SCIENTIFIC REPORTS

OPEN

Study of Gene Expression Profiles of Breast Cancers in Indian Women

Shreshtha Malvia[1], Sarangadhara Appala Raju Bagadi[1], Dibyabhaba Pradhan[2], Chintamani Chintamani[3], Amar Bhatnagar[4], Deepshikha Arora[5], Ramesh Sarin[6] & Sunita Saxena[1]

Received: 19 September 2017
Accepted: 25 June 2019
Published online: 10 July 2019

Breast cancer is the most common cancer among women globally. In India, the incidence of breast cancer has increased significantly during the last two decades with a higher proportion of the disease at a young age compared to the west. To understand the molecular processes underlying breast cancer in Indian women, we analysed gene expression profiles of 29 tumours and 9 controls using microarray. In the present study, we obtained 2413 differentially expressed genes, consisting of overexpressed genes such as *COL10A1*, *COL11A1*, *MMP1*, *MMP13*, *MMP11*, *GJB2*, and *CST1* and underexpressed genes such as *PLIN1*, *FABP4*, *LIPE*, *AQP7*, *LEP*, *ADH1A*, *ADH1B*, and *CIDEC*. The deregulated pathways include cell cycle, focal adhesion and metastasis, DNA replication, PPAR signaling, and lipid metabolism. Using PAM50 classifier, we demonstrated the existence of molecular subtypes in Indian women. In addition, qPCR validation of expression of metalloproteinase genes, *MMP1*, *MMP3*, *MMP11*, *MMP13*, *MMP14*, *ADAMTS1*, and *ADAMTS5* showed concordance with that of the microarray data; wherein we found a significant association of *ADAMTS5* down-regulation with older age (≥55 years) of patients. Together, this study reports gene expression profiles of breast tumours from the Indian subcontinent, throwing light on the pathways and genes associated with the breast tumourigenesis in Indian women.

Breast cancer is the most common cancer among women worldwide, representing nearly a quarter (25%) of all cancers with an estimated 2.1 million new cancer cases diagnosed in 2018[1]. Over the past two decades, there has been a rapid increase in breast cancer incidence throughout Asia, mainly South-Eastern Asia, including India[2–6]. Breast cancer is the most common cancer among Indian women in a majority of urban cancer registries at Delhi, Mumbai, Bangalore, Thiruvananthapuram (AAR ranges between 33-41/100000 women) and has rapidly overtaken cervical cancer[7].

In India, although age-adjusted incidence rate of breast cancer is lower (25.8 per 100 000) than the United States of America (93 per 100 000), age-wise distribution of incidence shows a higher percentage (46.7%) of breast cancer incidence among women below the age of 50 years compared to United States of America (19%)[8]. An incidence rate of 45.5% has been observed in Asian countries for this age group, suggesting a higher incidence of breast cancer in the younger age group in India and other Asian countries as compared to the western population[8]. The underlying causes may be attributed to demographic, genetic, and environmental factors alone or in combination, which may be contributing to the development of the disease at a younger age[9,10]. To our knowledge, there is a single report describing gene expression profiles of breast cancer from Indian patients, focusing mainly on estrogen receptor (ER) positive and ER-negative tumours profiles alone[11]. In the present study, we have analysed the gene signatures and molecular pathways involved in breast carcinogenesis in Indian women by transcriptome profiling.

Materials and Methods

Patients and tissue specimen. A total of ninety-seven (97), histologically confirmed breast cancer patients admitted at Safdarjung Hospital or Indraprastha Apollo Hospital, New Delhi, India, during 2008–2012, were enrolled for this study. The study was approved by institutional ethical committees of both Safdarjang Hospital and Indraprastha Apollo Hospital, New Delhi, and informed consent was taken from all the patients. All the

[1]Tumour Biology Division, ICMR-National Institute of Pathology, New Delhi, 110029, India. [2]Bioinformatics Cell, ICMR-National Institute of Pathology, New Delhi, 110029, India. [3]Department of Surgery, Safdarjung Hospital, New Delhi, 110029, India. [4]Department of Cancer Surgery, Safdarjung Hospital, New Delhi, 110029, India. [5]Department of Pathology, Indraprastha Apollo Hospital, New Delhi, 110076, India. [6]Department of Surgery, Indraprastha Apollo Hospital, New Delhi, 110076, India. Shreshtha Malvia and Sarangadhara Appala Raju Bagadi contributed equally. Correspondence and requests for materials should be addressed to S.A.R.B. (email: bsaraju@gmail.com) or S.S. (email: sunita_saxena@yahoo.com)

experiments were performed following relevant guidelines and regulations. The age of patients, ranged between 25–75 years, comprising of 41 premenopausal and 56 postmenopausal women. The patients were staged according to the American Joint Committee on Cancer (AJCC) guidelines. Expression of ER, progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2/neu) receptors had been determined by immunohistochemistry (IHC) as described elsewhere[12]. The tumour samples were stratified into, luminal –ER and or PR positive, and HER2/neu negative or ER positive with any PR and HER2/neu positive; basal-ER, PR, HER2/neu negative; and HER2/neu overexpressing tumours-ER, PR negative and HER2/neu positive. Of these enrolled patients, tumour tissue from 77 cases and 38 distant normal breast tissues were used for gene expression profiling and for validation by quantitative reverse transcription PCR (qPCR) (Supplementary Tables S1 and S2). The remaining cases were excluded from the study since they had either received prior therapy or had a history of other malignancies besides breast cancer or had poor RNA quality. All tissue samples were snap frozen in liquid nitrogen immediately after the modified radical mastectomy or after incision/trucut biopsy for RNA isolations and stored in RNA Later (Ambion, Austin, TX) at −80 °C.

Total RNA extraction. Total RNA was isolated using 'TRIzol' reagent (Thermofisher Scientific) following manufacturer's protocol. In brief, 50–100 mg of tissue samples were pulverized in liquid nitrogen and the powder obtained was lysed using 1 ml 'TRIzol', followed by 0.2 ml of chloroform, then, aqueous phase consisting of RNA was separated by centrifugation at 12,000 × g for 15 minutes at 4 °C. RNA was precipitated using an equal volume of isopropanol followed by centrifugation at 12,000 × g for 15 minutes at 4 °C. The RNA pellet was washed with 75% ethanol, air dried and resuspended in 50 µl DEPC treated water. Total RNA isolated from samples was further used for microarray and qPCR. The RNA samples were treated with DNase I (Qiagen, Hilden, Germany) and purified on RNeasy mini column (Qiagen, Hilden, Germany) before using for experiments to avoid genomic DNA contamination. In brief, after adjusting sample volume to 100 µl, 350 µl RLT buffer and 250 µl of absolute ethanol were added to it, the mixture was placed onto the column, 10 µl of DNase was added to the column for 30 minutes duration at room temperature, followed by two washes with 500 µl of buffer RPE and centrifugation at 10,000 rpm for 1 minute. Total RNA was eluted using 30 µl of RNase free water followed by centrifugation at 10,000 rpm for 3 minutes. Quantity and quality of the purified RNA were determined by Nanodrop (Thermofischer Scientific, U.S.A) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California) respectively. Tumour tissues with RNA integrity number (RIN) $\geq$7 were included in the study, in the case of controls, all the samples had RIN $\geq$7 except for two controls which had RIN of 6.8 and 6.1 respectively.

Gene expression profiling by microarray. Whole genome-wide expression profiling was done using HumanWG-6 v3.0 and HumanHT-12 v3 direct hybridization assay (Illumina, San Diego, CA) in 2 batches. The HumanWG-6 Bead Chip contains >48,000 probes, while HumanHT-12 chip contains the same panel of probes targeting more than 25,000 (human) genes from Reference Sequence (RefSeq) and UniGene database, from the National Center for Biotechnology Information (NCBI); but the later chip provides higher throughput processing of 12 samples per chip. In the present study a total of 29 tumour samples, including 12 Early-onset tumours (ET), from patients having age $\leq$40 years, and 17 Late-onset tumours (LT), from patients having age $\geq$55 years, along with 9 distant normal specimens as controls were used for gene expression profiling. Five hundred nanograms of total RNA was converted to complementary DNA (cDNA), followed by an *in vitro* transcription step to generate labeled complementary RNA(cRNA) using the 'Ambion Illumina Total Prep RNA Amplification Kit' (Ambion, Austin, TX) as per manufacturer's instructions. The labeled cRNA was hybridized to bead chip array and washed following manufacturer's protocols, which was scanned by 'Illumina Bead Array Reader, to obtain the raw data. The expression profiles of 29 cases and 9 controls have been deposited in NCBI's Gene Expression Omnibus (GEO) with GSE accession number GSE 89116.

Microarray data analysis. Raw data generated from the scanned slides was subjected to background correction followed by log2 transformation on Illumina's Genome Studio software. The data was quantile normalized using Linear Models for Microarray Data (LIMMA- v.3.36.1)[13] on Bioconductor package R.3.3.2[14]. Further, differentially expressed genes (DEGs) in Total tumours (TT), Early-onset tumours (ET) and Late-onset tumours (LT) compared to normal controls; and in various molecular subgroups viz. luminal, basal, HER2/neu were obtained using LIMMA. Molecular subtypes were predicted using 'molecular.subtyping' function on 'Genefu'[15] (v.2.12.0) using R as explained for 'Compare Molecular Subtype Classifications', in the 'Genefu' manual (bioconductor.org). Expression profiles obtained in the present study were compared with that of the western population, reported by Clarke *et al.*[16] and Maubant *et al.*[17] (GSE42568 and GSE65194 from the NCBI Gene Expression Omnibus).

Hierarchical clustering/gene ontology (GO) and network analysis. Unsupervised hierarchical clustering was performed for the DEGs by Cluster 3[18] software. The normalized probe intensities were median centered, Pearson correlation was used for similarity/distance measurement and centroid linkage clustering was performed. Further, JavaTree View Software[19] was used to view the clustering image. Gene ontology analysis was performed using Pathway Express software[20] (from Onto tools). Kyoto Encyclopedia of Genes and Genomes (KEGG) database[21] was used to determine specific pathways pertaining to differentially expressed genes. Further, Gene Set enrichment analysis (GSEA)[22] software was used to gain insights into the top 50 DEGs and also to generate a heatmap. Network analysis was done using 'NetworkAnalyst'[23], where protein-protein interactions networks were predicted using search tool for Recurring Instances of Neighboring Genes (STRING)[24].

