

# SCIENTIFIC REPORTS

OPEN

## Molecular portraits revealing the heterogeneity of breast tumor subtypes defined using immunohistochemistry markers

Xiaofeng Dai<sup>1,3</sup>, Yang Li<sup>2</sup>, Zhonghu Bai<sup>1,3</sup> & Xu-Qing Tang<sup>2</sup>

Received: 10 May 2015

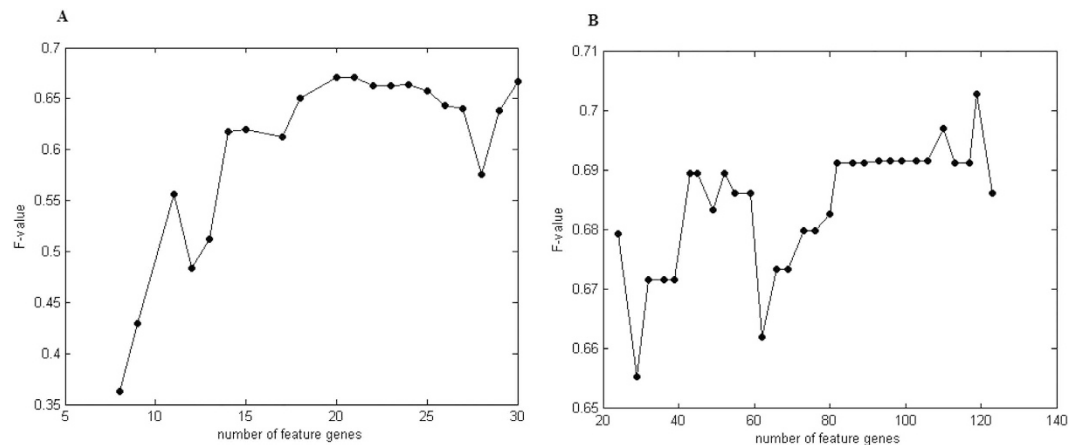
Accepted: 27 August 2015

Published: 25 September 2015

Breast cancer is highly heterogeneous. The subtypes defined using immunohistochemistry markers and gene expression profilings (GEP) are related but not equivalent, with inter-connections under investigated. Our previous study revealed a set of differentially expressed genes (diff-genes), containing 1015 mRNAs and 69 miRNAs, which characterize the immunohistochemistry-defined breast tumor subtypes at the GEP level. However, they may convey redundant information due to the large amount of genes included. By reducing the dimension of the diff-genes, we identified 119 mRNAs and 20 miRNAs best explaining breast tumor heterogeneity with the most succinct number of genes found using hierarchical clustering and nearest-to-center principle. The final signature panel contains 119 mRNAs, whose superiority over diff-genes was replicated in two independent public datasets. The comparison of our signature with two pioneering signatures, the Sorlie's signature and PAM50, suggests a novel marker, FOXA1, in breast cancer classification. Subtype-specific feature genes are reported to characterize each immunohistochemistry-defined subgroup. Pathway and network analysis reveal the critical roles of Notch signalings in [ER+|PR+|HER2-] and cell cycle in [ER+|PR+|HER2+] tumors. Our study reveals the primary differences among the four immunohistochemistry-defined breast tumors at the mRNA and miRNA levels, and proposes a novel signature for breast tumor subtyping given GEP data.

Breast cancer is a heterogeneous disease accompanied by differences in clinical, molecular and biological features<sup>1</sup>, which creates a challenge for prognosis and treatment<sup>2</sup>. Traditionally, immunohistochemistry (IHC) markers together with clinicopathologic indexes are used to classify breast cancer and predict disease outcome<sup>3</sup>. Increasing number of IHC molecules have been identified to play critical roles in breast tumor subtyping, among which estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) are the most commonly used. Based on these molecular markers, breast cancer can be classified into four basic subgroups, i.e., [ER+|PR+|HER2-] (positive ER and PR status, and negative HER2 status), [ER+|PR+|HER2+] (positive ER, PR and HER2 status), [ER-|PR-|HER2+] (negative ER and PR status, and positive HER2 status), [ER-|PR-|HER2-] (negative ER, PR and HER2 status)<sup>4</sup>. Over a decade ago, gene expression profiling (GEP) has been applied to capture breast tumor heterogeneity and subtyping. Sørlie *et al.*<sup>5-7</sup> firstly proposed the usage of "intrinsic" genes to classify breast tumors into four major subtypes, i.e., luminal A, luminal B, HER2 positive, basal-like, and the normal-like tumors in addition. Parker *et al.*<sup>8</sup> developed a classifier composed of 50 genes, namely PAM50, to identify the four major intrinsic subtypes. Each of the four major IHC-defined subtypes corresponds to a basic GEP-defined subgroup. Luminal A and luminal B are roughly equivalent

<sup>1</sup>School of Biotechnology, Jiangnan University, Wuxi 214122, China. <sup>2</sup>School of Science, Jiangnan University, Wuxi 214122, China. <sup>3</sup>National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi 214122, China. Correspondence and requests for materials should be addressed to X.Q.T. (email: txq5139@jiangnan.edu.cn)



