Gene expression profiles of inflammatory breast cancer reveal high heterogeneity across the epithelial-hybrid-mesenchymal spectrum

Running title: EMT heterogeneity in IBC

Priyanka Chakraborty[1], Jason T George[2,3], Wendy A Woodward[4,5], Herbert Levine[2,6], Mohit Kumar Jolly[1,*]

[1] Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India
[2] Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA
[3] Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77005, USA
[4] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[5] MD Anderson Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[6] Departments of Physics and Bioengineering, Northeastern University, Boston, MA 02115, USA

*Corresponding author: mkjolly@iisc.ac.in (M.K.J.)

Keywords: IBC, Gene expression signature, tumor heterogeneity, hybrid epithelial/ mesenchymal

Abstract

Inflammatory breast cancer (IBC) is a highly aggressive breast cancer that metastasizes largely via tumor emboli, and has a 5-year survival rate of less than 30%. No unique genomic signature has yet been identified for IBC nor has any specific molecular therapeutic been developed to manage the disease. Thus, identifying gene expression signatures specific to IBC remains crucial. Here, we compare various gene lists that have been proposed as molecular footprints of IBC using different clinical samples as training and validation sets and using independent training algorithms, and determine their accuracy in identifying IBC samples in three independent datasets. We show that these gene lists have little to no mutual overlap, and have limited predictive accuracy in identifying IBC samples. Despite this inconsistency, single-sample gene set enrichment analysis (ssGSEA) of IBC samples correlate with their position on the epithelial-hybrid-mesenchymal spectrum. This positioning, together with ssGSEA scores, improves the accuracy of IBC identification across the three independent datasets. Finally, we observed that IBC samples robustly displayed a higher coefficient of variation in terms of EMT scores, as compared to non-IBC samples. Pending verification that this patient-to-patient variability extends to intratumor heterogeneity within a single patient, these results suggest that higher heterogeneity along the epithelial-hybrid-mesenchymal spectrum can be regarded to be a hallmark of IBC and a possibly useful biomarker.

Introduction

Inflammatory breast cancer (IBC) is a rare (2-4% of breast cancer cases) but highly aggressive, locally advanced breast cancer with extremely poor prognosis and a 5-year survival rate of less than 30% [1]. At diagnosis, most IBC patients exhibit signs of lymph node metastasis; approximately 30% of patients have distant metastases as compared to 5% of patients in non-IBC breast cancers [2]. IBCs are often highly angiogenic and invasive, of high histological grade, and cause 7-10% of all breast cancer-associated deaths [1]. Histologically, IBC cells often do not present as a dominant mass, rather being diffused in clusters throughout the breast and skin, thereby leading to many false-negative imaging findings [3]. Decoding unique molecular underpinnings of this deadly disease remains an unmet clinical need.

The presence of tumor emboli in dermal-lymphatic vessels is a pathological hallmark of IBC. Consistently, IBC patients have a higher frequency and larger average size of clusters of circulating tumor cells (CTCs) as compared to non-IBC patients [4]. These clusters have a strong association with poor survival [4], reminiscent of the disproportionately high metastatic fitness of CTC clusters [5]. Aside from this major difference, no unique genomic signature has yet been conclusively identified for IBC, suggesting that other factors may be more important than genetic events for IBC: phenotypic and/or epigenetic heterogeneity, interaction of malignant cells within emboli or with cells from the tumor microenvironment [1,3]. Because of these uncertainties, no specific molecular therapeutic approaches have been yet proposed to manage IBC.

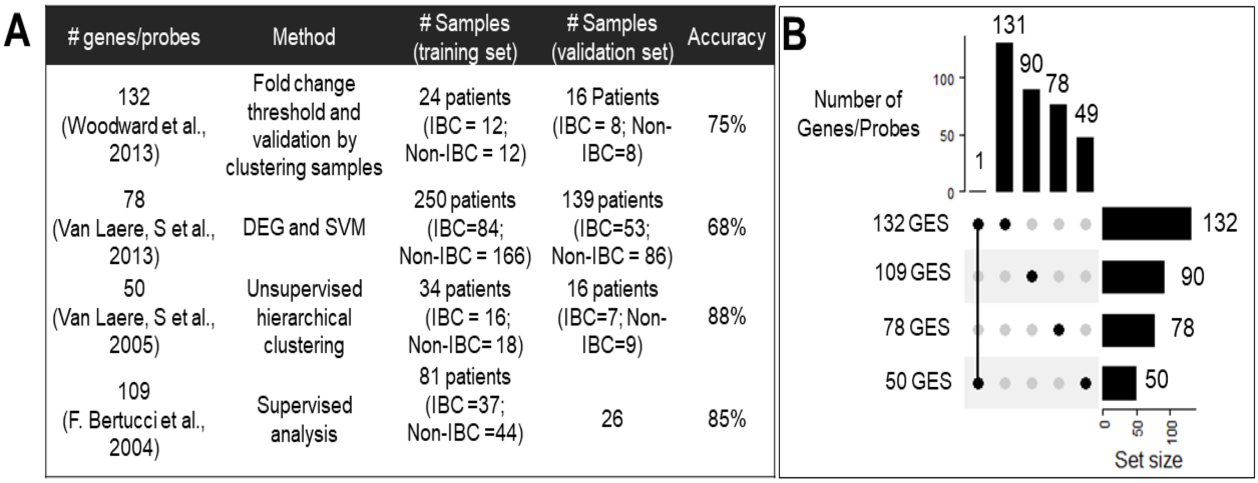
Extensive efforts have been undertaken to identify unique gene expression signatures of IBC. In 2013, the IBC World Consortium identified a 79-gene signature that focused on the inhibition of TGFβ signaling as molecular footprint of IBC [6]; this was consistent with experimental observations that weakened TGFβ signaling promoted collective cell invasion, a hallmark of IBC, while a strong activation of TGFβ signaling promotes individual invasion consistent with a full-blown EMT [7–9]. However, later, these differences were found to arise due to difference in incidence of HER2-positive subtype in IBC vs. non-IBC samples used [1]. Further efforts involving micro-dissected tumors led to identification of 132-gene signature associated with poor outcome [10], but this signature was seen in approximately 25% of breast cancer samples in TCGA which in fact has very few IBC samples, thus highlighting the limited ability of this signature in identifying IBC [1]. The only other study using micro-dissected tumors identified differences in gene expression in the stroma, instead of the tumor cells, in IBC vs. non-IBC cases [11]. Thus, a comprehensive gene expression signature of IBC remains to be identified.