Validation of differential gene expression by qPCR. Quantitative reverse transcription PCR was done for *MMP1*, *MMP3*, *MMP11*, *MMP13*, *MMP14*, *ADAMTS1*, *ADAMTS5*, *18sRNA*, *β-actin*, and *PSMC4* genes,

where *18sRNA*, *β-actin*, and *PSMC4* were used as endogenous controls for normalization of the qPCR data. QPCR validation of the above genes was done in 67 of the 77 tumours selected (depending on the availability of RNA) for the present study and 38 distant normal tissues along with 2 human mammary total RNA (Ambion, Austin, TX) (Supplementary Tables S1). The total RNA from the 38 distant normal tissues were combined to a single pool, which along with the 2 human mammary total RNAs (obtained from Ambion) were used as controls for determining the expression by qPCR.

The reverse transcription (RT) reaction was carried out using 1 μg of total RNA, random primers, and SuperScript III RT (Invitrogen, Thermo Fisher Scientific) at 50 °C for 50 minutes in a total reaction volume of 20 μL following the manufacturer's protocol. The cDNA generated by RT was diluted 5 folds and qPCR was carried out using 4.5 μL of the diluted cDNA, 5.0 μL of SYBR green mix (2X) and 0.25 pm of gene-specific primers in a total volume of 10 μL reaction. A non-RT control was also used during qPCRs to ensure lack of nonspecific amplification due to genomic DNA contamination. Most of the qPCR primers were designed (primer quest tool from Integrative DNA Technologies) such that they span exon junctions (except for *MMP11*, *MMP13*, *ADAMTS5*, and *ACTINB* genes) to avoid nonspecific amplification (Supplementary Table S3). The following cycling conditions were used for qPCRs, initial denaturation at 95 °C for 2 minutes, 40 cycles of 95 °C for 10 seconds and 60 °C for 1 minute on the StepOne Real-time PCR (ABI, Foster City, CA). All the samples were run in triplicate in a 96 well plate (ABI, Foster City, CA). The specificity of all the primers was confirmed by melting curve analysis on StepOne software v.2.3 (ABI, Foster City, CA). The mean Ct obtained for each gene was normalized with endogenous controls to obtain ΔCt, and further fold change (FC) was obtained by the ΔΔCt method on Data assist software v.3.01 (ABI, Foster City, CA).

Statistical analysis. Identification of differentially expressed genes (DEGs) was done by fitting gene-wise linear models to the gene expression data obtained from microarray experiments; the 'lmFit' function was used on LIMMA[13] for different groups (TT, ET, and LT compared to controls). False discovery rate was controlled by Benjamini Hochberg FDR[25] correction for a significant p-value cutoff of 0.05 and fold change $\geq \pm 1.5$. LIMMA was also used for determining DEGs (fold change $\geq \pm 2.0$ and p-value ≤ 0.05) from the western datasets, and for comparison with our data. Mann-Whitney U test was used to determine differential expression of MMP genes and their association with various clinicopathological parameters.

Results

Among 77 histopathologically confirmed cases of breast cancer enrolled for this study, 35 cases (45.45%) were below 40 years of age (ET) while, 42 cases (54.54%) were above 55 years of age (LT) (Supplementary Table S1). The tumours were staged into, stage I having 2 cases (2.59%), stage II having 32 cases (41.5%), stage III having 37 cases (48.05%) and stage IV having 3 cases (3.89%), while for 3 cases (3.89%) stage is not known. Molecular subtyping based on expression of ER, PR, and HER2/neu, yielded 33 cases (42.85%) to luminal subtype, 17 (22.07%) cases to basal subtype, and 23 cases (29.87%) to HER2/neu overexpressing subtype (Supplementary Table S2), and for 4 cases (5.19%) molecular subtype could not be determined as the expression of ER, PR and HER2/neu were not available for these cases.

Microarray analysis. Genome-wide expression profiling was done in 29 tumours, and 9 distant histologically confirmed normal control tissues, in the present study. The raw data obtained was analysed using LIMMA, a cutoff of FC ≥ 1.5 and p-value ≤ 0.05 were used for identification of differentially expressed genes. Volcano plots were drawn to get an overview of differential gene expression among different tumour groups (Fig. 1); it was observed that the proportion of down-regulated genes were more than up-regulated genes in TT, ET, and LT tumour groups.

Differential gene expression analysis. A total of 2413 differentially expressed genes (DEGs), including 991 up-regulated genes and 1422 down-regulated genes (Supplementary Table S4), were found in breast tumours (TT) compared to controls. The top up-regulated genes include *COL10A1*, *MMP11*, *GJB2*, *CST1*, *KIAA1199*, *MMP1*, *MMP13*, *CEACAM6*, *BUB1* and *ASPM* involved in cell cycle, focal adhesion and metastasis, while top down-regulated genes were *PLIN*, *KIAA1881*, *ADH1A*, *ADH1B*, *CIDEC*, *THRSP*, *GPD1*, *TIMP4*, *FABP4* and *C7* involved in lipid metabolism, and PPAR pathway (Table 1).

Hierarchical clustering and gene ontology. Unsupervised hierarchal clustering of the DEGs in breast tumours (TT), yielded two distinct clusters of up and down-regulated genes (Fig. 2, Supplementary Fig. S1). The topmost 50 genes were clustered separately for better visualisation of data (Supplementary Fig. S2). Further, pathway analysis was done to identify the biological pathways associated with breast cancer. The major pathways found to be deregulated include cell adhesion molecules, cell cycle, adherens junction, PPAR signalling, complement and coagulation cascades, focal adhesion, ECM-receptor interaction, DNA replication, adipocytokine signaling, pathways in cancer (Table 2). The pathways associated with up-regulated genes include cell cycle, systemic lupus erythematosus, DNA Replication, ECM-receptor interaction, p53 signalling (Supplementary Table S5) while the pathways associated with down-regulated genes include leukocyte transendothelial migration, cell adhesion molecules, adherens junction, complement and coagulation cascade, PPAR signaling, circadian rhythm, focal adhesion, adipocytokine signaling pathway, and tight junction (Supplementary Table S6).

Gene networking analysis. Gene network analysis was performed to identify the key regulatory genes among the DEGs found in the breast tumours (TTs). The topmost interactive up-regulated nodes include, *AURKB*, *CENPA*, *TOP2A*, *BUB1*, *CCNB2*, *MMP1*, and *SPP1* genes involved in cancer metastasis, cell cycle, and mitosis; and the down-regulated nodes include *CAV1*, *ACACB*, *NTRK2*, *KLF4*, and *MYH11* genes involved in regulation of Ras-ERK, fatty acid synthesis, MAP kinase and JAK2/STAT3/5, and ATP hydrolysis pathways (Fig. 3).

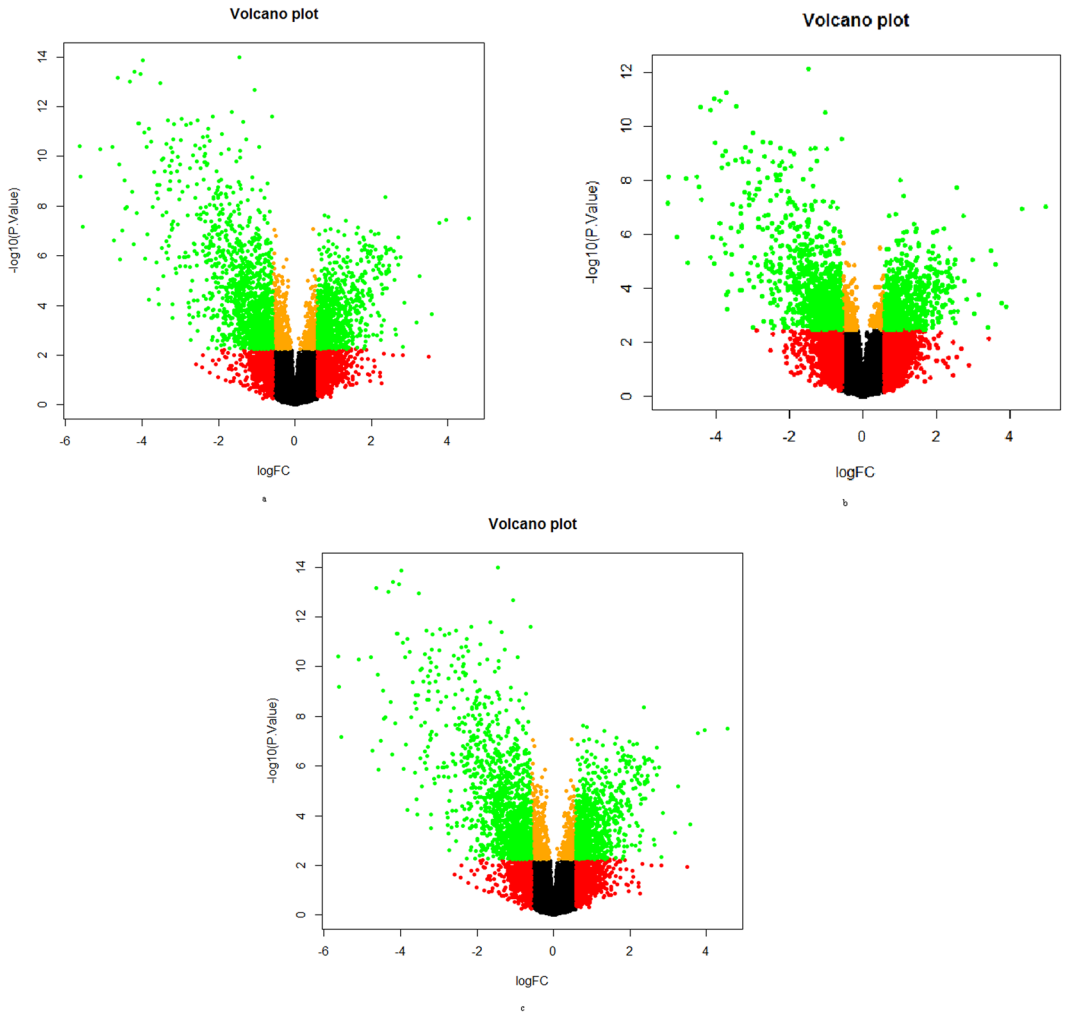

Figure 1. Volcano plots showing the distribution of gene expression by microarray in total breast tumours and early- and late-onset breast tumours as compared to controls. The plot shows gene expression profiles of breast tumours. The plot was obtained between negative log p-value (y-axis) and log fold change (x-axis). Each dot represents one gene, genes shown in green colour had significant fold change (FC $\geq$ 1.5, and adjusted p $\leq$ 0.05) while the remaining genes depicted in red, black and orange colour didn't reach significance. (**a**) Plot shows gene expression profiles of total tumours vs controls (**b**) Plot shows the gene expression profiles of early-onset tumours vs controls (**c**) Plot shows the gene expression profiles of late-onset tumours vs controls.