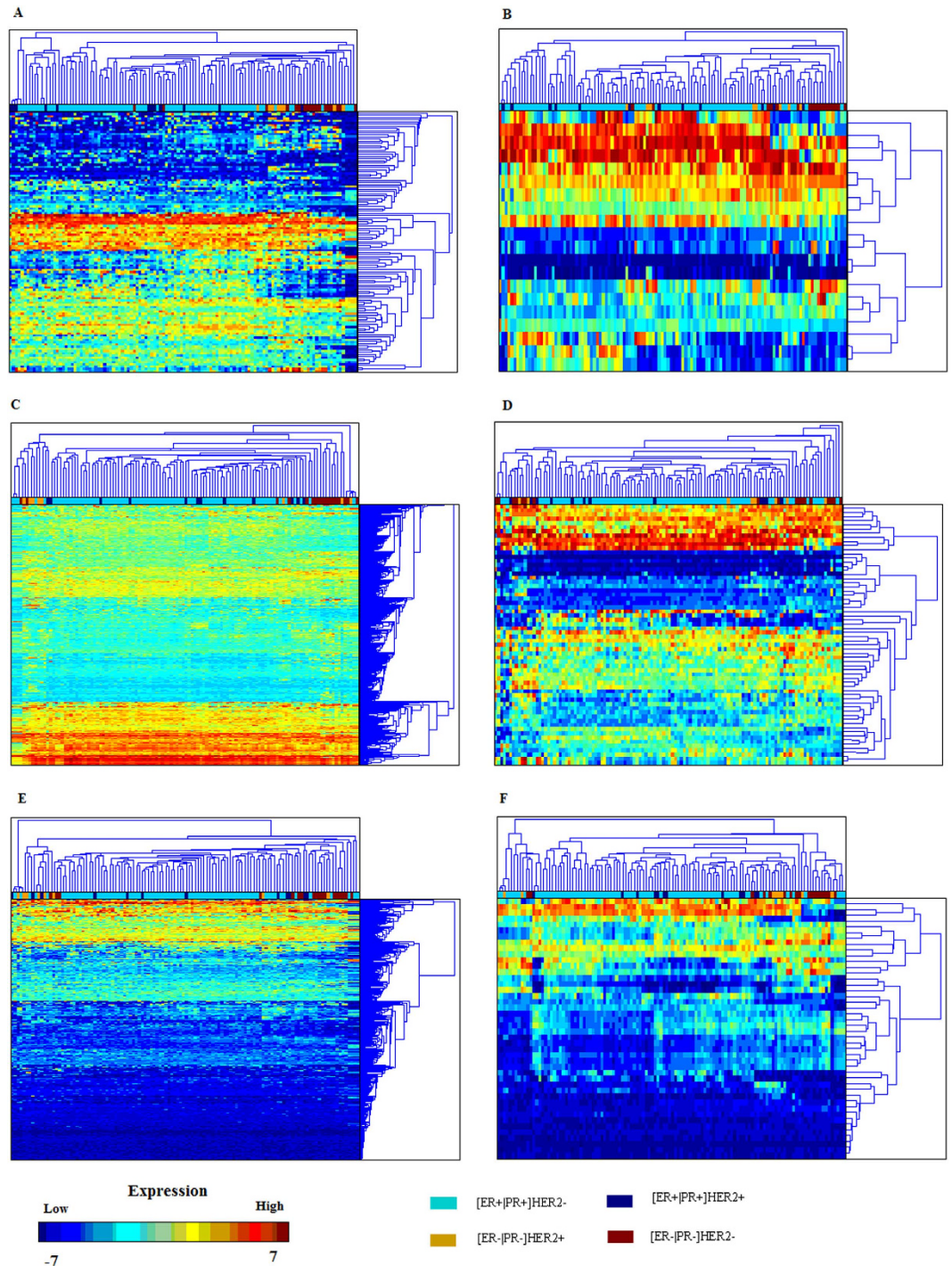
**Figure 1.** Selection of the number of (A) miRNA and (B) mRNA feature genes.

to [ER+|PR+]HER2<sup>-</sup> and [ER+|PR+]HER2<sup>+</sup> tumors, respectively, though a small percentage of [ER+|PR+]HER2<sup>-</sup> tumors with Ki67 positivity are reported to belong to the luminal B subtype<sup>9</sup>. HER2 positive tumors refer to [ER-|PR-]HER2<sup>+</sup> despite the different methods used on HER2 assessment. The [ER-|PR-]HER2<sup>-</sup> (also named triple negative tumors, TNP) subtype is mainly composed of basal-like tumors, which is highly heterogeneous including at least claudin-low<sup>10</sup>, metaplastic breast cancer<sup>11</sup> and interferon-rich tumors<sup>12</sup> in addition to core basal tumors as demonstrated by the accumulated evidence. Dai *et al.*<sup>4</sup> have reported a set of diff-genes, which is composed of 1015 mRNAs and 69 miRNAs that are differentially expressed among the four IHC-defined breast tumor subtypes. Though the diff-genes well capture the differences among these subtypes and could be used for GEP-based subtyping in principle, the large amount of genes included in the gene set may not be feasible for clinical use. Also, only a small subset of genes are, in general, relevant and the useful information may be masked by the other genes that are either redundant or noisy. We are thus inspired to identify the representatives of the diff-genes, aiming at obtaining the best subtyping accuracy with the most succinct number of genes. Subtype-specific feature genes are also revealed to characterize the differences among these IHC-defined subtypes. Network and pathway analysis were conducted to uncover the interconnections and functional roles of these signature genes. Our study reveals the core differences explaining the heterogeneity of the four basic subtypes defined using ER, PR and HER2 status at the mRNA and miRNA expression levels. It bridges the gap between IHC and GEP in differentiating breast tumor subtypes and could be used for subtyping of such tumors given gene expression data.

## Results

**Identification and performance assessment of the signature genes.** The diff-genes presented in<sup>4</sup> are the differentially expressed genes among four breast tumor subtypes defined using ER, PR and HER2 status. It is comprised of 1015 mRNAs and 69 miRNAs, which were reduced to 119 mRNA and 20 miRNA, namely the feature genes (Supplementary Table 1 and Supplementary Table 2), by maximizing the F-values in this paper (Fig. 1). The number of feature genes was determined for each breast tumor subtype (Supplementary Figure 1). Altogether, 13 (out of 379), 19 (out of 65), 16 (out of 152), 18 (out of 777) feature mRNAs, 10 (out of 30), 3, 5, 11 (out of 58) feature miRNAs were selected for [ER+|PR+]HER2<sup>-</sup>, [ER+|PR+]HER2<sup>+</sup>, [ER-|PR-]HER2<sup>+</sup>, [ER-|PR-]HER2<sup>-</sup> tumors, respectively (Supplementary Table 3).