Here, we compare the utility of multiple proposed IBC gene expression signatures by their ability to identify IBC vs. non-IBC cases. Our results here reveal shortcomings in the consistency and predictive utility of these signatures. We subsequently show that single-sample gene set enrichment analysis (ssGSEA) of IBC samples correlate with their EMT scores as calculated via three independent metrics that quantify the spectrum of epithelial-hybrid-mesenchymal phenotypes. Next, we show that while mean EMT scores of IBC samples were not consistently high or low when compared with corresponding non-IBC samples, IBC samples robustly displayed a comparatively higher coefficient of variation in terms of their EMT score. These results suggest that higher heterogeneity along epithelial-hybrid mesenchymal spectrum can be regarded a hallmark of IBC.

Results

Available IBC gene signatures do not distinguish robustly between IBC and non-IBC

First, we collated the four available gene lists identified to be unique to IBC. These four gene signatures showed high accuracy in classifying IBC samples from non-IBC (nIBC) samples in their respective studies. Each signature was comprised of a different number of genes/probes – 132, 78, 50 and 109 (denoted as 132 GES, 78 GES, 50 GES and 109 GES henceforth) [6,10,12,13]. The 109 GES signature consists of 109 different probe sets which uniquely mapped to 90 genes; in all other gene signatures, all probe sets mapped to an equal number of genes. These gene signatures were identified via distinct data-driven algorithms, each utilizing a single dataset and variable number of samples in the respective training and validation sets. These signatures exhibited varied accuracy in identifying IBC cases (**Fig 1A**). Investigating the intersection of common genes amongst each signature revealed minimal or no overlap. Aside from one common gene identified by 132 GES and 50 GES, all other signature elements were unique (**Fig 1B**).



# genes/probes	Method	# Samples (training set)	# Samples (validation set)	Accuracy
132 (Woodward et al., 2013)	Fold change threshold and validation by clustering samples	24 patients (IBC = 12; Non-IBC = 12)	16 Patients (IBC = 8; Non-IBC=8)	75%
78 (Van Laere, S et al., 2013)	DEG and SVM	250 patients (IBC=84; Non-IBC = 166)	139 patients (IBC=53; Non-IBC = 86)	68%
50 (Van Laere, S et al., 2005)	Unsupervised hierarchical clustering	34 patients (IBC = 16; Non-IBC= 18)	16 patients (IBC=7; Non-IBC=9)	88%
109 (F. Bertucci et al., 2004)	Supervised analysis	81 patients (IBC =37; Non-IBC =44)	26	85%

Fig 1: IBC gene signatures A) Summary of all available IBC gene signatures **B)** Overlap in different IBC gene signatures

Based on high accuracy values (68% - 88%) of these gene lists in their ability to identify IBC in corresponding datasets, we hypothesized that each gene list might be capable of correctly classifying IBC and nIBC cases in three datasets containing gene expression data of clinically annotated IBC and nIBC samples (see Methods). We first performed principal component analysis (PCA) on all genes present in these datasets. In each case, PCA showed no clear separation between IBC and nIBC (**Fig 2A, i; Fig S1A i; S2A, i**), thus highlighting the high transcriptional heterogeneity observed in IBC and nIBC. Surprisingly, none of the four IBC gene lists (132 GES, 78 GES, 50 GES and 109 GES) performed consistently better than the all-genes approach in being able to segregate between IBC and non-IBC samples (**Fig 2A, ii-v; Fig S1A, B ii-iv**), with the only possible exception being the performance of 132 GES in one of the three datasets (**Fig 2A, iii**). This segregation was not improved by using non-linear methods such as uMAP as well (**Fig 2B, S1B, ii; S2B, ii**), suggesting that the mRNA levels of these genes are unable to resolve IBC and nIBC.

ssGSEA scores of IBC gene signatures helps in separation of IBC and nIBC samples

As both linear and non-linear combinations of all four gene lists failed to segregate the IBC from nIBC samples in these datasets, we next examined whether as a group, these genes are enriched in IBC samples or not. We used single-sample GSEA (ssGSEA), an extension of Gene Set Enrichment Analysis (GSEA), which calculates separate enrichment scores for each pair of a sample and a gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are cumulatively up- or down-regulated within a given sample [14]. We thus tested whether IBC gene signatures are relatively enriched in IBC samples.

