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108541).

Declarations

Ethics approval and consent to participate

Breast specimens were obtained from the Susan G. Komen Tissue Bank at the IU Simon Comprehensive Cancer Center (KTB) and donated upon informed consent by healthy women volunteers. Subjects were recruited under a protocol approved by the Indiana University Institutional Review Board (IRB protocols number 1011003097 and 1607623663).

Consent for publication

Not applicable.

Competing interests

The authors have no conflicts of interest to disclose.

Author details

[1]Susan G. Komen Tissue Bank at the IU Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA. [2]Department of Medicine, Hematology/Oncology Division, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [3]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA. [4]Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [5]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [6]Department of Anatomy, Cell Biology, & Physiology, Indiana University, Bloomington, IN 47405, USA. [7]Department of Surgery, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

Received: 10 September 2021 Accepted: 25 January 2022
Published online: 09 February 2022

References

1. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4.
2. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017;49(12):1767–78.
3. Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, et al. A population-based study of genes previously implicated in breast cancer. N Engl J Med. 2021.
4. Breast Cancer Association C, Dorling L, Carvalho S, Allen J, Gonzalez-Neira A, Luccarini C, et al. Breast Cancer Risk Genes—Association Analysis in More than 113,000 Women. N Engl J Med. 2021;384(5):428–39.
5. Feng H, Gusev A, Pasaniuc B, Wu L, Long J, Abu-Full Z, et al. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. Genet Epidemiol. 2020;44(5):442–68.
6. Ferreira MA, Gamazon ER, Al-Ejeh F, Aittomaki K, Andrulis IL, Anton-Culver H, et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat Commun. 2019;10(1):1741.
7. Gao G, Pierce BL, Olopade OI, Im HK, Huo D. Trans-ethnic predicted expression genome-wide association analysis identifies a gene for estrogen receptor-negative breast cancer. PLoS Genet. 2017;13(9):e1006727.
8. Hoffman JD, Graff RE, Emami NC, Tai CG, Passarelli MN, Hu D, et al. Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet. 2017;13(3):e1006690.
9. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet. 2018;50(7):968–78.
10. Kothari C, Ouellette G, Labrie Y, Jacob S, Diorio C, Durocher F. Identification of a gene signature for different stages of breast cancer development that could be used for early diagnosis and specific therapy. Oncotarget. 2018;9(100):37407–20.
11. Cantone I, Fisher AG. Epigenetic programming and reprogramming during development. Nat Struct Mol Biol. 2013;20(3):282–9.
12. Hansmann T, Pliushch G, Leubner M, Kroll P, Endt D, Gehrig A, et al. Constitutive promoter methylation of BRCA1 and RAD51C in patients with familial ovarian cancer and early-onset sporadic breast cancer. Hum Mol Genet. 2012;21(21):4669–79.
13. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683–92.
14. Marino N, German R, Rao X, Simpson E, Liu S, Wan J, et al. Upregulation of lipid metabolism genes in the breast prior to cancer diagnosis. NPJ Breast Cancer. 2020;6:50.