The performance of these feature genes in subtyping tumor samples was compared with the original diff-genes<sup>4</sup> using the HEBCS dataset (Fig. 2). The clustering accuracies (measured by F-value and Rand index) were summarized in Table 1, with the patterns displayed in Fig. 2. The F-value and Rand-index using the mRNA feature genes are higher than the corresponding diff-genes in the HEBCS dataset (Table 1: F-value 0.7029 vs. 0.6599; Rand-index 0.7272 vs. 0.6577, Fig. 2A vs. Fig. 2C), and so as to miRNAs (Table 1: F-value 0.6712 vs. 0.5682; Rand-index 0.6898 vs. 0.5, Fig. 2B vs. Fig. 2D). The mRNA feature genes have higher accuracy than that of the miRNA feature genes (Table 1: F-value 0.7029 vs. 0.6712; Rand-index 0.7272 vs. 0.6898, Fig. 2A vs. Fig. 2B). The performance of the unified mRNA and miRNA feature genes ('the signature') has the same F-value and Rand-index as the mRNA feature genes. We, thus, include only mRNAs in the signature to make it as concise as possible. The performance improvement of the signature over mRNA diff-genes was replicated using GSE22220 (Table 1: F-value 0.8449 vs. 0.7084; Rand-index 0.7454 vs. 0.6175, Supplementary Figure 2A vs. Supplementary Figure 2B). The performance of the signature was compared with the Sorlie's signature<sup>5</sup>, the first widely accepted gene list differentiating breast tumor subtypes (Table 1: F-value 0.7029 vs. 0.63; Rand-index 0.7272 vs. 0.5981, Fig. 2A vs. Fig. 2E), and PAM50, the most well-known gene panel for GEP subtyping (Table 1: F-value 0.7029 vs. 0.618; Rand-index 0.7272 vs. 0.6003, Fig. 2A vs. Fig. 2F), using HEBCS. Such performance



**Figure 2.** Hierarchical clustering of breast tumor samples in HEBCS using (A) mRNA feature genes (the signature), (B) miRNA feature genes, (C) mRNA diff-genes, (D) miRNA diff-genes, (E) Sorlie's signature, (F) PAM50 genes. Molecules shown in red (increased expression) and green (decreased expression) identified in different sets of genes.

superiority over the Sorlie's signature (Table 1: F-value 0.8449 vs. 0.683; Rand-index 0.7454 vs. 0.5305, Supplementary Figure 2A vs. Supplementary Figure 2C) and PAM50 (Table 1: F-value 0.8449 vs. 0.7316; Rand-index 0.7454 vs. 0.6364, Supplementary Figure 2A vs. Supplementary Figure 2D) was also observed for the signature using GSE22220. Similar clustering accuracies were obtained for the signature genes, the Sorlie's signature and PAM50 using the TCGA dataset (Table 1, Supplementary Figure 3).

The feature mRNAs and miRNAs (feature genes identified by using the diff-genes of all subtypes) were compared with the unified subtype-specific feature mRNAs and miRNAs (unified genes containing the feature genes identified from the diff-genes of each subtype). Out of the 119 mRNA feature genes and

Dataset	Gene	Dimension	F-value	Rand-index	Purpose
HEBCS	miRNA diff-genes	69	0.5682	0.5	Identification
	miRNA feature genes	20	0.6712	0.6898	Identification
	mRNA diff-genes	1015	0.6599	0.6577	Identification
	mRNA feature genes (the signature)	119	0.7029	0.7272	Identification
	The unified subtype-specific genes	139	0.7029	0.7272	Identification
	Sorlie's signature	456	0.63	0.5981	Comparison
	PAM50	50	0.618	0.6003	Comparison
GSE22220	mRNA diff-genes	1015	0.7084	0.6175	Validation
	mRNA feature genes (the signature)	119	0.8449	0.7454	Validation
	Sorlie's signature	456	0.683	0.5305	Comparison
	PAM50	50	0.7316	0.6364	Comparison
TCGA	mRNA diff-genes	1015	0.7225	0.7044	Validation
	mRNA feature genes (the signature)	119	0.7237	0.7032	Validation
	Sorlie's signature	456	0.7189	0.7028	Comparison
	PAM50	50	0.7304	0.7068	Comparison

**Table 1. Comparison of clustering accuracy between feature genes, diff-genes and well-known signature genes using different datasets.**

mRNA			miRNA		
<i>ALCAM</i>	<i>CAMK2N1</i>	<i>EFHD1</i>	<i>HS_239</i>	<i>hsa-miR-130b*</i>	<i>hsa-miR-135b</i>
<i>SPARCL1</i>	<i>DCTN4</i>	<i>GRP</i>	<i>hsa-miR-101*</i>	<i>hsa-miR-33b</i>	<i>hsa-miR-135a</i>
<i>C19orf33</i>	<i>DHRS2</i>		<i>hsa-miR-184</i>	<i>hsa-miR-521</i>	<i>hsa-miR-411</i>