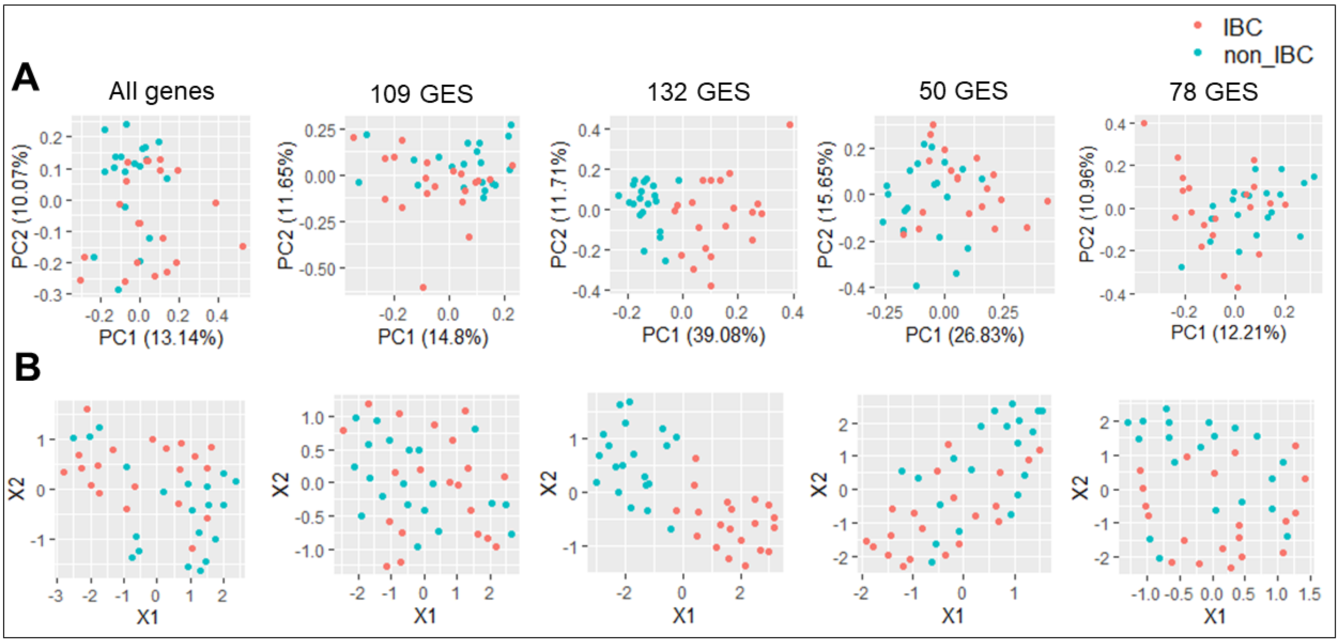
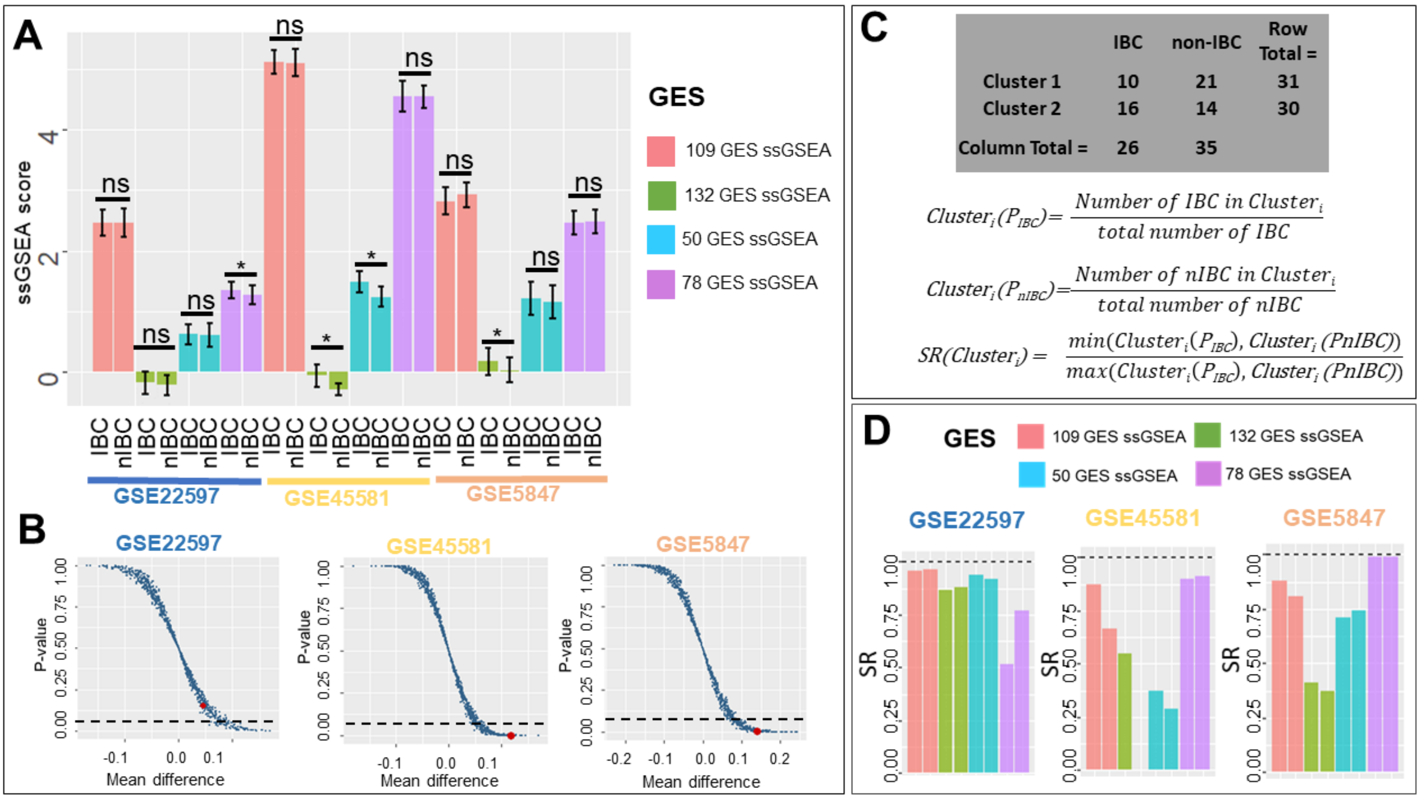


Fig 2: IBC gene signature expression in GSE45581 using dimension reduction methods A) PCA **B)** uMAP

We calculated ssGSEA enrichment scores for all four different gene sets and compared these scores between IBC and nIBC samples across three corresponding datasets. This comparison showed some instances of statistically significant differences in ssGSEA scores of IBC and nIBC samples: a) ssGSEA scores of 78 GES for IBC samples was higher than that of nIBC samples in GSE22597, b) ssGSEA scores of 50 GES and 132 GES was relatively higher for IBC samples in GSE5847, and c) ssGSEA scores of 132 GES were comparatively higher for IBC samples in GSE45581 (**Fig 3A**). However, none of the four gene sets showed consistent and statistically significant differences between IBC and nIBC across the three datasets.