Comparison of DEGs between Indian and western patients. Gene expression signatures found in the present study were further compared with that of the gene expression profiles of breast tumours derived from the western population (GSE 65194 and GSE 42568). A total of 5062 DEGs, including 3789 up-regulated and 1273 down-regulated genes were found with western data set. Comparison of DEGs between Indian patients (found in the present study) with that of western, showed 715 genes (Supplementary Table S7) that are common between both the data sets, while 558 DEGs were associated with Indian patients (Supplementary Table S8). Pathway analysis of common DEGs among the two population showed deregulation of leukocyte transendothelial migration (*ESAM*, and *MYL7)*, cytokine receptor interaction (*IL17B*, *CNTFR*, *FIGF*, *MPL*, and *CCL21)*, and adherens junction (*PVRL3*, *PVRL4*, and *TCF7l1*) pathway.

DEGs in early- and late-onset breast cancer. Analysis of DEGs amongst ET and LT groups showed 1685 DEGs in ET and 2379 DEGs in LT compared to controls (Table 3, Supplementary Tables S9 and S10). When DEGs between ET and LT were compared, a majority of common and few uniquely expressed genes were found between the two groups; though there was a difference in terms of fold expression of the genes. When we used ANOVA to identify genes significantly associated with ET and LT, 420 genes were found to be associated with ET, 1114 genes were found to be associated with LT while, 1265 genes were common between both the groups (Supplementary Fig. S3). Pathways analysis of the DEGs revealed the involvement of similar pathways in ET and LT groups (Supplementary Tables S11 and S12) but, the genes associated with each of these pathways were found to be different. Cellular processes such as leukocyte transendothelial migration, cell adhesion molecules, PPAR signaling, cell cycle, ECM-receptor interaction pathways are some of the pathways that were altered in ET and LT.

GENE	(FC)	Adjusted.p-value	Accession
Up-regulated			
COL10A1	23.2815	5.96E-06	NM_000493.2
MMP11	15.45492	6.55E-06	NM_005940.3
GJB2	13.68954	8.27E-06	NM_004004.3
CST1	11.94889	0.004912	NM_001898.2
KIAA1199	9.509539	0.000351	NM_018689.1
MMP1	9.096327	0.008953	NM_002421.2
MMP13	7.297567	0.002203	NM_002427.2
CEACAM6	7.014203	0.042617	NM_002483.3
BUB1	6.795974	9.11E-05	NM_004336.2
ASPM	6.500159	2.39E-05	NM_018136.2
Down-regulated			
C7	−22.871	1.39E-05	NM_000587.2
FABP4	−23.8769	0.000104	NM_001442.1
TIMP4	−24.1717	1.03E-07	NM_003256.2
GPD1	−24.9122	3.83E-10	NM_005276.2
THRSP	−26.7033	3.01E-05	NM_003251.2
CIDEC	−27.4842	3.18E-08	NM_022094.2
ADH1B	−33.9347	3.6E-08	NM_000668.3
ADH1A	−46.9269	1.12E-05	NM_000667.2
KIAA1881	−48.6453	2.71E-07	Hs.567652

Table 1. Top ten up-regulated and down-regulated genes between breast tumours and controls.

Networking analysis in ET shown *BRAF* gene as a top overexpressed node while *SMAD3* gene was an top under-expressed node. The top nodes getting up-regulated exclusively in LT were *MCM2*, and *MAD2L1*, while *PXN*, and *SOCS3* were top down-regulated nodes.

DEGs in various stages of breast cancer. We analysed DEGs between lower stage (stage I and II) and advanced stage (stage III and IV) tumours, where 1386 DEGs associated with lower stage (stage I and II) and 200 DEGs associated with advanced stage (stage III and IV), and 1336 DEGs common between both the groups were found (Supplementary Tables S13 and S14). Gene ontology analysis of these DEGs showed that in advanced stage tumours, pathways such as cell adhesion (*VCAN*), ECM receptor interaction (*COL1A2*, *COL3A1*, *ITGA11*, and *TNN*) and pathways in cancer (*JUN*, and *MYC*) getting deregulated.

Molecular subtypes. Molecular subtyping based on gene expression profiles was done using Prediction Analysis of Microarray 50 (PAM50) classifier. Using 'Genefu' we could stratify tumour samples into, luminal-A consisting of 7 cases (24%), luminal-B consisting of 6 cases (20%), basal consisting of 9 cases (31%), HER2/neu overexpressing tumours consisting of 6 cases (20%) and normal-like subtype consisting of 1 case (3%). Further, unsupervised hierarchical clustering of PAM 50 genes yielded distinct gene clusters corresponding to each molecular subtypes (Fig. 4, Supplementary Fig. S4). This confirms the existence of molecular subtypes in breast tumours of Indian women, similar to that reported in the western population. In addition, we obtained DEGs associated with each subtype (Supplementary Table S15) consisting of 198, 415, 842, 705 and 39 (Fig. 5) genes unique to luminal-A, luminal-B, basal, HER2/neu overexpressing and normal-like subtypes respectively.

Validation of gene expression profiles. Gene expression profiling showed deregulation of several members of metallopeptidase activity/extracellular matrix activity genes in breast tumours and they formed part of top 50 DEGs. Hence we validated gene expression of *MMP1*, *MMP3*, *MMP11*, *MMP13*, *MMP14*, *ADAMTS1* and *ADAMTS5* genes belonging to metallopeptidase activity, using qPCR in 67 tumours. Overexpression of *MMP1*, FC = 15.4 ($p = 0.05$), *MMP11*, FC = 6.8 ($p = 0.03$) and *MMP13*, FC = 12.3 ($p = 0.018$) genes in breast tumours reached statistical significance, compared to their expression in controls (Supplementary Table S16); while, *MMP3* (FC = 6.8, $p = 0.214$) and *MMP14* (FC = 0.48, $p = 0.7$) genes did not show specific pattern of expression among tumours. Adamalysins genes/ A distintegrin and metalloproteinase with thrombospondin motifs (ADAMTS) family genes, *ADAMTS1* (FC = −9.4, $p = 0.009$), and *ADAMTS5* (FC = −5.7, $p = 0.05$) were significantly down-regulated in tumours, as compared to controls (Supplementary Table S16, Fig. 6). Further, it was found that the down-regulated expression of *ADAMTS5* was significantly associated with LT (FC = −6.5, $p = 0.013$, Supplementary Table S17). Though expression of various MMP genes was found to be relatively higher in ET compared to LT, it did not reach statistical significance. We further, analysed the differential expression of metallopeptidase genes for their association with various clinicopathological features. Overexpression of *MMP11* gene ($p = 0.04$) was significantly associated with the metastasis, while overexpression of *MMP1* was associated with loss of ER ($p = 0.01$), and PR ($p = 0.006$), on the other hand, overexpression of *MMP13* was found to be associated with overexpression of HER2/neu in patients ($p = 0.023$). The qPCR data confirmed the overexpression of several MMP genes in breast tumours that was observed using microarray experiments.

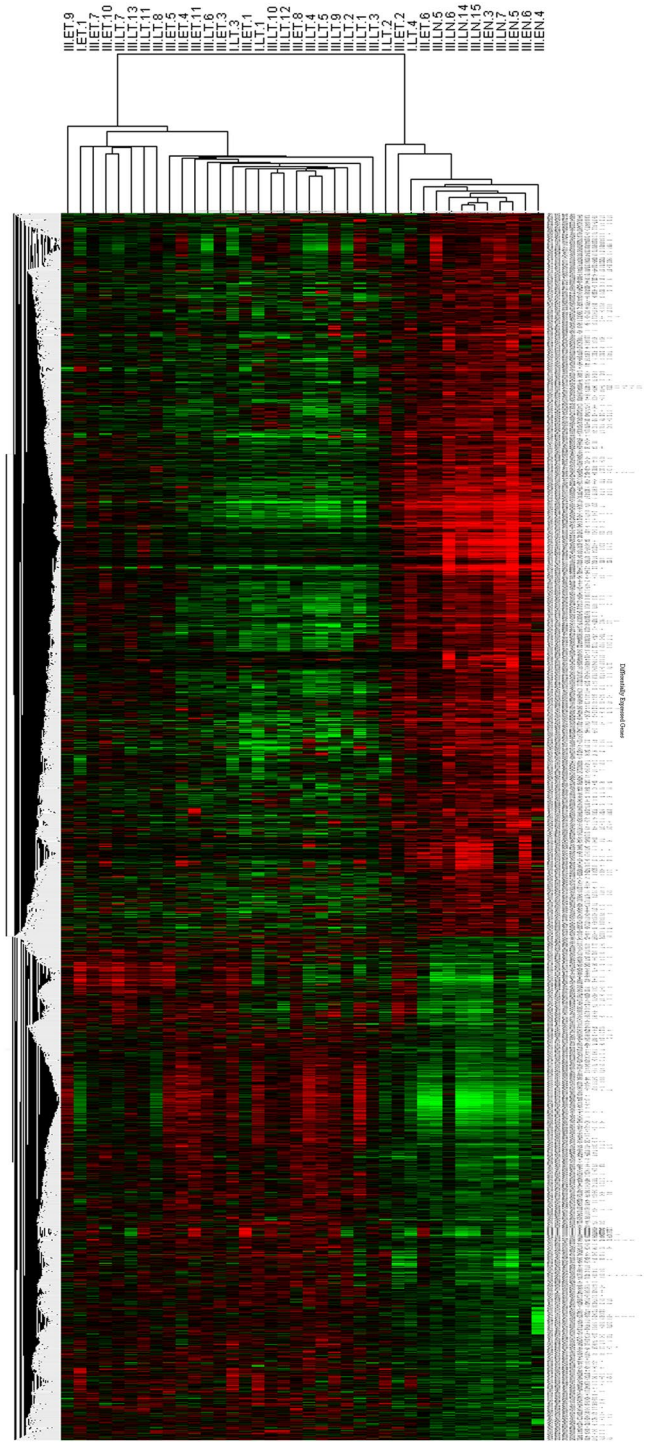

Figure 2. Unsupervised hierarchical clustering of differentially expressed genes. Heatmap showing the hierarchical clustering of tumours based on their gene expression. 2413 genes were found to be differentially expressed in tumours (FC $\geq$ 1.5, and adjusted p-value $\leq$ 0.05) forming distinct up-regulated and dow-nregulated clusters. Red colour represents up-regulation and green colour represents down-regulation. The differentially expressed genes are mentioned on the y-axis, and sample IDs are mentioned on the x-axis.