**Table 2. Overlapping genes between the feature genes and the unified subtype-specific genes.**

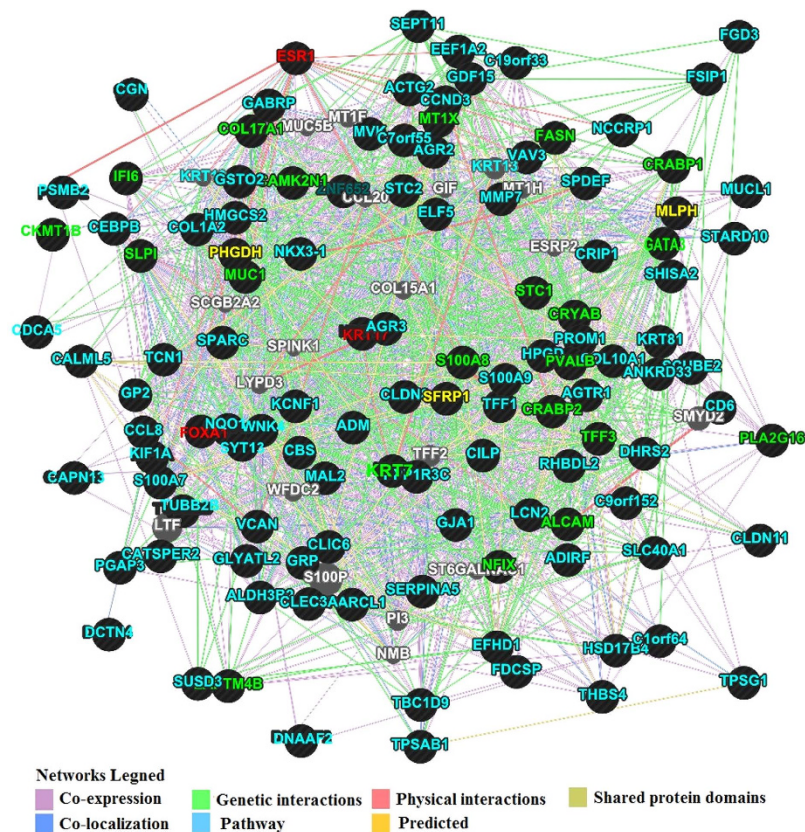
62 unified subtype-specific feature mRNAs, 8 overlapped; and out of the 20 miRNA feature genes and 25 unified subtype-specific feature miRNAs, 9 overlapped. These overlapping genes, as listed in Table 2, might be the key molecules differentiating breast tumor subtypes.

**Pathway and disease analysis of the signature genes.** Several cancer core pathways were found enriched in the signature genes, miRNA targets, subtype-specific feature genes or their union (Supplementary Table 4). In particular, cell adhesion molecules including VCAN, ALCAM, CLDN11, CLDN8, and CD6, were enriched in the signature genes ( $p = 0.004$ ). The unified subtype-specific genes were present in the p53 pathway ( $p = 0.024$ ). The targets of the miRNA feature genes were mostly involved in cell cycle ( $p = 0.03$ ), mTOR ( $p = 0.043$ ) and VEGF ( $p = 0.044$ ) signalings. Among the four IHC-defined subtypes, genes of [ER+|PR+]HER2+ were enriched in DNA replication ( $p = 0.026$ ), Notch signaling ( $p = 0.034$ ) and the TGF $\beta$  pathway ( $p = 0.056$ ). We also checked the diseases relevant to the signature genes and feature genes of each subtype, with various cancers significantly enriched especially for the [ER-|PR-]HER2+ subtype (Supplementary Table 5).

Genemania was used to study the networks of the signature genes (Fig. 3), and the subtype-specific feature genes (Supplementary Figure 4). The sum of different links of the subtype-specific feature genes and the signature genes were summarized in Table 3. The feature genes of [ER+|PR+]HER2- tumors are involved in many known pathways and harbor many physical interactions. Those of [ER+|PR+]HER2+ tumors have the most shared protein domains, The feature genes of [ER-|PR-]HER2+ tumors are enriched by co-expressed genes, and [ER-|PR-]HER2- specific genes have the most co-localized genes among others. Genetic interaction is equally common among the feature genes of [ER+|PR+]HER2-, [ER+|PR+]HER2+ and [ER-|PR-]HER2- tumors, except for the [ER-|PR-]HER2+ subtype where the genetic interaction is rare. Co-expression is the most common among other interactions in the signature genes.

These signature genes are densely connected, among which several, such as ESR1, FOXA1, NQO1, GATA1, ALDH3B2, keratins, are well-known players driving the heterogeneity and carcinogenesis of breast tumors.





**Figure 3.** Network of the signature genes constructed using GeneMANIA. Red colored genes are shared between the signature, the Sorlie's signature and PAM50 genes; yellow colored genes are shared between the signature and the Sorlie's signature; green colored genes are shared between the signature and PAM50 genes.

Links	[ER+ PR+] HER2–	[ER+ PR+] HER2+	[ER– PR–] HER2+	[ER– PR–] HER2–	The signature
Co-expression	54	40	86	9	1141
Co-localization		8		44	133
Genetic interaction	20	31		26	310
Pathway	48				8
Physical interaction	47	6			25
Shared-protein domain		95	7		66
Total links	185	180	93	79	1694

**Table 3.** Link properties in the network of signature genes and subtype-specific feature genes.

## Discussion

The mRNA and miRNA feature genes perform better than the original diff-genes reported in<sup>4</sup> in differentiating the four IHC-defined tumor subtypes using HEBCS (Table 1: F value 0.7029 vs. 0.6599 for mRNA, 0.6712 vs. 0.5682 for miRNA; Rand Index  $p = 0.7272$  vs. 0.6577 for mRNA,  $p = 0.6898$  vs. 0.5 for miRNA), indicating that irrelevant genes have been efficiently removed from the signature which add little information but noise. MiRNAs perform less accurately than mRNAs, and do not contribute additional information to the signature on top of mRNAs. This, on one hand, may be caused by the complex and indirect influences of miRNAs on the phenotypic differences among breast tumor subtypes and, on the other hand, suggests the same pathways involved by the feature miRNAs and mRNAs (i.e., the targets of miRNAs share the same signaling with mRNAs). Actually, none of the validated miRNA targets overlaps with the signature mRNA genes, and so as to their enriched pathways which were retrieved from “KEGG Mapper—Search&Color Pathway” with the default parameter setting (Supplementary Table 4). KEGG database collects manually drawn pathway maps representing our current knowledge on molecular interactions and reaction networks. These seemingly inconsistent results imply that these genes,

though being different and annotated to different KEGG pathways, may be involved in the same or alternative signaling with novel functional roles to be discovered.