Next, we performed a permutation test to check whether the statistical differences observed in the mean ssGSEA scores of IBC and non-IBC samples were specific to these gene lists. This test was performed for 132 GES, because it showed significant results in terms of being enriched in IBC samples vs. non-IBC ones ($p<0.05$) for two out of three datasets. We chose the same number of genes (132) randomly out of the entire list of genes available to calculate the ssGSEA scores of IBC and non-IBC samples and tested whether the means of ssGSEA scores were significantly different for IBC and non-IBC samples. We repeatedly generated 1000 such instances and compared it to the ssGSEA scores obtained for 132 GES for each instance. This experiment showed that across the three datasets, for a large number of such randomly chosen gene lists, the difference in mean values of ssGSEA scores was not statistically significant (**Fig 3B**). This analysis indicated that the predefined 132 GES is quite likely to better distinguish between IBC and non-IBC compared to a randomly chosen gene set, but the extent of utility of this signature needs further validation.



	IBC	non-IBC	Row Total =
Cluster 1	10	21	31
Cluster 2	16	14	30
Column Total =	26	35	

$$Cluster_i(P_{IBC}) = \frac{Number\ of\ IBC\ in\ Cluster_i}{total\ number\ of\ IBC}$$

$$Cluster_i(P_{nIBC}) = \frac{Number\ of\ nIBC\ in\ Cluster_i}{total\ number\ of\ nIBC}$$

$$SR(Cluster_i) = \frac{min(Cluster_i(P_{IBC}),\ Cluster_i\ (PnIBC))}{max(Cluster_i(P_{IBC}),\ Cluster_i\ (PnIBC))}$$

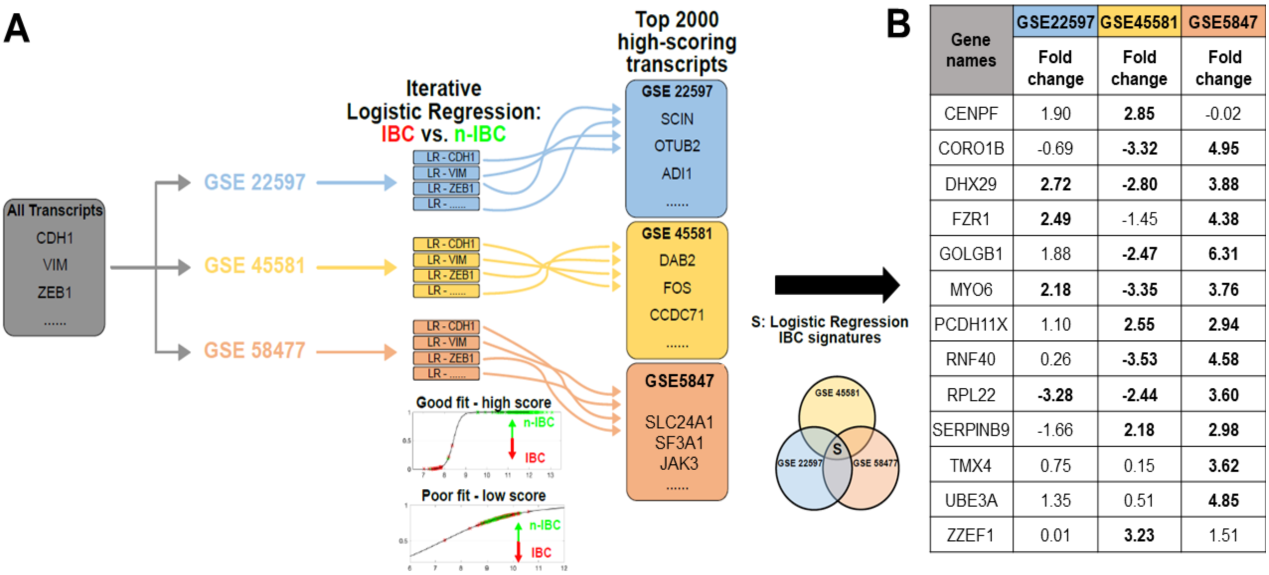
Fig 3: ssGSEA scores of IBC gene signatures. A) ssGSEA scores in IBC and nIBC samples across three datasets (* p< 0.05, two-tailed Student's t-test; error bars represent standard deviation). **B)** Scatter plot showing ssGSEA scores calculated from 1000 random combinations of 132 genes each. Each data point in the plot represents data for a randomly generated gene list; the single point highlighted in red is calculated from 132 GES. x-axis is Mean difference = mean($ssGSEA_{IBC}$) - mean($ssGSEA_{nIBC}$) and y-axis is p-value estimated by *two-tailed Student's t-test* of ssGSEA score across IBC and nIBC group (dotted lines denote p-value = 0.05). **C)** SR (sample ratio) method to measure accuracy of clustering based on ssGSEA scores (GSE5847 clusters) **D)** SR (sample ratio) of different gene signature ssGSEA scores across three datasets.

After comparing ssGSEA scores, we next tested their ability to sort samples into two clusters and checked whether those two clusters corresponded to IBC and nIBC samples. We performed k-means (k=2) clustering on all four ssGSEA scores to cluster the samples and measured the accuracy of clustering into IBC/nIBC based on the SR (sample ratio) values (**Fig 3C**). A perfect clustering would mean that one of two clusters identified via k-means contains all IBC samples in that dataset, and the other contains all nIBC samples. To quantify the effectiveness of clustering into IBC/nIBC, we first calculated cluster IBC and nIBC scores for both clusters; these are the percentage of IBC (and nIBC) samples in the entire dataset that get classified into this cluster. The closer one of these percentages is to 100% and the closer the other percentage is to 0%, the more exclusive (or accurate) the clustering. Thus, we calculated the sample ratio (SR) by dividing the smaller of these two cluster scores by the larger of these scores. For a given cluster, the further the SR value is from 1 or the closer the SR value is to 0, the closer that cluster is to contain either all IBC or nIBC samples. For GSE22597, clusters formed on basis of 78GES had the lowest SR values compared to clusters formed by 132 GES, 50GES or 109 GES. For GSE45581, the SR values of 132 GES were shown to be the least; for GSE5847, 132 GES performed the best (**Fig 3C**).