Discussion

Breast cancer incidence is increasing globally (24.2%) as well as in India (15.46%) and has become the most common cancer among Indian women[7,8]. To gain insight into the molecular mechanisms involved in the pathogenesis of breast cancer in Indian women, we have carried out gene expression profiling, wherein we have analysed the gene expression profiles associated with breast tumours and those associated with age and tumour stage. So far, there has been a single report where gene expression profiles of breast tumours were studied in Indian women,

Rank	Pathway Name	Impact Factor	Input Genes in Pathway	List of Genes	Corrected p-value
1	Cell adhesion molecules (CAMs)	235.367	21	*CADM3*, *CLDN15*, *CDH5*, *CLDN11*, *CLDN5*, *ESAM*, *ICAM2*, *JAM2*, *NLGN1*, *PECAM1*, *PTPRM*, *PVRL3*, *SELP*, *PVRL2*, *CD34*, *ITGA9*, *C10ORF54*, *ITGA7*, *LAMA2*, *LAMA3* and *NEGR1*	0.011143
2	Cell cycle	22.121	34	*E2F2*, *E2F3*, *CDC2*, *E2F5*, *CDCA5*, *PKMYT1*, *CDC45L*, *ORC1L*, *CHEK1*, *PTTG1*, *CCNE2*, *CCNE1*, *RAD21*, *MCM7*, *CDKN2C*, *BUB1*, *CCNA2*, *MYC*, *CDCA5*, *ESPL1*, *CDC20*, *MCM2*, *CDC25C*, *MCM4*, *CDC25A*, *MCM6*, *CDKN1C*, *CCNB1*, *CCNB2*, *MAD2L1*, *TTK*, *PLK1*, *GSK3B*, *and BUB1B*	8.12E-10
3	Adherens junction	20.559	15	*LEF1*, *PTPRB*, *PTPRM*, *WASF2*, *SORBS1*, *SSX2IP*, *PVRL3*, *PVRL2*, *PVRL4*, *CDH5*, *PRR4*, *CADM3*, *ANG*, *TCF7L2*, and *CDH23*	0.00458
4	PPAR signalling pathway	15.598	20	*ACADL*, *ACSL1*, *ADIPOQ*, *ANGPTL4*, *AQP7*, *CD36*, *EHHADH*, *FABP4*, *GK*, *LPL*, *MMP1*, *OLR1*, *PCK1*, *PLIN*, *PLTP*, *PPARG*, *SORBS1*, *RXRA*, *PPARGC1A*, and *NR1H3*	3.54E-06
5	Complement and coagulation cascades	15.474	19	*F12*, *C7*, *A2M*, *F10*, *F13A1*, *C6*, *SERPING1*, *C4BPA*, *PLAUR*, *VWF*, *THBD*, *SERPINF2*, *F3*, *CD59*, *CFH*, *TFPI*, *CFI*, *CFD*, and *PROS1*	1.40E-05
6	Focal adhesion	14.939	40	*COL3A1*, *ITGA11*, *LOXL4*, *CAV2*, *MYL7*, *CAV1*, *TLN2*, *FIGF*, *PXN*, *MYL9*, *LAMB2*, *ARHGAP5*, *TNN*, *PDGFD*, *COL11A1*, *SPP1*, *PIK3R2*, *THBS4*, *FN1*, *BRAF*, *IGF1*, *FLNC*, *COL5A2*, *COL5A1*, *LAMA2*, *VWF*, *VEGFC*, *ITGA9*, *LAMA4*, *PPP1CA*, *LAMA3*, *CCND2*, *JUN*, *GSK3B*, *ITGA7*, *COL1A2*, *RELN*, *COL1A1*, *COL24A1*, and *MYLK*	5.09E-06
7	ECM-receptor interaction	14.846	23	*IBSP*, *COL3A1*, *ITGA11*, *COL5A2*, *COL5A1*, *HMMR*, *LAMA2*, *VWF*, *LAMA4*, *LAMB2*, *CD36*, *LAMA3*, *ITGA7*, *COL1A2*, *RELN*, *TNN*, *SV2B*, *COL1A1*, *COL24A1*, *COL11A1*, *SPP1*, *FN1*, and *THBS4*	1.69E-06
8	DNA replication	10.812	12	*DNA2*, *RFC4*, *MCM7*, *POLE2*, *RFC2*, *PRIM2*, *MCM2*, *POLA2*, *RNASEH2A*, *MCM4*, *FEN1*, and *MCM6*	4.31E-05
9	Adipocytokine signaling pathway	8.699	13	*LEP*, *IRS2*, *ACSL1*, *CD36*, *SOCS3*, *LEPR*, *RXRA*, *ACACB*, *POMC*, *ADIPOQ*, *PPARGC1A*, *SLC2A1*, and *PCK1*	0.010329
10	Pathways in cancer	9.026	50	*ACVR1C*, *E2F2*, *E2F3*, *FGF7*, *STAT5A*, *SLC2A1*, *STAT5B*, *PPARG*, *MYC*, *FGF2*, *MMP1*,*CCNE2*, *WNT2*, *FOS*, *FIGF*, *CCNE1*, *TCF7L1*, *WNT11*, *LEF1*, *DAPK2*, *KIT*, *ZBTB16*, *IL6*, *BRAF*, *RXRA*, *FZD5*, *CEBPA*, *BMP2*, *VEGFC*, *EVI1*, *JUN*, *EPAS1*, *GNAI1*, *FOXO1*, *FN1*, *TGFBR2*, *TCF7L2*, *FZD4*, *LAMB2*, *LAMA2*, *PIK3R2*, *IGF1*, *BIRC5*, *LAMA4*, *LAMA3*, *GSK3B*, *SMAD3*, *NRAS*, *WNT7B*, and *BAX*	6.97E-04

Table 2. Gene ontology analysis of DEGs found in the breast tumours.

however, the authors have analysed gene expression profiles of ER positive and negative tumours, where they have found four hormone-responsive genes as DEGs[11]. To our knowledge, this is the first study describing comprehensive gene expression profiles in Indian women and demonstrating the existence of molecular subtypes using gene expression profiles in breast tumours.

The present study identified 2413 differentially expressed genes comprising of top up-regulated genes such as *COL10A1*, *COL11A1*, *MMP1*, *MMP13*, *MMP11*, *GJB2*, *CST1*, *KIAA1199*, *CEACAM6*, and *BUB1*; top down-regulated genes comprising of *PLIN1*, *FABP4*, *LIPE*, *AQP7*, *LEP*, *ADH1A*, *ADH1B*, *CIDEC*, *THRSP*, *GPD1*, *TIMP4*, and *KIAA1881* (Supplementary Fig. S2, Supplementary Table S2). Among the top DEGs, up-regulated expression of genes such as *COL10A1*, *MMP1*, *MMP11*, and *BUB1*; down-regulated expression of genes such as *ADH1B*, *CIDEC*, *FABP4*, *AQP7*, *RBP4*, *CDO1*, *FIGF*, and *LPL* were reported to be differentially expressed in breast and or other cancers by various authors using microarray profiling in western population[26–38], showing concordance with the present study. In the present study, we found up-regulation of cell cycle genes such as *BUB1*, *CCNA2*, *CCNB2*, and *CDC2*; up-regulated expression of *BUB1*, *CCNA2*, *CCNB2* and *CDC2* has also been reported in breast[39,40] and several other cancers[41–46] using microarray and were found to be associated with poor prognosis of the disease[44,47–49]. Overexpression of the above cell cycle genes may be contributing to the uncontrolled proliferation of the tumour cells and hence may serve as biomarkers and targets for therapy.

Overexpression of genes involved in DNA replication such as *MCM2*, *MCM6*, *MCM10*, and *RAD21* was observed in the present study, increased expression of *MCM2*, *MCM6*, and *MCM10* have been reported in breast and several other epithelial malignancies by transcriptome profiling, and was associated with poor prognosis[50–52]. Furthermore, the focal adhesion genes such as *COL1A1*, *COL10A1*, and *COL11A1* were also found to be up-regulated in the present study and are also reported to be up-regulated in various cancers including breast tumours[37,53–55]. In cancer cells, collagen gene expression is known to increase drug resistance by inhibiting drug penetration as well as cause an increase in apoptosis resistance, thus, in turn, promoting tumour progression[37,56–58].

Figure 3. Showing gene network analysis of differentially expressed genes in breast tumours. Interactive gene networks were identified based on their position in the network by two measures; degree centrality, where the importance of a node is dependent on the number of connections to other nodes and betweenness centrality, which measures the number of shortest paths going through a node. Nodes with a higher degree are hubs of the network, and the size of the nodes is based on their degree values, with a bigger size accounting for larger degree values. The colour of the nodes is related to the expression of genes, where up-regulated nodes are shown in red and down-regulated nodes in green colour while the grey coloured nodes are those genes that are not present in our data set but are part of the PPI network (The network analysis was done with DEGs having FC $\geq \pm 5$). Among the gene networks, *AURKB*, *CENPA*, *TOP2A*, *BUB1*, *CCNB2*, *MMP1*, and *SPP1* were the most interactive nodes.