The presented signature outperforms the Sorlie's signature<sup>6</sup> (which pioneers the field using gene expression profiling for breast tumor subtyping, Fig. 2E) and PAM50 genes (which is commonly applied for GEP-based breast tumor subtyping, Fig. 2F), with increased accuracy and moderate number of genes included as tested using HEBCS data (Table 1). The superiority of the signature over the diff-genes, Sorlie's signature and PAM50 was replicated using GSE22220, demonstrating the generality and correctness of our observations. However, no significant difference regarding the classification accuracy was observed among the signature, diff-genes, the Sorlie's signature and PAM50 using TCGA. This indicates that GEP-based clustering accuracy, though dominated by the genes included in the signature, is affected by the gene expression levels assessed, and the performance of the signature is at least as good as the Sorlie's signature and PAM50.

Among the genes included in the signature, 25 and 6 are in common with the Sorlie's signature and PAM50 genes, respectively, among which 3 are shared among all three datasets (Supplementary Table 6, Supplementary Figure 5). Several overlapping genes especially the ones present in all datasets (ESR1, FOXA1, KRT17) are known to play critical roles in the subtyping and carcinogenesis of breast tumors. For example, ESR1 is a discriminative factor between ER positive and ER negative tumors that mediates the biological effects of estrogens through direct binding to the estrogen response elements (EREs) of the target genes<sup>13</sup>; FOXA1 is associated with the methylation of the promoter of tumor suppressor genes and thus suggested as a potential demethylation target for the prevention and treatment of breast cancer<sup>14</sup>; cytokeratins such as KRT17 and KRT7 are basal markers and known to be up-regulated in circulating tumor cells<sup>15</sup>; and GATA3 is a transcriptional activator highly expressed in the luminal epithelial cells of the breast and lowly expressed in invasive carcinomas<sup>16</sup>, whose low expression is associated with ER negativity, PR positivity and HER2 over-expression<sup>17</sup>. Among the three genes shared by all signatures, two have already been applied for tumor subtyping, i.e., ESR1 is the primary marker classifying breast tumors into ER positive and ER negative subgroups and KRT17 plays crucial roles in differentiating the basal-like subtype from the other triple negative tumors, indicating that FOXA1 may be a novel immunohistochemistry marker for breast tumor classification.

The feature genes, selected from the unified diff-genes, have 8 mRNAs (6.7% of mRNA feature genes) and 9 miRNAs (45% of miRNA feature genes) overlapping with the unified subtype-specific feature genes (Table 2). The relatively small percentage of genes selected using both methods suggests the high heterogeneity of breast tumors and that these overlapping ones may play the key roles in distinguishing breast tumor subtypes. Most of these overlapping mRNA genes and miRNA targets are known to play critical roles in cancers or tumor cell lines. For example, ALCAM<sup>18–20</sup> is associated with breast cancer migration and progression; GRP<sup>21,22</sup> has mitogenic effects on some human breast cancer cell lines; SPARCL1<sup>23,24</sup> is relevant to aggressive and invasive tumors and drives disease recurrence of prostate cancers; DHRS2<sup>25</sup> encodes for Hep27 that is part of the molecular pathway regulating cell cycle and apoptosis in osteosarcoma and MCF7 breast cancer cells; CAMK2N1 plays a tumor suppressive role in prostate cancer and is suggested as a biomarker and therapeutic target of such tumors<sup>26</sup>. Has-miR-33b is known to target genes involved in cancer pathways such as MAPK, Wnt and Nf-kB signalings<sup>27</sup>. A direct target of has-miR-184, SND1, is suggested as a therapeutic target for malignant glioma<sup>28</sup>. Has-miR-135a/b modulate apoptosis via targeting MCL1 in lung cancer cell lines<sup>29</sup>. Interestingly, hsa-miR-135a and hsa-miR-135b share the same set of mRNA targets and play crucial roles in distinguishing breast tumors by ER positivity<sup>4</sup>, suggesting their non-redundant roles in distinguishing ER positive and ER negative breast tumors. Furthermore, hsa-miR-135b is characteristic of [ER–|PR–]HER2– tumors while has-miR-135a symbolizes the [ER+|PR+]HER2+ subtype in addition to [ER–|PR–]HER2–, implying an underlying connection between [ER+|PR+]HER2+ and [ER–|PR–]HER2– tumors, which are both aggressive.

The networks of the subtype-specific feature genes reveal the hub components representing each of these IHC-defined subtypes. NOTCH1, a key component present in [ER+|PR+]HER2– tumors, symbolizes the importance of Notch signaling in such cancers, which is an evolutionarily conserved mechanism that mediates communications between cells<sup>30</sup>. CDKN2A could induce cell cycle arrest in G1 and G2 phases<sup>31</sup>, whose presence in [ER+|PR+]HER2+ tumors suggests the representative roles of cell cycle signaling on tumors of this subtype.

## Conclusion

By reducing the dimensionality of the differentially expressed genes among IHC-defined subtypes presented in<sup>4</sup>, we report a 119-gene signature that captures the characteristics of these subtypes with improved accuracy and reduced number of genes. The feature genes of each subtype, including both mRNAs and miRNAs, are also presented, which explain the heterogeneity of the four basic IHC-defined subtypes. Comparison of our signature with the Sorlie's signature and PAM50 suggests the crucial roles played by FOXA1 in breast cancer classification. Network analysis reveals the critical roles of Notch signaling in [ER+|PR+]HER2– and cell cycle in [ER+|PR+]HER2+. We present a set of signature genes rather than a tumor subtyping tool here, which better captures the differences among breast cancer subtypes than the genes included in the Sorlie's signature and PAM50. It could be made available for breast tumor subtyping by relating a given sample to the centroid of each subtype determined using the expression of the signature genes from the training data, which would be our next step. As a reduced gene set

of the diff-genes from<sup>4</sup>, the signature inherits the advantages of diff-genes. It bridges the gap between immunohistochemistry markers and gene expression profiling in breast tumor subtyping in addition to its integration of information at mRNA and miRNA levels. On top of that, the signature improves the subtyping accuracy and reduces the experimental cost, which better explains the heterogeneity of breast cancer and avails in the diagnosis of breast cancer patients as compared with the diff-genes reported in<sup>4</sup>.