We considered the actual sample groups and cluster numbers as two categorical variables. Statistical significance of enrichment of IBC and nIBC samples across these two clusters for each dataset was calculated based on a Fisher-exact test. 78 GES was the best performer in GSE22597, 50 GES in GSE45581, and 132 GES in GSE5847 (**Fig 3C-D**).

Logistic Regression iteratively identifies an IBC signature

After exploring all the available IBC gene signatures, next we tried to define IBC signature by applying logistic regression (LR) to the three different datasets. LR is a machine learning-based classification scheme that can predict the probability of a given sample to belong to one of the two (or more) categories. The LR approach has been used to bin samples into epithelial, mesenchymal or hybrid epithelial/mesenchymal categories [15]. Applying LR to each dataset individually, all transcripts are ranked based on their ability to distinguish between IBC and non-IBC samples in the corresponding dataset. The generation of the specific IBC signatures yielded reasonable preliminary models of IBC vs. non-IBC (**Fig 4A**). Predictor goodness-of-fit and predictive accuracies correlated within each dataset, with the highest values observed in GSE45581 and the lowest in GSE22597 (**Table S1**). First, we identified which transcripts were best able to resolve IBC from non-IBC samples in individual datasets. While top transcripts varied across each dataset in their ability to separate IBC and non-IBC samples, with deviances ranging from 18.3 to 89.7 in the top ten from each dataset, their predictive accuracies were quite comparable – 75%-80% (**Table S1**).



Gene names	GSE22597	GSE45581	GSE5847
	Fold change	**Fold change**	**Fold change**
CENPF	1.90	**2.85**	-0.02
CORO1B	-0.69	**-3.32**	**4.95**
DHX29	**2.72**	**-2.80**	**3.88**
FZR1	**2.49**	-1.45	**4.38**
GOLGB1	1.88	**-2.47**	**6.31**
MYO6	**2.18**	**-3.35**	**3.76**
PCDH11X	1.10	**2.55**	**2.94**
RNF40	0.26	**-3.53**	**4.58**
RPL22	**-3.28**	**-2.44**	**3.60**
SERPINB9	-1.66	**2.18**	**2.98**
TMX4	0.75	0.15	**3.62**
UBE3A	1.35	0.51	**4.85**
ZZEF1	0.01	**3.23**	1.51

Fig 4: Identification of IBC-relevant Gene Signatures via Iterative Logistic Regression. A) Logistic regression was applied to each of the transcripts for GSE227597 (blue), GSE45581 (yellow), and GSE58477 (red). Transcripts were sorted based on their ability to resolve IBC and n-IBC samples. The intersection of the top 2000 transcripts were used to define the LR-specific IBC signature, S. **B)** LR signature fold-change between IBC vs. nIBC across three datasets. The bold-cases show when the

fold-change value (mean (IBC)/mean (nIBC)) was statistically significant ($p < 0.05$) using a Students' two-tailed t-test.

These results suggest that IBC can be reasonably identified, quite simply by using a single predictor, provided the analysis is restricted to a given dataset. On the other hand, generalizing across datasets is less straightforward. In an attempt to define an IBC signature based on the mutual intersection of three datasets, we looked for overlap in top ranked transcripts from each dataset; top 200 transcripts revealed no common genes. In top 2000 transcripts, 13 of them were found to be common across the three datasets. Third, we calculate the fold-change in levels of these 13 genes in IBC vs. non-IBC samples in these datasets. None of the 13 transcripts showed consistent upregulation or downregulation in IBC samples across the three datasets.

While our iterative approach represents the maximal resolvability achievable for each dataset using the iterative LR approach, the discrepancies across datasets seen in our earlier analysis do not completely disappear. For instance, lack of resolvability was also seen in IBC and non-IBC when using the LR-derived gene lists in PCA across the three datasets (**Fig S3 B-D**). Thus, increasing the predictive accuracy further may be a useful goal for future pursuits of a universal IBC signature. Such an approach is possible by utilizing multivariate LR models, but it requires significantly more training data for IBC and non-IBC cases, to prevent overfitting.

Correlation between IBC gene signature ssGSEA scores and EMT scores

It has been proposed that IBC cells exhibit a partial EMT behavior, given the retention of E-cadherin levels and the trait of collective cell migration through tumor emboli [16]. Thus, following the assessment of IBC gene signatures, we quantified the EMT-ness of IBC and nIBC samples based on three different EMT scoring metrics – KS, 76GS and MLR [17].