15. Latimer JJ, Johnson JM, Kelly CM, Miles TD, Beaudry-Rodgers KA, Lalanne NA, et al. Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer. Proc Natl Acad Sci USA. 2010;107(50):21725–30.
16. Ziegler Y, Laws MJ, Sanabria Guillen V, Kim SH, Dey P, Smith BP, et al. Suppression of FOXM1 activities and breast cancer growth in vitro and in vivo by a new class of compounds. NPJ Breast Cancer. 2019;5:45.
17. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
18. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
19. Piazza R, Ramazzotti D, Spinelli R, Pirola A, De Sano L, Ferrari P, et al. OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. Sci Rep. 2017;7:46290.
20. Kumar B, Prasad M, Bhat-Nakshatri P, Anjanappa M, Kalra M, Marino N, et al. Normal breast-derived epithelial cells with luminal and intrinsic subtype-enriched gene expression document interindividual differences in their differentiation cascade. Cancer Res. 2018;78(17):5107–23.
21. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017;543(7643):72–7.
22. Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012;44(11):1236–42.
23. Kufe DW. MUC1-C oncoprotein as a target in breast cancer: activation of signaling pathways and therapeutic approaches. Oncogene. 2013;32(9):1073–81.
24. Jing X, Liang H, Hao C, Yang X, Cui X. Overexpression of MUC1 predicts poor prognosis in patients with breast cancer. Oncol Rep. 2019;41(2):801–10.
25. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G. Pan-cancer landscape of aberrant DNA methylation across human tumors. Cell Rep. 2018;25(4):1066–80 e8.
26. Chen X, Zhang J, Dai X. DNA methylation profiles capturing breast cancer heterogeneity. BMC Genomics. 2019;20(1):823.
27. de Almeida BP, Apolonio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer. 2019;19(1):219.
28. Xu Z, Sandler DP, Taylor JA. Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the sister study. J Natl Cancer Inst. 2020;112(1):87–94.
29. Johnson KC, Houseman EA, King JE, Christensen BC. Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. Breast Cancer Res. 2017;19(1):81.
30. Sehl ME, Henry JE, Storniolo AM, Ganz PA, Horvath S. DNA methylation age is elevated in breast tissue of healthy women. Breast Cancer Res Treat. 2017;164(1):209–19.
31. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
32. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004;23(7):1111–30.
33. Branigan TB, Kozono D, Schade AE, Deraska P, Rivas HG, Sambel L, et al. MMB-FOXM1-driven premature mitosis is required for CHK1 inhibitor sensitivity. Cell Rep. 2021;34(9):108808.
34. Fulcher LJ, Bozatzi P, Tachie-Menson T, Wu KZL, Cummins TD, Bufton JC, et al. The DUF1669 domain of FAM83 family proteins anchor casein kinase 1 isoforms. Sci Signal. 2018;11:531.
35. Cipriano R, Graham J, Miskimen KL, Bryson BL, Bruntz RC, Scott SA, et al. FAM83B mediates EGFR- and RAS-driven oncogenic transformation. J Clin Invest. 2012;122(9):3197–210.
36. Cipriano R, Miskimen KL, Bryson BL, Foy CR, Bartel CA, Jackson MW. Conserved oncogenic behavior of the FAM83 family regulates MAPK signaling in human cancer. Mol Cancer Res. 2014;12(8):1156–65.
37. Parameswaran N, Bartel CA, Hernandez-Sanchez W, Miskimen KL, Smigiel JM, Khalil AM, et al. A FAM83A positive feed-back loop drives survival and tumorigenicity of pancreatic ductal adenocarcinomas. Sci Rep. 2019;9(1):13396.
38. Bartel CA, Jackson MW. HER2-positive breast cancer cells expressing elevated FAM83A are sensitive to FAM83A loss. PLoS ONE. 2017;12(5):e0176778.
39. Richtmann S, Wilkens D, Warth A, Lasitschka F, Winter H, Christopoulos P, et al. FAM83A and FAM83B as prognostic biomarkers and potential new therapeutic targets in NSCLC. Cancers (Basel). 2019;11(5).
40. Zheng YW, Li ZH, Lei L, Liu CC, Wang Z, Fei LR, et al. FAM83A promotes lung cancer progression by regulating the wnt and hippo signaling pathways and indicates poor prognosis. Front Oncol. 2020;10:180.
41. Snijders AM, Lee SY, Hang B, Hao W, Bissell MJ, Mao JH. FAM83 family oncogenes are broadly involved in human cancers: an integrative multi-omics approach. Mol Oncol. 2017;11(2):167–79.
42. Zhang J, Sun G, Mei X. Elevated FAM83A expression predicts poorer clincal outcome in lung adenocarcinoma. Cancer Biomark. 2019;26(3):367–73.