## Material and Method

**Materials.** The three public data sets employed in<sup>4</sup> for diff-gene discovery, i.e., HEBCS, GSE22220, and TCGA were used in this study to identify and validate the signature genes.

HEBCS is composed of the mRNA (GSE24450) and miRNA (GSE43040) data from the GEO database<sup>32</sup>. This dataset harbors 24660 mRNAs (Illumina HumanHT-12\_V3 Expression BeadChips) and 1104 miRNAs (IlluminaHumanMI\_V2 BeadChips) for 183 primary breast tumor samples from the department of Oncology of the Helsinki University Central Hospital (HUCH) and department of Surgery<sup>4,33</sup>. The samples were grouped into four subtypes, i.e., [ER+|PR+]HER2-, [ER+|PR+]HER2+, [ER-|PR-]HER2+ and [ER-|PR-]HER2-, based on the status of ER, PR and HER2<sup>4</sup>. 1015 mRNAs and 69 miRNAs were identified differentially expressed among the four IHC defined subgroups.

GSE22220 consists of mRNA (GSE22219) and miRNA (GSE22216) data from GEO<sup>32</sup>. GSE22219 contains 24332 probes (Illumina Human Ref-8\_V1 expression Bead Chips) for 216 patients, and GSE22216 contains 734 probes (Illumina HumanMI\_V1 BeadChips) for 207 samples. These samples were grouped into ER+ and ER- tumors in<sup>4</sup> based on its available IHC information.

TCGA data (level 3) was retrieved from the TCGA portal at <http://tcga.cancer.gov/dataportal>, which contains 17814 mRNAs (Agilent 244K Custom Gene Expression G4502A-07-3) for 451 samples and 1046 miRNAs (IlluminaGA\_miRNASeq) for 315 patients<sup>4</sup>. These primary solid tumor samples were classified into the four IHC-characterized subtypes as defined in the HEBCS data.

All datasets were pre-processed following instructions in<sup>4</sup>.

## Methods

**Hierarchical Clustering and accuracy assessment.** Hierarchical clustering (HC) was applied to identify samples sharing similar expression levels according to a given set of genes. In the iterative process of HC, each sample is a point in a  $|G|$  dimensional space, and all samples are clustered based on a certain similarity measure and the distance of these genes as measured according to their expression levels. The average linkage clustering algorithm was employed due to its efficiency in analyzing differential expression among samples.

Two well-known external evaluation indexes, i.e., Rand index and F-value<sup>34</sup>, were applied to assess the clustering accuracy provided with the knowledge on the ground-truth of the data structure.

Rand index considers the relationship between pairwise samples. Define the original and clustered set are  $U$  and  $V$ , respectively, there are four situations considering sample pair analysis, i.e.,

$$\begin{aligned} a &= |\{i, j\} | C_U(i) = C_U(j) \wedge C_V(i) = C_V(j) | \\ b &= |\{i, j\} | C_U(i) = C_U(j) \wedge C_V(i) \neq C_V(j) | \\ c &= |\{i, j\} | C_U(i) \neq C_U(j) \wedge C_V(i) = C_V(j) | \\ d &= |\{i, j\} | C_U(i) \neq C_U(j) \wedge C_V(i) \neq C_V(j) | \end{aligned} \quad (1)$$

Based on the consistency and deviation, the Rand index is defined as

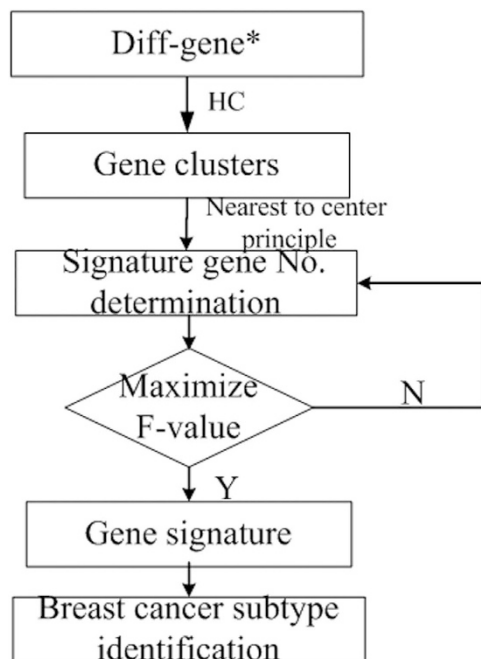
$$R(U, V) = \frac{a + d}{a + b + c + d}, \quad R(U, V) \in [0, 1]. \quad (2)$$

F-measure applies the concept of 'precision' and 'recall' from information retrieval here. They are defined as  $P_i = \text{Prec}(N_{ij}, U_i) = \frac{|N_{ij}|}{|U_i|}$ ,  $R_j = \text{Rec}(N_{ij}, V_j) = \frac{|N_{ij}|}{|V_j|}$ , where  $N_{ij}$  is the intersection set between classes  $U_i$  and  $V_j$ , and  $| * |$  (\* represents  $U_i$ ,  $V_j$  and  $N_{ij}$ , respectively) denotes the number of the elements in each of these sets. F-value is determined by

$$F(U_i) = \frac{2P_iR_j}{P_i + R_j}, \quad (3)$$

where  $F(U_i) \in [0, 1]$ .

**Signature gene identification.** It is assumed that samples sharing similar expression profiles of the genes characterizing their heterogeneity (namely the signature genes) are likely to form a subtype that is phenotypically distinct from the other samples. To remove the irrelevant genes masking the roles of the signature genes in differentiating breast tumor subtypes and make the gene panel as succinct as possible, two steps were applied to the diff-genes<sup>4</sup> which were differentially expressed among IHC-defined subgroups.