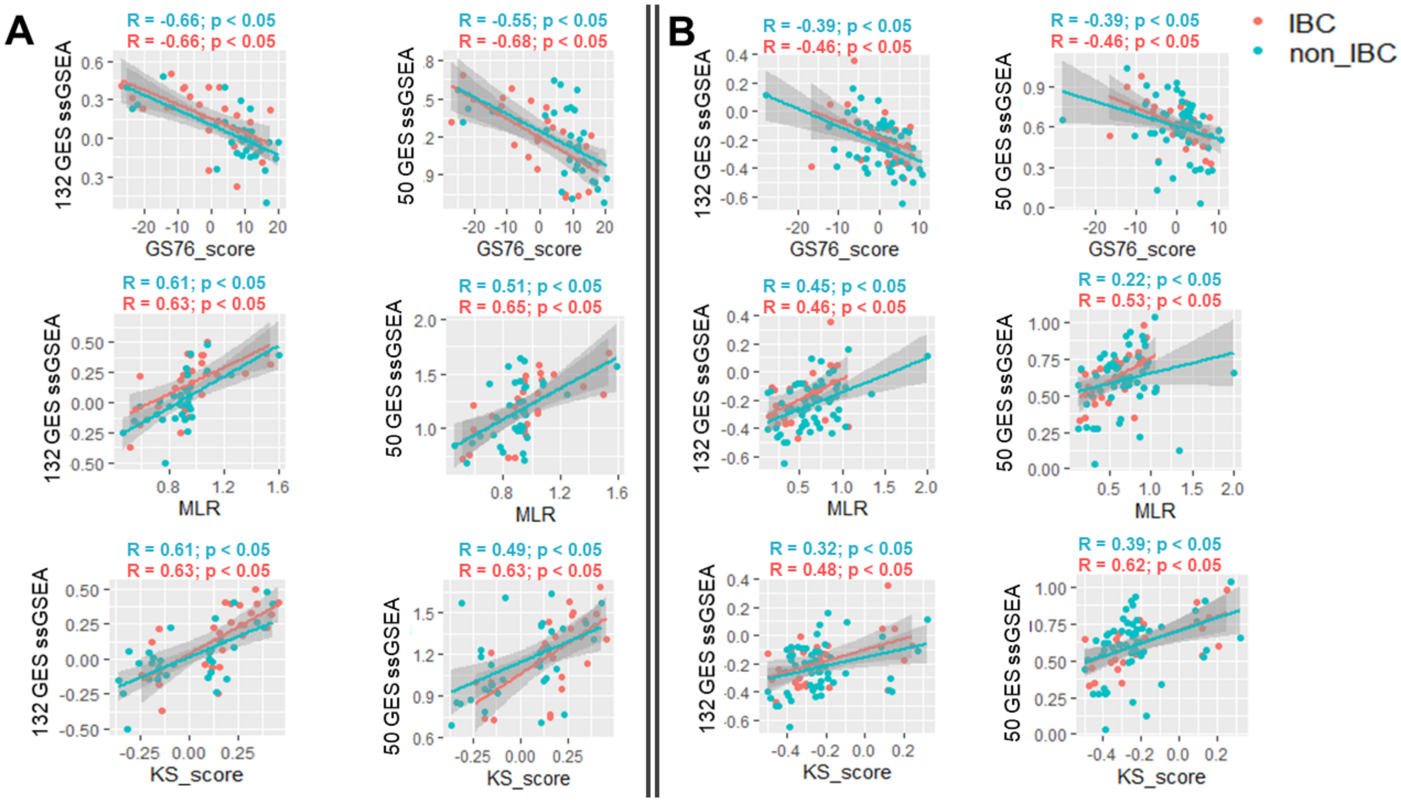


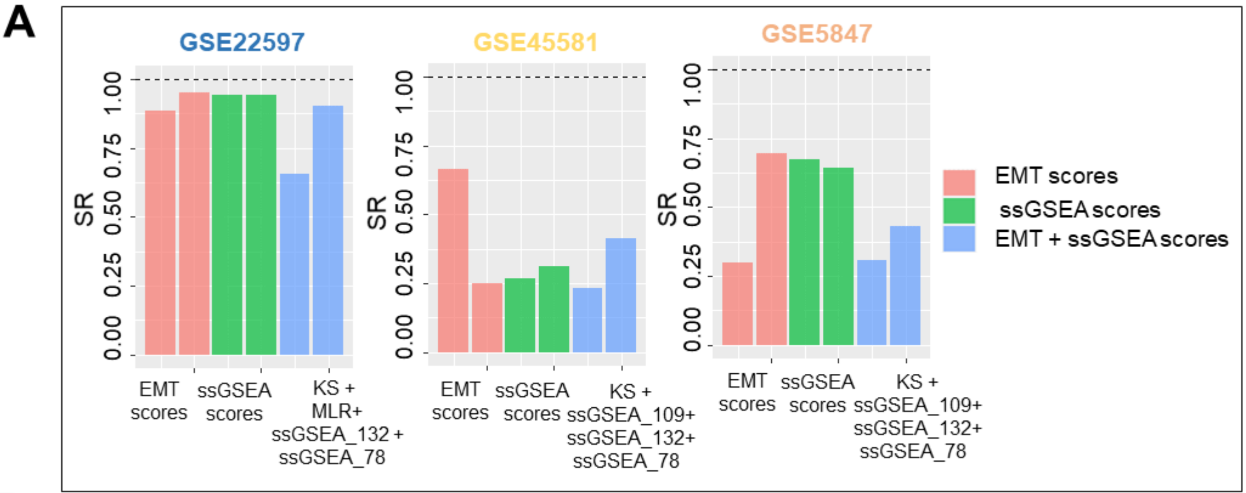
Fig 5: Correlation between ssGSEA score (132 GES and 50 GES) and EMT scoring methods. A) GSE5847 **B)** GSE22597. Pearson's correlation R and p-values highlighted above each scatter plot

These metrics score EMT on a continuum, based on the transcriptomics of individual samples. While KS and MLR score the samples on a scale of [-1, 1] and [0, 2] respectively, the 76GS metric has no pre-defined scale. The higher the MLR or KS score, the more mesenchymal the sample is; the higher the KS score, the more epithelial the sample is. Thus, KS and MLR scores of samples in a given dataset correlate positively with one another; both of them correlate negatively with 76GS scores, as seen across multiple datasets (**Table S2**)

Here, we used these metrics to estimate where IBC and nIBC samples lie in the entire epithelial-hybrid-mesenchymal spectrum, followed by an assessment of correlation of EMT scores with their corresponding ssGSEA enrichment scores. Two ssGSEA enrichment scores (132 GES and 50 GES) showed consistently very high and significant correlations with all three EMT scoring metrics in two datasets (GSE45581, GSE5847). Overall, a higher enrichment in IBC signature is associated with a more EMT-like phenotype. These correlations were maintained across both IBC and nIBC samples (**Fig 5A, B, Fig S6-S8**). However, this consistency is lost when using 78GES or 109GES in these datasets as well as in the case of GSE22597 using any of the four ssGSEA scores (132 GES, 50GES, 78 GES, 109 GES) (**Fig S3, S4**). Taken together, these results suggest that some of the IBC gene lists may enrich for mesenchymal samples instead of segregating IBC/ nIBC. The heterogeneity of both IBC and nIBC samples along epithelial-hybrid-mesenchymal spectrum may contribute to compromising the accuracy of these gene lists in identifying IBC.