In the present study genes such as *PLIN1*, *FABP4*, *LIPE*, *LEP*, *CIDEC*, *THRSP*, *AQP7*, *ADH1A*, *ADH1B*, *GPD1*, and *TIMP4* were found to be down-regulated, involved mainly in lipid metabolism, lipolysis, oxidoreductase activity, and PPAR pathways. In concordance with the above findings, down-regulated expression of lipid metabolism genes such as *LEP*, *CIDEC*, *THRSP*, *PLIN1*, *GPD1*, and *FABP4* genes were also reported at the transcript level by various authors in breast[54,59–64] and other cancers such as gastric[65], hepatocellular[66] and keratoacanthomas[67]. Similar to that observed in the present study, down-regulated expression of aquaporin, *AQP7* gene belonging to water channel family, *TIMP4* belonging to mettalloproteinases inhibitor family member was also reported in breast and hepatocellular carcinoma at transcript level[68–70]. Contrary to the down-regulation observed in the present study, up-regulation of *FABP4*, *LEP*, *CIDEC* genes was reported at transcript and or protein level[35,71–74] in lung, thyroid, colorectal, and tongue squamous cell carcinoma; these differences may be attributed to tissue-specific differences in gene expression, differences due to the techniques employed in the studies, which need to be established by experimental validation.

Networking analysis was done to identify genes involved in regulation of gene expression in cancer cells, *AURKB*, *CENPA*, *TOP2A*, *BUB1*, *CCNB2*, *MMP1*, and *SPP1* were identified as top hub genes from the up-regulated genes; suggesting these genes might be playing key regulatory role in breast carcinogenesis through deregulation of cell cycle and in invasion/metastasis. Similarly, genes such as *CAV1*, *ACACB*, *NTRK2*, *KLF4*, and *MYH11* were the key down-regulated hub genes suggesting a possible role of their decreased expression in breast carcinogenesis.

Comparison of gene expression profiles of Indian patients with that of western patients led to the identification of 558 genes specifically found to be deregulated in Indian patients, suggesting some differences in the gene sets between these populations. The differences in DEGs among the two populations may be partly due to differences in platforms, experimental procedures or genetic makeup of the populations. Up-regulated expression of *COL10A1*, *MMP11*, *CST1*, *GJB2*, *MMP1*, *MMP13*, and *CEACAM6*; down-regulated expression of *ADH1B*, *CIDEC*, *THRSP*, *GPD1*, *TIMP4*, *FABP4*, and *SCARA5* genes was common in breast cancers of the two populations. The similarity in the DEGs between the Indian and western patients suggests a similarity in the molecular events associated with breast carcinogenesis. Further, we compared DEGs obtained in the present study with that of Lebanese population[54], where several genes were found to be common between the two populations. Up-regulated expression of *COL11A1*, *GJB2*, *MMP13*, *EPYC*, *CEP55*, and *MELK* and down-regulated expression of *PLIN*, *TIMP4*, *LEP*, *LYVE1*, *SDPR*, *FIGF*, and *LPL* was observed in common with the Lebanese population from

Genes	(FC)_Early-onset tumours	Adjusted p-value	Accession
(a) Top 10 genes in early-onset tumours			
Up-regulated			
COL10A1	31.47292482	1.46E-07	NM_000493.2
MMP11	20.16883881	1.62E-07	NM_005940.3
CST1	15.09417634	0.000998	NM_001898.2
MMP1	13.78096344	0.001196	NM_002421.2
KIAA1199	12.24468581	2.69E-05	NM_018689.1
GJB2	11.13842735	2.76E-07	NM_004004.3
S100P	10.67661335	0.011113	NM_005980.2
MMP13	8.92393836	0.000343	NM_002427.2
PITX1	8.241605007	0.00272	NM_002653.3
TUBB3	7.974669152	2.51E-05	NM_006086.2
Down-regulated			
ALDH1L1	−17.95091861	1.08E-12	NM_012190.2
G0S2	−21.35394129	9.76E-09	NM_015714.2
GPD1	−21.60426989	6.70E-13	NM_005276.2
TIMP4	−22.50528229	2.24E-09	NM_003256.2
CIDEC	−23.24428146	3.64E-10	NM_022094.2
ADH1A	−27.50733965	1.64E-07	NM_000667.2
ADH1B	−28.39338049	4.42E-10	NM_000668.3
THRSP	−33.98850427	1.47E-06	NM_003251.2
PLIN	−39.64517112	2.95E-10	NM_002666.3
KIAA1881	−40.50151762	5.61E-09	Hs.567652
(b) Top 10 genes in late-onset tumours			
Up-regulated			
COL10A1	18.81883032	1.46E-07	NM_000493.2
GJB2	15.83475077	2.76E-07	NM_004004.3
MMP11	12.80727039	1.62E-07	NM_005940.3
CST1	10.13195192	0.000998	NM_001898.2
CEACAM6	8.845061876	0.012935	NM_002483.3
KIAA1199	7.955397998	2.69E-05	NM_018689.1
BUB1	7.409286293	6.92E-06	NM_004336.2
ASPM	7.250017904	1.01E-06	NM_018136.2
FAM83D	7.144854585	1.29E-05	NM_030919.1
CKAP2L	6.929207833	6.64E-06	NM_152515.2
Down-regulated			
TIMP4	−25.42177657	2.24E-09	NM_003256.2
C7	−27.25711853	5.32E-07	NM_000587.2
GPD1	−27.54770546	6.70E-13	NM_005276.2
C2ORF40	−28.62109977	2.00E-08	NM_032411.1
CIDEC	−30.93481391	3.64E-10	NM_022094.2
FABP4	−36.04296806	1.88E-06	NM_001442.1
ADH1B	−38.48568016	4.42E-10	NM_000668.3
KIAA1881	−55.36142413	5.61E-09	Hs.567652
PLIN	−58.36652422	2.95E-10	NM_002666.3
ADH1A	−68.41751514	1.64E-07	NM_000667.2

Table 3. Top ten up-regulated and down-regulated genes between (a) Early-onset breast tumours and controls and (b) Late -onset- tumours and controls.

the top 50 genes found in the present study. This is pointing towards a possible existence of greater similarity in the molecular pathogenesis of breast cancers amongst the Asian population compared to the western population.

Comparison of expression profiles of ET (≤40 years) and LT (≥55 years) yielded few genes that are unique between the two groups, 7 genes *B4GALNT1*, *S100P*, *KLK4*, *HIST3H2A*, *DRD4*, *PCSK1N*, and *BAPX1* were significantly overexpressed in early-onset tumours compared to late-onset tumours. Overexpression *B4GALNT1* causes deregulation of glycosphingolipid biosynthesis and is reported to be up-regulated in breast cancer stem cells[75] at the transcript level, similarly, *S100P*[76], *KLK4*[77,78], *DRD4*[79], and *BAPX1*[80] genes were also reported to be up-regulated in breast and other cancers at mRNA level. Several of these genes are known to induce invasion and

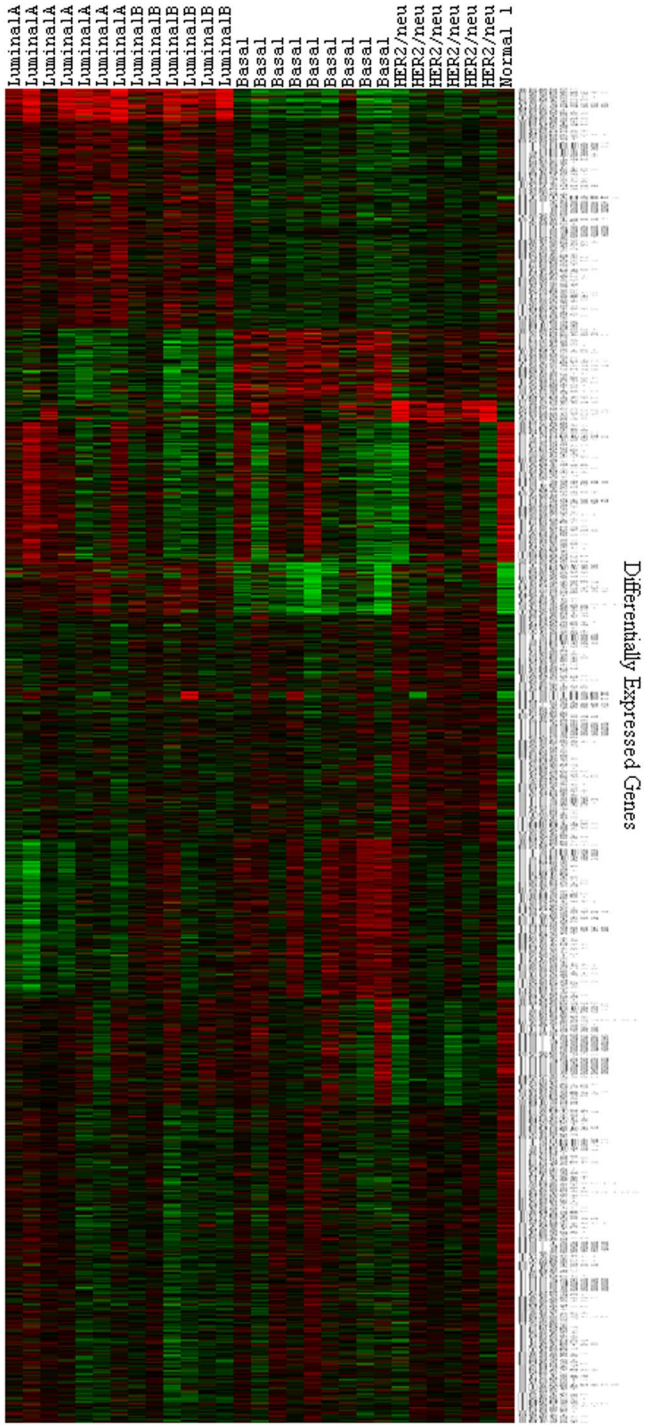

Figure 4. Heatmap showing hierarchical clustering of predicted molecular subtypes. Molecular subtypes were predicted using PAM50 classifier in breast tumours, consisting of subtypes viz. luminal A, luminal B, HER2/neu, basal and normal-like ($FC \geq 1.5$, and adjusted p-value ≤ 0.05). Genes pertaining to each subtype formed distinct clusters. Red colour represents up-regulation and green colour depicts down-regulation. The subtypes are mentioned on x-axis while differentially expressed genes are mentioned on the y-axis.

metastasis[81–84], breast cancer in young patients is known to be aggressive[85–88], the overexpression of these genes may be thus contributing to the aggressive behavior of the early-onset cancers. Anders *et al.*[89–91] have analysed gene expression profiles between early-onset patients and late-onset patients, where 693 DEGs were found initially, later the significance was lost when they have corrected the gene differences for subtypes and for ER and histological grades. However, such segregated analysis could not be carried out in the present study due to the small sample size. Further, we compared gene expression profiles between lower and advanced stages of tumours; we identified 200 genes uniquely deregulated in advanced stage cancers, involving pathways such as cell adhesion

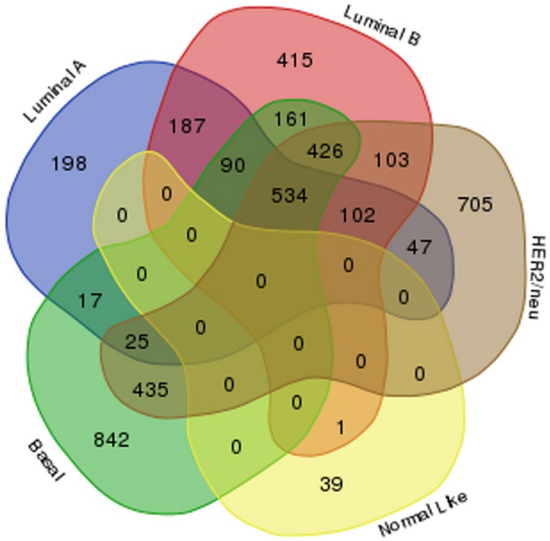

Figure 5. Venn diagram showing the common and unique genes belonging to each molecular subtypes in breast tumours. Venn diagram showing differentially expressed genes unique in basal subtype (842) followed by HER2/neu (705), luminal B & A (415, 198) and normal-like subtypes (39).