**Figure 4.** The process for signature gene identification. Genes marked by \*are taken from<sup>4</sup>.

First, determine the number of feature genes ( $N(C_r)$ ) for each cluster by the cluster cohesiveness (which measures the closeness of a cluster). The cohesiveness of a class is given by

$$Co(C_r) = \left( \frac{2 \cdot \sum_i \sum_{j>i} d_{ij}}{|C_r| \cdot (|C_r| - 1)} \right)^{-1} \quad i, j \leq n, \quad (4)$$

where  $|C_r|$  denotes the number of samples in class  $C_r$ , and  $d_{ij}$  denotes the distance between  $C_i^r$  and  $C_j^r$  where the Euclidean distance is applied. The cohesiveness index reflects the similarity of the gene expression profiles within a class with a positive correlation. The number  $N(C_r)$  of feature genes in class  $C_r$  is determined by maximizing the F-value according to

$$N(C_r) = \frac{K}{Co(C_r)} \quad 1 \leq r \leq n, \quad (5)$$

where  $K(K > 0)$  is the cohesive strength, and  $\min(K) = \max\{Co(C_r)\}$  as each group has at least one signature gene.  $N(C_r)$  and  $Co(C_r)$  are negatively correlated as the more diverse a gene cluster is the more genes are needed to characterize<sup>34</sup>.

Second, select the signature genes for each cluster based on the nearest-to-center principle. Genes in class  $C_r$  could be divided into  $N(C_r)$  subclasses using HC. The center  $Cen(C_{ri})$  of the subclass  $C_{ri}$  can be given by

$$Cen(C_{ri}) = \frac{\sum_j C_j^{ri}}{|C_{ri}|}, \quad (6)$$

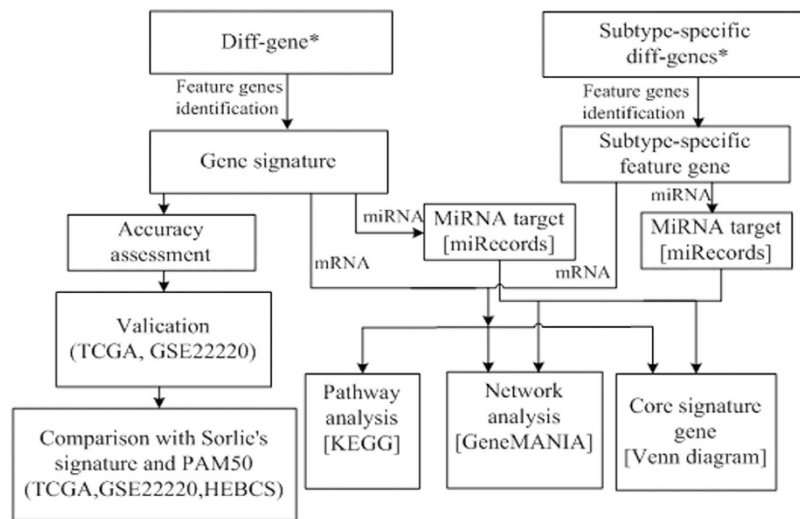
where  $C_j^{ri}$  means that gene  $j$  belongs to the subclass  $C_{ri}$  of class  $C_r$ . One feature gene is selected from each subclass by

$$del(C_{ri}) = \left\{ C_j^{ri} \mid \min\{C_j^{ri} - Cen(C_{ri})\} \right\}, \quad (7)$$

where  $del(C_{ri})$  is the agent selected from subclass  $C_{ri}$  that is closest to the fictitious center  $Cen(C_{ri})$ . The representative genes for each subtype or a particular type of data are called the feature genes, and the final gene panel selected for characterizing breast tumor heterogeneity and subtyping is named 'the signature'.

The process for signature gene identification is illustrated in Fig. 4.





**Figure 5. Work flow of the whole process deciphering the heterogeneity of breast tumor subtypes.** Genes marked by \*are taken from<sup>4</sup>. Tools used in the analysis are shown in the square brackets, and the datasets used are listed in the round brackets.

**Network and pathway analysis using gene signature.** To investigate the intrinsic heterogeneity of breast cancer, metabolic pathway and network analysis were applied to the obtained signature genes. MiRecords<sup>35</sup>, a resource integrating experimentally validated miRNA targets having systematic documentation of experimental support and predicted miRNA targets produced by 11 established prediction algorithms (predicted algorithms = 4), was used to find the targets of the feature miRNAs. DAVID<sup>36</sup> (similarity term overlap = 4; similarity threshold = 0.85; group members = 3; multiple linkage threshold = 0.5 and EASE = 1) and KOBAS<sup>37</sup> (statistical method is hypergeometric test/Fisher's exact test; FDR correction method is Benjamini and Hochberg; small term cutoff default = 5) were used to interpret the enrichment of gene ontology, metabolic pathway and relevant disease of these feature mRNAs and miRNA targets. The gene network was constructed using GeneMANIA<sup>38</sup> (co-expression, co-localization, genetic interactions, pathway, physical interactions, predicted and shared protein domains were selected; automatically selected weighting method was used) to further elucidate the functional roles of the feature genes and the characteristics of each subtype. The whole process for identifying the signature genes and deciphering the heterogeneity of breast cancer subtypes is illustrated in Fig. 5.

## References

- Simpson, P. T., Reis-Filho, J. S., Gale, T. & Lakhani, S. R. Molecular evolution of breast cancer. *PATH* **205**, 248–254, doi: 10.1002/path.1691 (2005).
- Dunnwald, L. K., Rossing, M. A. & Li, C. I. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *BCR* **9**, R6, doi: 10.1186/bcr1639 (2007).
- Blows, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLOS MED* **7**, e1000279, doi: 10.1371/journal.pmed.1000279 (2010).
- Dai, X., Chen, A. & Bai, Z. Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling. *Sci. Rep* **4**, 6566, doi: 10.1038/srep06566 (2014).
- Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS* **98**, 10869–10874, doi: 10.1073/pnas.191367098 (2001).
- Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS* **100**, 8418–8423, doi: 10.1073/pnas.0932692100 (2003).
- Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *JCO* **27**, 1160–1167, doi: 10.1200/JCO.2008.18.1370 (2009).
- Cheang, M. C. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *JNCI* **101**, 736–750, doi: 10.1093/jnci/djp082 (2009).
- Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *BCR* **12**, R68, doi: 10.1186/bcr2635 (2010).
- Gibson, G. R., Qian, D., Ku, J. K. & Lai, L. L. Metaplastic breast cancer: clinical features and outcomes. *The American surgeon* **71**, 725–730 (2005).
- Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics* **7**, 96, doi: 10.1186/1471-2164-7-96 (2006).
- Buffa, F. M. *et al.* microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *CR* **71**, 5635–5645, doi: 10.1158/0008-5472.CAN-11-0489 (2011).
- Zheng, L. *et al.* FOXA1 positively regulates gene expression by changing gene methylation status in human breast cancer MCF-7 cells. *Int J Clin Exp Pathol* **8**, 96–106 (2015).
- Lu, J. *et al.* Isolation of circulating epithelial and tumor progenitor cells with an invasive phenotype from breast cancer patients. *IJC* **126**, 669–683, doi: 10.1002/ijc.24814 (2010).