Combination of EMT score and ssGSEA score helps in better separation of IBC and nIBC samples

Next, we asked whether IBC and nIBC can be separated using a combination of two different dimensions - ssGSEA enrichment score (using one or more of the four gene lists – 78 GES, 132 GES, 50 GES, 109 GES) and EMT score (using one or more of the EMT scoring metrics). To test the performance of different classifier combinations in sample segregation, we used different numbers and combinations of ssGSEA enrichment scores and EMT scores to perform k-means clustering. We used one, two, three, four, five, six, and seven different scores in all possible combination to cluster the samples into two groups. The clustering accuracy was again measured based on both SR value and a Fisher-exact test. This exercise showed that a combination of EMT scores and ssGSEA scores performs best in terms of clustering IBC and nIBC into separate groups (**Table S3**). Again, these combination of scores were not consistent across different datasets but it was always one or more ssGSEA scores with the combination of one or more EMT scores. Based on the SR value and Fisher exact test – (1) Combination of 78 GES and KS in GSE22597 (2) Combination of 50 GES, 132 GES, and KS in GSE45581 and (3) Combination of 132 GES, MLR, and KS in GSE5847 were the best performers in terms of clustering (**Fig 6A**).



B

Dataset	Sample	COV			Mean		
		76GS	MLR	KS	76GS	MLR	KS
GSE22597	IBC	112.3	0.52	1.08	-0.7	0.96	0.16
	nIBC	253.6	0.49	0.758	7.09	0.89	0.01
GSE45581	IBC	2.94	0.26	0.99	-0.05	0.55	-0.20
	nIBC	2.41	0.22	0.67	0.02	0.46	-0.23
GSE5847	IBC	17.2	1.25	20.38	-1.73	0.18	-0.15
	nIBC	1.5	0.74	1.20	1.93	0.13	-0.21

Fig 6: EMT scoring based identification of IBC. A) Clustering accuracy based on the combination of ssGSEA and EMT scores **B**) EMT Score mean and COV (coefficient of variation) across IBC and nIBC samples across the three independent datasets

EMT score COV (coefficient of variance) is higher in IBC samples

Finally, after establishing the importance of EMT scores in segregation of IBC and nIBC samples, we compared these scores across IBC and nIBC. There was no significant difference in mean scores of IBC vs. nIBC across the three datasets (**Fig S9**), with overlapping variance between the two groups. However, the within-group coefficient of variance (COV), a better measure than variance to assess the dispersion around the mean, was consistently higher in 8 out of 9 (3 EMT scores x 3 datasets) total cases (**Fig 6B**). This result shows IBC samples are more heterogenous as compared to nIBC in terms of their positioning on EMT spectrum.

Discussion

Identifying unique genomic or transcriptomic signatures for IBC has been a challenge, and this lack of consensus limits potential molecular therapeutic approaches to treat this rare but deadly disease [1]. The term 'inflammatory' for IBC originated from its physical appearance, which mimics an acute inflammation of the breast [18]. However, a useful association between the inflammatory phenotype and the cell's omics has not yet been established.

An earlier study that attempted to characterize IBC based on clinical presentation as a distinct molecular entity concluded that “molecular subtype and inflammatory character are two independent features of breast cancers” [19], as the major molecular subtypes described for non-IBC were also found to exist within IBC. Nevertheless, there were multiple studies aimed at defining the molecular signature of IBC using genome-wide gene expression profiling. These studies used several different unsupervised and supervised methods to identify features related to IBC [6,10,12,13]. These gene lists have little to no overlap, and none of them so far have been interpretable as one or more common biological processes or pathways. The main objective of our study was not to predict the status of a new sample as IBC or non-IBC, but to validate the previously defined IBC gene signatures using available datasets consisting of IBC and non-IBC samples. These gene signatures were defined in individual datasets containing both IBC and non-IBC samples using specific statistical models, and we investigated if these gene lists are capable of identifying IBC and non-IBC in independent datasets. Here, we compare these previously defined gene signatures in their ability to classify IBCs from non-IBC, based on a analysis of three separate microarray datasets, none of which were directly involved in identifying the four gene lists (132 GES, 109 GES, 78 GES, 50 GES). Additionally, we also elucidate the EMT status of IBC samples using three different transcriptomics-based EMT scores. To the best of our understanding, this is the first study that contrasts different IBC gene signatures and EMT scoring across IBC datasets.

Mechanistic studies utilizing *in vitro* and *in vivo* models have revealed some markers for IBC, such as P-cadherin [20]. This molecule has also been proposed as a marker of the hybrid epithelial/mesenchymal (E/M) phenotype [21] due to its role in promoting collective cell migration and invasion [22,23] as well as tumor-initiating properties [24]; both of these properties are considered as hallmarks of hybrid E/M phenotype(s) [25]. P-cadherin (CDH3) is also a transcriptional target of NP63α [26], another potential ‘phenotypic stability factor’ (PSFs) for a hybrid E/M phenotype [16]. Similar to other PSFs [27–29], overexpression of P-cadherin associates with poor clinical outcome in invasive breast carcinomas [30]. Another pathway that has been reported to be enriched in IBC is the IL-6 pathway [16] which can promote Notch-JAG1 signaling [31]. JAG1 was reported as one of the top upregulated genes in collectively migrating cells [32] and its knockdown severely inhibited emboli formation in SUM149 IBC cells [31]. Moreover, given the role of IL-6 in mediating tumor-stroma crosstalk in IBC [33], it is possible that IL-6 mediates cell-cell communication both among IBC cells and with stroma. Despite these promising mechanistic insights, no accurate predictive signature for IBC exists.