(*VCAN*), ECM receptor interaction (*COL1A2*, *COL3A1*, *ITGA11*, and *TNN*) and pathways in cancer (*JUN*, and *MYC*) which may be contributing to increased proliferation, migration and increased angiogenesis in advanced stage of tumours.

Molecular subtyping and hierarchical clustering of gene expression profiles of breast tumours using PAM50 molecular signatures, yielded distinct clusters corresponding to each molecular subtype, showing the existence of molecular subtypes in these patients. Among these patients, 44% tumours falling into luminal subtype, 31% into basal subtype and 20% into HER2/neu overexpressing subtype, and 3% into normal-like subtype, which is more or less similar to that reported in the western population[28,29]. To our knowledge this is the first study acknowledging the existence of molecular subtypes from the Indian subcontinent based on gene expression profiles; earlier, Kumar *et al.*[92] reported the existence of molecular subtypes based on the expression of ER, PR, HER2/neu and cytokeratins at protien level, however, transcriptome-based subtyping has not been demonstrated so far.

Matrix metalloproteinases (MMPs) belong to Zn2+-dependent endopeptidases family, capable of catalyzing the hydrolysis of collagen, forming a major part of extracellular matrix (ECM) remodelling[93,94]. Metallopeptidases genes were one of the functional class of genes found to have deregulated expression in breast cancers in the present study and hence we validated some of the genes by qPCR. QPCR analysis confirmed the up-regulated expression of *MMP1*, *MMP13*, and *MMP11* genes and down-regulated expression of *ADAMTS1* and *ADAMTS5* genes in breast cancers observed in the microarray experiments. Overexpression of *MMP1*, *MMP11*, and *MMP13* was also reported in breast[95] and several other cancers such as gastric, oral, colorectal, oesophageal and nasopharyngeal at the transcript and or protein level[30,54,96–103]. Upregulated expression of MMPs has been reported in cancer, vascular diseases and many different types of inflammatory diseases[104], their overexpression results in increased invasion and metastasis in cancer cells[105]. *A disintegrin and metalloproteinase with thrombospondin motif* (*ADAMTS*), superfamily genes play, an important role in ECM assembly and degradation, several of them act as tumour or metastasis suppressors by influencing cell proliferation, migration, apoptosis, and angiogenesis[106]. In the present study, *ADAMTS1* and *ADAMTS5* genes were observed to be underexpressed, down-regulated expression and antitumour acitivity of these genes were reported in breast cancers[107] and gastric carcinoma carcinoma[108] respectively at both transcript and protein. An association between overexpression of *MMP1* transcripts, with loss of ER ($p = 0.01$), and PR ($p = 0.006$) was found in the present study, Nakopoulou *et al.*[109] has also found a similar inverse association with PR expression at the protein level, supporting findings of the present study. Further, overexpression of *MMP13* was found to be associated with overexpression of HER2/neu in patients ($p = 0.023$), Zhang *et al.*[110] have also reported similar association in breast cancer with protein level expression. Interestingly, down-regulation of *ADAMTS5* was found to be associated with late-onset tumours ($\geq$55 years) compared to ET ($\leq$40 years), ($FC = -6.5$ in LT and $FC = -4.5$ in ET, $p = 0.013$), suggesting the involvement of loss of this gene in the molecular pathogenesis with late-onset breast cancer. To our knowledge, the down-regulated expression of *ADAMTS5* in breast tumours and its association with late-onset breast cancer (old age of patient) is reported for the first time in the present study. Together, deregulated expression of these matrix remodeling factors in breast tumours may be contributing to the degradation of ECM and invasion and metastasis in breast tumours, suggesting a pivotal role played by these genes in breast tumorigenesis. However, up-regulation of *MMP3* and *MMP14* genes didn't reach statistical significance, unlike found in microarray data. The discrepancy between microarray and qPCR data could be due to a different number of samples analysed by each method, to some extent, tumour heterogeneity might have also contributed to such differences.

The present study describes comprehensive gene expression profiles of breast tumours from Indian women and the presence of molecular subtypes in this population. Genes involved in cell cycle, ECM, metastasis were

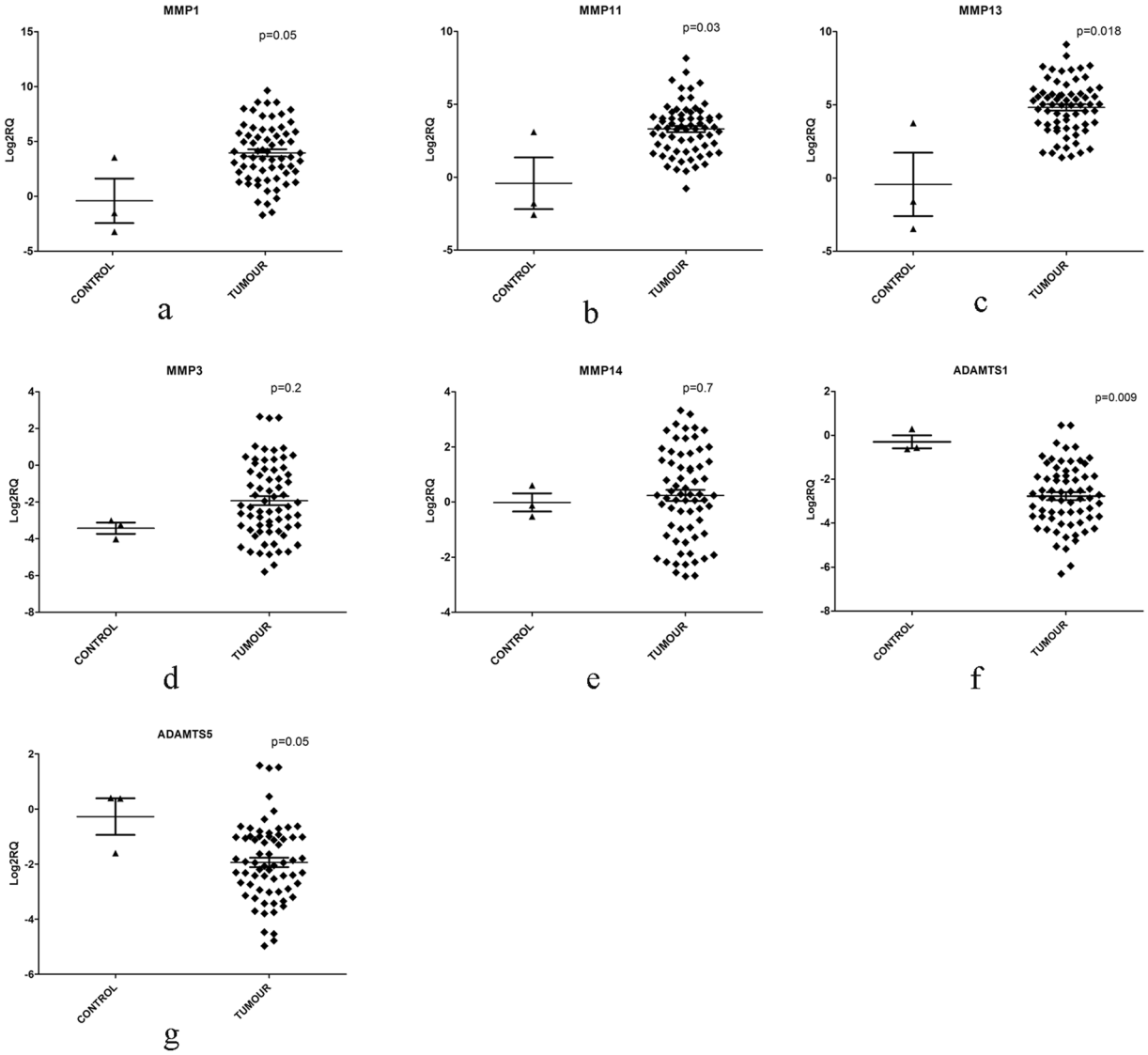

Figure 6. Validation of gene expression of MMPs by quantitative reverse transcription PCR. Scatter plots showing the up-regulation of (**a**) *MMP1* (p = 0.05), (**b**) *MMP11* (p = 0.03), (**c**) *MMP13* (p = 0.018) (**d**) *MMP3* (p = 0.214), (**e**) *MMP14* (p = 0.722) and down-regulation of *ADAMTS1* (p = 0.009), (**g**) *ADAMTS5* (p = 0.05) in breast tumours compared to controls. The values are the mean of log fold change normalized to endogenous controls, along with the standard error (shown by vertical bars) as obtained by Mann-Whitney U test.

some of the essential pathways found to be up-regulated, on the other hand genes involved in lipid metabolism, PPAR were some of the pathways that were found down-regulated. Genes belonging to cell adhesion, cell cycle, ECM receptor interaction pathways were deregulated in early-onset breast cancers. This study confirmed the presence of molecular subtypes in breast tumours based on gene expression profiles, for the first time from Indian patients. Comparison with western data has revealed the presence of several deregulated genes that are common between Indian and western patients suggesting a similarity in the molecular mechanisms; however, a higher similarity was with that of the Asian population. Comparison of gene expression profiles in early- and late-onset tumours showed several common DEGs between the groups, but with differences in fold change of their gene expression. Further, significant down-regulation of *ADAMTS5* in old age patients had been reported for the first time in breast cancer patients.