43. Rong L, Li H, Li Z, Ouyang J, Ma Y, Song F, et al. FAM83A as a potential biological marker is regulated by miR-206 to promote cervical cancer progression through PI3K/AKT/mTOR pathway. Front Med (Lausanne). 2020;7:608441.
44. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 2005;7(6):R953–64.
45. Lee SY, Meier R, Furuta S, Lenburg ME, Kenny PA, Xu R, et al. FAM83A confers EGFR-TKI resistance in breast cancer cells and in mice. J Clin Invest. 2012;122(9):3211–20.
46. Grant S. FAM83A and FAM83B: candidate oncogenes and TKI resistance mediators. J Clin Invest. 2012;122(9):3048–51.
47. Hayward DG, Fry AM. Nek2 kinase in chromosome instability and cancer. Cancer Lett. 2006;237(2):155–66.
48. Chen S, Huang J, Liu Z, Liang Q, Zhang N, Jin Y. FAM83A is amplified and promotes cancer stem cell-like traits and chemoresistance in pancreatic cancer. Oncogenesis. 2017;6(3):e300.
49. Fry AM, O'Regan L, Sabir SR, Bayliss R. Cell cycle regulation by the NEK family of protein kinases. J Cell Sci. 2012;125(Pt 19):4423–33.
50. Hayward DG, Clarke RB, Faragher AJ, Pillai MR, Hagan IM, Fry AM. The centrosomal kinase Nek2 displays elevated levels of protein expression in human breast cancer. Cancer Res. 2004;64(20):7370–6.
51. Kokuryo T, Yokoyama Y, Yamaguchi J, Tsunoda N, Ebata T, Nagino M. NEK2 Is an effective target for cancer therapy with potential to induce regression of multiple human malignancies. Anticancer Res. 2019;39(5):2251–8.
52. Fang Y, Zhang X. Targeting NEK2 as a promising therapeutic approach for cancer treatment. Cell Cycle. 2016;15(7):895–907.
53. Gormally MV, Dexheimer TS, Marsico G, Sanders DA, Lowe C, Matak-Vinkovic D, et al. Suppression of the FOXM1 transcriptional programme via novel small molecule inhibition. Nat Commun. 2014;5:5165.
54. Cho YH, Yazici H, Wu HC, Terry MB, Gonzalez K, Qu M, et al. Aberrant promoter hypermethylation and genomic hypomethylation in tumor, adjacent normal tissues and blood from breast cancer patients. Anticancer Res. 2010;30(7):2489–96.
55. Esteller M, Silva JM, Dominguez G, Bonilla F, Matias-Guiu X, Lerma E, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. J Natl Cancer Inst. 2000;92(7):564–9.
56. Flanagan JM, Munoz-Alegre M, Henderson S, Tang T, Sun P, Johnson N, et al. Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients. Hum Mol Genet. 2009;18(7):1332–42.
57. Potapova A, Hoffman AM, Godwin AK, Al-Saleem T, Cairns P. Promoter hypermethylation of the PALB2 susceptibility gene in inherited and sporadic breast and ovarian cancer. Cancer Res. 2008;68(4):998–1002.
58. Eswaran J, Patnaik D, Filippakopoulos P, Wang F, Stein RL, Murray JW, et al. Structure and functional characterization of the atypical human kinase haspin. Proc Natl Acad Sci U S A. 2009;106(48):20198–203.
59. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7–30.
60. Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. Genome Biol. 2020;21(1):221.
61. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82.
62. McClintock AH, Golob AL, Laya MB. Breast cancer risk assessment: a stepwise approach for primary care providers on the front lines of shared decision Making. Mayo Clin Proc. 2020;95(6):1268–75.

63. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
65. Maiuri AR, Peng M, Podicheti R, Sriramkumar S, Kamplain CM, Rusch DB, et al. Mismatch repair proteins initiate epigenetic alterations during inflammation-driven tumorigenesis. Cancer Res. 2017;77(13):3467–78.
66. Maiuri AR, Savant SS, Podicheti R, Rusch DB, O'Hagan HM. DNA methyltransferase inhibition reduces inflammation-induced colon tumorigenesis. Epigenetics. 2019;14(12):1209–23.
67. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagn. 2003;5(2):73–81.
68. Kramer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30(4):523–30.
69. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017;19(8):649–58.
70. Nakshatri H, Kumar B, Burney HN, Cox ML, Jacobsen M, Sandusky GE, et al. Genetic ancestry-dependent differences in breast cancer-induced field defects in the tumor-adjacent normal breast. Clin Cancer Res. 2019;25(9):2848–59.
71. Sandusky GE, Mintze KS, Pratt SE, Dantzig AH. Expression of multidrug resistance-associated protein 2 (MRP2) in normal human tissues and carcinomas using tissue microarrays. Histopathology. 2002;41(1):65–74.

Publisher's Note