16. Lacroix, M. & Leclercq, G. About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. *Mol Cell Endocrinol* **219**, 1–7, doi: 10.1016/j.mce.2004.02.021 (2004).
17. Mehra, R. *et al.* Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *CR* **65**, 11259–11264, doi: 10.1158/0008-5472.CAN-05-2495 (2005).
18. Jezierski, A., Matysiak, W. & Moty, T. ALCAM/CD166 protects breast cancer cells against apoptosis and autophagy. *Med Sci Monit* **12**, BR263–BR273 (2006).
19. Kulasingham, V. *et al.* Activated leukocyte cell adhesion molecule: a novel biomarker for breast cancer. *IJC* **125**, 9–14, doi: 10.1002/ijc.24292 (2009).
20. Ihnen, M. *et al.* Predictive impact of activated leukocyte cell adhesion molecule (ALCAM/CD166) in breast cancer. *BREAST CANCER RES TR* **112**, 419–427, doi: 10.1007/s10549-007-9879-y (2008).
21. Halmos, G., James, L., Wittliff & Schally, A. V. Characterization of Bombesin/Gastrin-releasing Peptide Receptors in Human Breast Cancer and Their Relationship to Steroid Receptor Expression. *CR* **55**, 280–228 (1995).
22. Van de Wiele, C. *et al.* Gastrin-releasing peptide receptor imaging in human breast carcinoma versus immunohistochemistry. *JNM* **49**, 260–264, doi: 10.2967/jnumed.107.047167 (2008).
23. Esposito, I. *et al.* Tumor-Suppressor Function of SPARC-Like Protein 1/Hevin in Pancreatic Cancer. *NEOPLASIA* **9**, 8–17, doi: 10.1593/neo.06646 (2007).
24. Hurley, P. J. *et al.* Secreted protein, acidic and rich in cysteine-like 1 (SPARCL1) is down regulated in aggressive prostate cancers and is prognostic for poor clinical outcome. *PNAS* **109**, 14977–14982, doi: 10.1073/pnas.1203525109 (2012).
25. Gabrielli, F. & Tofanelli, S. Molecular and functional evolution of human DHRS2 and DHRS4 duplicated genes. *Gene* **511**, 461–469, doi: 10.1016/j.gene.2012.09.013 (2012).
26. Wang, T. *et al.* The tumor suppressive role of CAMK2N1 in castration-resistant prostate cancer. *ONCOTARGET* **5**, 3611–3621 (2014).
27. Cui, F. M. *et al.* Radon-induced alterations in micro-RNA expression profiles in transformed BEAS2B cells. *J TOXICOL ENV HEAL A* **76**, 107–119, doi: 10.1080/15287394.2013.738176 (2013).
28. Emdad, L. *et al.* Suppression of miR-184 in malignant gliomas upregulates SND1 and promotes tumor aggressiveness. *Neuro Oncol*, doi: 10.1093/neuonc/nou220 (2014).
29. Zhou, L. *et al.* miR-135a/b modulate cisplatin resistance of human lung cancer cell line by targeting MCL1. *POR* **19**, 677–683, doi: 10.1007/s12253-013-9630-4 (2013).
30. Stylianou, S., Clarke, R. B. & Brennan, K. Aberrant activation of notch signaling in human breast cancer. *CR* **66**, 1517–1525, doi: 10.1158/0008-5472.CAN-05-3054 (2006).
31. Dębniak, T. *et al.* Germline mutation and large deletion analysis of the CDKN2A and ARF genes in families with multiple melanoma or an aggregation of malignant melanoma and breast cancer. *IJC* **110**, 558–562, doi: 10.1002/ijc.20163 (2004).
32. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *NAR* **30**, 207–210 (2002).
33. Fagerholm, R. *et al.* NAD (P) H: quinone oxidoreductase 1 NQO1\* 2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer. *Nat Genet* **40**, 844–853 (2008).
34. Halkidi, M. & Vazirgiannis, M. *Clustering validity assessment using multi representatives. Proceedings of SETN conference, Thessaloniki, Greece.* New York: Springer-Verlag Telos (2002).
35. Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *NAR* **37**, D105–D110, doi: 10.1093/nar/gkn851 (2009).
36. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *NAR* **37**, 1–13, doi: 10.1093/nar/gkn923 (2009).
37. Xie, C. *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *NAR* **39**, W316–W322, doi: 10.1093/nar/gkr483 (2011).
38. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *NAR* **38**, W214–W220, doi: 10.1093/nar/gkq537 (2010).

## Acknowledgments

This work was supported by two projects from the National Natural Science Foundation of China (Grand No. 31471251, Grand No. 11371174), and the Jiangnan University Research Support Foundation for Young Scientists (Grand No. 5922050205150370). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Contributions

X.D. and X.Q.T. designed the study. X.D. and Y.L. implemented the analysis and prepared the draft. Z.B. contributed in finalizing the manuscript. All authors have read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Dai, X. *et al.* Molecular portraits revealing the heterogeneity of breast tumor subtypes defined using immunohistochemistry markers. *Sci. Rep.* **5**, 14499; doi: 10.1038/srep14499 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>