Limitations. The current study describes gene expression profiles from Indian breast cancer patients, yet it has some limitations. In the present study, we have analysed gene expression profiles of 29 tumours and 9 controls, however, to extrapolate this outcome to the breast cancer patients in the Indian subcontinent, gene expression profiles in larger patients set covering various geographical regions in India is warranted. Another limitation of the study is that genes which were found to be differentially expressed in microarray and qPCR, could not be validated at the protein level due to limited resources.

References

1. The, L. GLOBOCAN 2018: counting the toll of cancer. *Lancet* **392**, 985 (2018).
2. Shin, H. R. *et al*. Recent trends and patterns in breast cancer incidence among Eastern and Southeastern Asian women. *Cancer Causes Control* **21**, 1777–1785 (2010).
3. Wang, N. *et al*. Time trends of cancer incidence in urban beijing, 1998–2007. *Chin J Cancer Res* **23**, 15–20 (2011).
4. Jung, K. W. *et al*. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2008. *Cancer Res Treat* **43**, 1–11 (2011).
5. Leong, S. P. *et al*. Is breast cancer the same disease in Asian and Western countries? *World J Surg* **34**, 2308–2324 (2010).
6. Malvia, S., Bagadi, S. A., Dubey, U. S. & Saxena, S. Epidemiology of breast cancer in Indian women. *Asia Pac J Clin Oncol* (2017).
7. Anonymous. *Three Year Report of Population Based Cancer Registries 2012–2014*., (Indian Council of Medical Research(ICMR), Bangalore, India, 2016).
8. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359–386 (2015).
9. Chouchane, L., Boussen, H. & Sastry, K. S. Breast cancer in Arab populations: molecular characteristics and disease management implications. *Lancet Oncol* **14**, e417–424 (2013).
10. Chopra, B. *et al*. Age shift: Breast cancer is occurring in younger age groups - Is it true? *Clinical Cancer Investigation. Journal* **3**, 526–529 (2014).
11. Thakkar, A. D. *et al*. Identification of gene expression signature in estrogen receptor positive breast carcinoma. *Biomark Cancer* **2**, 1–15 (2010).
12. Mishra, A. K. *et al*. Expression of androgen receptor in breast cancer & its correlation with other steroid receptors & growth factors. *The Indian journal of medical research* **135**, 843–852 (2012).
13. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
14. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314 (1996).
15. Gendoo, D. M. *et al*. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097–1099 (2016).
16. Clarke, C. *et al*. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* **34**, 2300–2308 (2013).
17. Maubant, S. *et al*. Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. *PLoS One* **10**, e0122333 (2015).
18. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868 (1998).
19. Saldanha, A. J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
20. Khatri, P., Draghici, S., Ostermeier, G. C. & Krawetz, S. A. Profiling gene expression using onto-express. *Genomics* **79**, 266–270 (2002).
21. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
22. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550 (2005).
23. Xia, J., Benner, M. J. & Hancock, R. E. NetworkAnalyst–integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res* **42**, W167–174 (2014).
24. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* **28**, 3442–3444 (2000).
25. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279–284 (2001).
26. Perou, C. M. *et al*. Molecular portraits of human breast tumours. *Nature* **406** (2000).
27. Sorlie, T. *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98** (2001).
28. Sorlie, T. *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* **100**, 8418–8423 (2003).
29. Hu, Z. *et al*. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96 (2006).
30. Fu, J., Khaybullin, R., Zhang, Y., Xia, A. & Qi, X. Gene expression profiling leads to discovery of correlation of matrix metalloproteinase 11 and heparanase 2 in breast cancer progression. *BMC Cancer* **15**, 473 (2015).
31. Wang, J. *et al*. Alcohol consumption and breast tumor gene expression. *Breast Cancer Research* **19**, 108 (2017).
32. Lee, S. *et al*. Differentially expressed genes regulating the progression of ductal carcinoma *in situ* to invasive breast cancer. *Cancer Res* **72**, 4574–4586 (2012).
33. Kondrakhin, Y. V., Sharipov, R. N., Keld, A. E. & Kolpakov, F. A. Identification of differentially expressed genes by meta-analysis of microarray data on breast cancer. *In silico biology* **8**, 383–411 (2008).
34. Wang, Y., Zhang, Y., Huang, Q. & Li, C. Integrated bioinformatics analysis reveals key candidate genes and pathways in breast cancer. *Mol Med Rep* **17**, 8091–8100 (2018).
35. Boldrup, L. *et al*. Gene expression changes in tumor free tongue tissue adjacent to tongue squamous cell carcinoma. *Oncotarget* **8**, 19389–19402 (2017).
36. Meller, S. *et al*. CDO1 promoter methylation is associated with gene silencing and is a prognostic biomarker for biochemical recurrence-free survival in prostate cancer patients. *Epigenetics* **11**, 871–880 (2016).
37. Chapman, K. B. *et al*. COL10A1 expression is elevated in diverse solid tumor types and is associated with tumor vasculature. *Future oncology* **8**, 1031–1040 (2012).
38. Norton, N. *et al*. Assessment of Tumor Heterogeneity, as Evidenced by Gene Expression Profiles, Pathway Activation, and Gene Copy Number, in Patients with Multifocal Invasive Lobular Breast Tumors. *PLoS One* **11**, e0153411 (2016).
39. Zhou, H., Lv, Q. & Guo, Z. Transcriptomic signature predicts the distant relapse in patients with ER+ breast cancer treated with tamoxifen for five years. *Molecular medicine reports* **17**, 3152–3157 (2018).
40. Wang, Z. *et al*. Biological and Clinical Significance of MAD2L1 and BUB1, Genes Frequently Appearing in Expression Signatures for Breast Cancer Prognosis. *PLoS One* **10**, e0136246 (2015).
41. Yan, H. *et al*. Aberrant expression of cell cycle and material metabolism related genes contributes to hepatocellular carcinoma occurrence. *Pathol Res Pract* (2017).
42. Huang, R. & Gao, L. Identification of potential diagnostic and prognostic biomarkers in non-small cell lung cancer based on microarray data. *Oncology letters* **15**, 6436–6442 (2018).
43. Davidson, B. *et al*. BUB1 mRNA is significantly co-expressed with AURKA and AURKB mRNA in advanced-stage ovarian serous carcinoma. *Virchows Archiv: an international journal of pathology* **464**, 701–707 (2014).
44. Zhang, C. *et al*. Combined analysis identifies six genes correlated with augmented malignancy from non-small cell to small cell lung cancer. *Tumour Biol* **37**, 2193–2207 (2016).
45. Bednarek, K. *et al*. Recurrent CDK1 overexpression in laryngeal squamous cell carcinoma. *Tumour Biol* **37**, 11115–11126 (2016).
46. Han, Y., Jin, X., Zhou, H. & Liu, B. Identification of key genes associated with bladder cancer using gene expression profiles. *Oncology letters* **15**, 297–303 (2018).

47. Tong, H. *et al*. Transcriptomic analysis of gene expression profiles of stomach carcinoma reveal abnormal expression of mitotic components. *Life Sci* **170**, 41–49 (2017).
48. Shubbar, E. *et al*. Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome. *BMC Cancer* **13**, 1 (2013).
49. Chae, S. W. *et al*. Overexpressions of Cyclin B1, cdc2, p16 and p53 in human breast cancer: the clinicopathologic correlations and prognostic implications. *Yonsei medical journal* **52**, 445–453 (2011).
50. Tachibana, K. E., Gonzalez, M. A. & Coleman, N. Cell-cycle-dependent regulation of DNA replication and its relevance to cancer pathology. *The Journal of pathology* **205**, 123–129 (2005).
51. Kwok, H. F. *et al*. Prognostic significance of minichromosome maintenance proteins in breast cancer. *American journal of cancer research* **5**, 52–71 (2015).
52. Al-Ejeh, F. *et al*. Meta-analysis of the global gene expression profile of triple-negative breast cancer identifies genes for the prognostication and treatment of aggressive breast cancer. *Oncogenesis* **3**, e100 (2014).
53. Pepin, F. *et al*. Gene-expression profiling of microdissected breast cancer microvasculature identifies distinct tumor vascular subtypes. *Breast cancer research: BCR* **14**, R120 (2012).
54. Makoukji, J. *et al*. Gene expression profiling of breast cancer in Lebanese women. *Sci Rep* **6**, 36639 (2016).
55. Tian, Z. Q. *et al*. Identification of Commonly Dysregulated Genes in Non-small-cell Lung Cancer by Integrated Analysis of Microarray Data and qRT-PCR Validation. *Lung* **193**, 583–592 (2015).
56. Januchowski, R. *et al*. Increased Expression of Several Collagen Genes is Associated with Drug Resistance in Ovarian Cancer Cell Lines. *Journal of Cancer* **7**, 1295–1310 (2016).
57. Wu, Y. H., Chang, T. H., Huang, Y. F., Huang, H. D. & Chou, C. Y. COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene* **33**, 3432–3440 (2014).
58. Li, J., Ding, Y. & Li, A. Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer. *World journal of surgical oncology* **14**, 297 (2016).
59. Zhou, C. *et al*. Prognostic significance of PLIN1 expression in human breast cancer. *Oncotarget* **7**, 54488–54502 (2016).
60. Kim, S., Lee, Y. & Koo, J. S. Differential expression of lipid metabolism-related proteins in different breast cancer subtypes. *PLoS One* **10**, e0119473 (2015).
61. Karim, S. *et al*. Low expression of leptin and its association with breast cancer: A transcriptomic study. *Oncology reports* **36**, 43–48 (2016).
62. Shi, Y. *et al*. Integrative Comparison of mRNA Expression Patterns in Breast Cancers from Caucasian and Asian Americans with Implications for Precision Medicine. *Cancer research* **77**, 423–433 (2017).
63. Merdad, A. *et al*. Transcriptomics profiling study of breast cancer from Kingdom of Saudi Arabia revealed altered expression of Adiponectin and Fatty Acid Binding Protein4: Is lipid metabolism associated with breast cancer? *BMC Genomics* **16**(Suppl 1), S11 (2015).
64. Zhou, C. *et al*. Identification of glycerol-3-phosphate dehydrogenase 1 as a tumour suppressor in human breast cancer. *Oncotarget* **8**, 101309–101324 (2017).
65. Jin, Y. & Da, W. Screening of key genes in gastric cancer with DNA microarray analysis. *European journal of medical research* **18**, 37 (2013).
66. Zhong, C. Q. *et al*. FABP4 suppresses proliferation and invasion of hepatocellular carcinoma cells and predicts a poor prognosis for hepatocellular carcinoma. *Cancer medicine* **7**, 2629–2640 (2018).
67. Ra, S. H. *et al*. Keratoacanthoma and squamous cell carcinoma are distinct from a molecular perspective. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc* **28**, 799–806 (2015).
68. Chen, X. F., Li, C. F., Lu, L. & Mei, Z. C. Expression and clinical significance of aquaglyceroporins in human hepatocellular carcinoma. *Molecular medicine reports* **13**, 5283–5289 (2016).
69. Zhu, L. *et al*. Significant prognostic values of aquaporin mRNA expression in breast cancer. *Cancer management and research* **11**, 1503–1515 (2019).
70. Zhao, Y. G. *et al*. Endometase/matrilysin-2 in human breast ductal carcinoma *in situ* and its inhibition by tissue inhibitors of metalloproteinases-2 and -4: a putative role in the initiation of breast cancer invasion. *Cancer research* **64**, 590–598 (2004).
71. Tang, Z. *et al*. Elevated expression of FABP3 and FABP4 cooperatively correlates with poor prognosis in non-small cell lung cancer (NSCLC). *Oncotarget* **7**, 46253–46262 (2016).
72. Lee, D. *et al*. Expression of fatty acid binding protein 4 is involved in the cell growth of oral squamous cell carcinoma. *Oncology reports* **31**, 1116–1120 (2014).
73. Akinci, M. *et al*. Leptin levels in thyroid cancer. *Asian journal of surgery* **32**, 216–223 (2009).
74. Koda, M., Sulkowska, M., Kanczuga-Koda, L., Surmacz, E. & Sulkowski, S. Overexpression of the obesity hormone leptin in human colorectal cancer. *Journal of clinical pathology* **60**, 902–906 (2007).
75. Liang, Y. J. *et al*. Differential expression profiles of glycosphingolipids in human breast cancer stem cells vs. cancer non-stem cells. *Proc Natl Acad Sci USA* **110**, 4968–4973 (2013).
76. Prica, F., Radon, T., Cheng, Y. & Crnogorac-Jurcevic, T. The life and works of S100P - from conception to cancer. *American journal of cancer research* **6**, 562–576 (2016).
77. Yang, F. *et al*. Tissue kallikrein-related peptidase 4 (KLK4), a novel biomarker in triple-negative breast cancer. *Biological chemistry* **398**, 1151–1164 (2017).
78. Schmitt, M. *et al*. Emerging clinical importance of the cancer biomarkers kallikrein-related peptidases (KLK) in female and male reproductive organ malignancies. *Radiology and oncology* **47**, 319–329 (2013).
79. Pornour, M., Ahangari, G., Hejazi, S. H., Ahmadkhaniha, H. R. & Akbari, M. E. Dopamine receptor gene (DRD1-DRD5) expression changes as stress factors associated with breast cancer. *Asian Pacific journal of cancer prevention: APJCP* **15**, 10339–10343 (2014).
80. Yamamoto, M., Cid, E., Bru, S. & Yamamoto, F. Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. *PLoS One* **6**, e17149 (2011).
81. Tabrizi, M. E. A. *et al*. S100P enhances the motility and invasion of human trophoblast cell lines. *Sci Rep* **8**, 11488 (2018).
82. Papagerakis, P. *et al*. Clinical significance of kallikrein-related peptidase-4 in oral cancer. *Anticancer research* **35**, 1861–1866 (2015).
83. Zhuo, D., Li, X. & Guan, F. Biological Roles of Aberrantly Expressed Glycosphingolipids and Related Enzymes in Human Cancer Development and Progression. *Frontiers in physiology* **9**, 466 (2018).
84. Ouyang, S. *et al*. Bapx1 mediates transforming growth factor-beta- induced epithelial-mesenchymal transition and promotes a malignancy phenotype of gastric cancer cells. *Biochemical and biophysical research communications* **486**, 285–292 (2017).
85. Nixon, A. J. *et al*. Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage I or II breast cancer. *Journal of clinical oncology* **12**, 888–894 (1994).
86. Host, H. & Lund, E. Age as a prognostic factor in breast cancer. *Cancer* **57**, 2217–2221 (1986).
87. Adami, H. O., Malker, B., Holmberg, L., Persson, I. & Stone, B. The relation between survival and age at diagnosis in breast cancer. *The New England journal of medicine* **315**, 559–563 (1986).
88. Bonnier, P. *et al*. Age as a prognostic factor in breast cancer: relationship to pathologic and biologic features. *International journal of cancer* **62**, 138–144 (1995).

89. Anders, C. K. *et al*. Age-specific differences in oncogenic pathway deregulation seen in human breast tumors. *PLoS One* **3**, e1373 (2008).
90. Anders, C. K. *et al*. Breast carcinomas arising at a young age: unique biology or a surrogate for aggressive intrinsic subtypes? *J Clin Oncol* **29**, e18–20 (2011).
91. Anders, C. K., Johnson, R., Litton, J., Phillips, M. & Bleyer, A. Breast Cancer Before Age 40 Years. *Seminars in oncology* **36**, 237–249 (2009).
92. Kumar, N., Patni, P., Agarwal, A., Khan, M. A. & Parashar, N. Prevalence of molecular subtypes of invasive breast cancer: A retrospective study. *Med J Armed Forces India* **71**, 254–258 (2015).
93. Lu, P., Takai, K., Weaver, V. M. & Werb, Z. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harb Perspect Biol* **3** (2011).
94. Amar, S., Smith, L. & Fields, G. B. Matrix metalloproteinase collagenolysis in health and disease. *Biochim Biophys Acta* (2017).
95. Ren, F. *et al*. Overexpression of MMP Family Members Functions as Prognostic Biomarker for Breast Cancer Patients: A Systematic Review and Meta-Analysis. *PLoS One* **10**, e0135544 (2015).
96. Decock, J. *et al*. Matrix metalloproteinase expression patterns in luminal A type breast carcinomas. *Dis Markers* **23**, 189–196 (2007).
97. Xu, J. *et al*. Matrix metalloproteinase expression and molecular interaction network analysis in gastric cancer. *Oncology letters* **12**, 2403–2408 (2016).
98. Tanis, T. *et al*. The role of components of the extracellular matrix and inflammation on oral squamous cell carcinoma metastasis. *Arch Oral Biol* **59**, 1155–1163 (2014).
99. Kohrmann, A., Kammerer, U., Kapp, M., Dietl, J. & Anacker, J. Expression of matrix metalloproteinases (MMPs) in primary human breast cancer and breast cancer cell lines: New findings and review of the literature. *BMC Cancer* **9**, 188 (2009).
100. Zhang, X. *et al*. Insights into the distinct roles of MMP-11 in tumor biology and future therapeutics (Review). *Int J Oncol* **48**, 1783–1793 (2016).
101. Lee, J. Y. *et al*. Gene Expression Profiling of Breast Cancer Brain Metastasis. *Sci Rep* **6**, 28623 (2016).
102. Osako, Y. *et al*. Regulation of MMP13 by antitumor microRNA-375 markedly inhibits cancer cell migration and invasion in esophageal squamous cell carcinoma. *Int J Oncol* **49**, 2255–2264 (2016).
103. Ou, B. *et al*. CCR4 promotes metastasis via ERK/NF-kappaB/MMP13 pathway and acts downstream of TNF-alpha in colorectal cancer. *Oncotarget* **7**, 47637–47649 (2016).
104. Fanjul-Fernandez, M., Folgueras, A. R., Cabrera, S. & Lopez-Otin, C. Matrix metalloproteinases: evolution, gene regulation and functional analysis in mouse models. *Biochim Biophys Acta* **1803**, 3–19 (2010).
105. Jablonska-Trypuc, A., Matejczyk, M. & Rosochacki, S. Matrix metalloproteinases (MMPs), the main extracellular matrix (ECM) enzymes in collagen degradation, as a target for anticancer drugs. *Journal of enzyme inhibition and medicinal chemistry* **31**, 177–183 (2016).
106. Sun, Y., Huang, J. & Yang, Z. The roles of ADAMTS in angiogenesis and cancer. *Tumour Biol* **36**, 4039–4051 (2015).
107. Freitas, V. M. *et al*. Decreased expression of ADAMTS-1 in human breast tumors stimulates migration and invasion. *Mol Cancer* **12**, 2 (2013).
108. Huang, J. *et al*. ADAMTS5 acts as a tumor suppressor by inhibiting migration, invasion and angiogenesis in human gastric cancer. *Gastric cancer: official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association* (2018).
109. Nakopoulou, L. *et al*. Matrix metalloproteinase-1 and -3 in breast cancer: correlation with progesterone receptors and other clinicopathologic features. *Human pathology* **30**, 436–442 (1999).
110. Zhang, B. *et al*. Tumor-derived matrix metalloproteinase-13 (MMP-13) correlates with poor prognoses of invasive breast cancer. *BMC Cancer* **8**, 83 (2008).

Acknowledgements

This work was funded by Department of Biotechnology (DBT), India, authors thank DBT for funding the experimental work. We would also like to thank Sandor Proteomics Pvt. Ltd. and Mr. Nitin for their help in filtration of the initial microarray data analysis. The authors thank the Indian Council of Medical Research (ICMR), Bioinformatics Cell and Dr. A.K. Jain for allowing us to utilize the facility.

Author Contributions

S.M. collected clinical samples and history, performed experiments, analysed data, prepared the figures, and drafted the manuscript. S.A.R.B. conceived the study and performed experiments, analysed data, interpreted the results, and drafted and edited the manuscript. D.P. did the bioinformatic analysis of microarray data. S.S. conceived the study, data analysis, edited the manuscript. C.C., A.B., D.A. and R.S. recruited the patients and provided clinical specimens.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-46261-1.

Competing Interests: The authors declare no competing interests.

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