Our results highlight that the proposed IBC gene lists so far (109 GES, 78 GES, 50 GES and 132 GES), despite showing a good accuracy in their corresponding validation datasets, show quite limited success in segregating IBC from non-IBC samples in independent cases. Various reasons may contribute to this result: inconsistency in the identification of IBC in different clinical samples, possible contamination by stromal cells in the samples investigated, and/or lack of metrics other than gene enrichment. To further indicate this failure of transferability, we used logistic regression (LR) methods to identify predictors (genes) that can best segregate between IBC and non-IBC, but no common trend was seen even in the top 2000 predictors collated from each of the three clinical datasets investigated. Put together, there exists a need to examine alternative metrics to be able to accurately identify IBC samples and perhaps gene expression on its own is insufficient to distinguish between IBC and non-IBC. One potential way to overcome the existing limitation may be to apply multivariate LR, but more training data to identify IBC from non-IBC samples would be required there to prevent any overfitting. Another approach would be to incorporate additional modalities of data, e.g. proteomics, to distinguish between IBC and nIBC.

Given the extensive literature on the role of a partial or complete EMT in collective migration and metastasis in breast cancer [32,34,35], we investigated if a more mesenchymal phenotype correlated with enrichment of ssGSEA scores for IBC gene lists in breast cancer samples. Indeed, ssGSEA scores for two IBC gene lists (50GES, 132 GES) correlated significantly with more mesenchymal samples, irrespective of whether those samples belonged to IBC or non-IBC. However, the ssGSEA scores for the 78GES list correlated with a more epithelial phenotype specifically for IBC samples, reminiscent of 78GES being associated with attenuated TGFβ signaling that may drive collective migration. Put together, one way to interpret these results may be that an 'intermediate' EMT associates with IBC, but at large, this inconsistency demonstrates a complex relation between IBC and EMT-ness and strengthens the idea of EMT- related heterogeneity in IBC. It is worth noting that these gene lists have been identified based on primary tumors and the expression signatures of primary tumors may contain little information about whether circulating tumor cells (CTCs) migrate individually or collectively [36]. Further characterization of emboli or clusters of CTCs for IBC will help deconvolute the contribution of EMT for IBC metastasis. Moreover, single-cell analysis of primary tumors and CTCs of IBC shall help in identifying various immune cell subsets in IBC which may drive disease progression. The composition and/or spatial localization of immune cells is likely to yield better insights into immune ecology of highly aggressive disease such as IBC [37,38].

The EMT scores for IBC as well as non-IBC samples did not indicate a 'full-blown' EMT (i.e. MLR score > 1.5 or equivalently KS score > 0.6), thus strengthening our previous observations that a complete EMT is not required for metastasis [39]. Intriguingly, we did observe a higher heterogeneity in EMT scores for IBC samples as compared to non-IBC samples. It remains to be ascertained whether the inter-tumor heterogeneity in EMT scores in IBC is also reflected as high intra-tumor heterogeneity as well. If that turns out to be the case, a higher intra-tumor heterogeneity along the EMT spectrum can be considered as a potential biomarker for IBC.

While genomic heterogeneity in tumors has been extensively studied, quantifying non-genetic (i.e. phenotypic) heterogeneity has been possible only recently through investigating cell-to-cell variability in isogenic populations [40–43]. Higher phenotypic heterogeneity may encourage cancer invasion [44] as well as the evolution of therapy resistance [45]. It may arise from network topology features such as mutually inhibitory feedback loops [46], for instance, the loop between RKIP and BACH1 [47] or that between AMPK and AKT [48]. Our earlier attempts have highlighted network topology based features such as hierarchical organization as a marker for IBC [39], endorsing that quantitative metrics to dissecting phenotypic heterogeneity in IBC may be better poised to highlight the IBC hallmarks than searching for gene signatures.

Acknowledgements This work was supported by Ramanujan Fellowship (SB/S2/RJN-049/2018) awarded to MKJ by SERB, DST, Government of India.

Author contributions PC and JTG performed research, HL and WW analyzed data, MKJ designed and supervised research. All authors participated in writing and editing of the manuscript.

Methods

All the analyses have been performed on R 3.4.4 version and data was plotted using ggplot2 package.

Datasets: Three separate IBC datasets were used in this study. GSE5847, GSE22597 and GSE45581 datasets were downloaded from NCBI GEO website.

Principle component analysis (PCA): PCA was performed using prcomp function available in R and plotted using factstoextra R package.

ssGSEA analysis: ssGSEA analysis for various different gene sets were performed using GSVA R Bioconductor package with “ssgsea” option for method argument.

k-means clustering: K-means clustering was performed using R package “cluster” and centers were set as two to get two separate clusters.

EMT scoring: Three different EMT scoring methods – KS, MLR, 76GS were used to score samples separately in the three datasets [17].

Statistical analysis: All the pairwise comparison significance was tested using student’s t-test. Significance of the enrichment of IBC and nIBC samples across clusters were tested using fisher exact test.

Permutation/Randomization test

Permutation test was performed to test the significance of a gene signature as compared to the random set of genes. To determine the efficiency of an IBC gene signature on the basis of a distribution firstly the values were calculated using original expression values then same number of genes were chosen randomly. This process was repeated 1000 times to generate a null distribution of ssGSEA scores. Next, these ssGSEA scores compared across IBC and nIBC groups to obtain difference in mean values and significance based on t-test. This test was used to show that the original expression shows a higher difference in the mean of ssGSEA score and a lower p-value as compared to most of the random cases.

Resolution IBC vs. n-IBC via iterative logistic regression: Samples from GSE22597 were first identified and categorized into IBC and n-IBC samples as previously reported, with all other samples omitted from analysis. The predictor set was comprised of all transcripts and for each individual transcript binomial logistic regression was fitted to the categorical IBC status. The output corresponds to each transcript a generalized residual sum of squared error, or deviance, with smaller values corresponding to better fit. The transcripts were then sorted in increasing order of deviance. For each of top ten predictors coding for genes, a leave-one-out assessment of prediction ability was performed. In each case, the logistic regression model was again constructed, this time on all but one sample. The corresponding statistical model was then used to predict the IBC status of the withheld sample. This procedure was repeated iteratively, withholding a distinct sample each time, and the results of the prediction were aggregated to estimate predictive accuracy.

This process was repeated for two additional datasets (GSE45581 and GSE58477). The top ten predictors for each dataset, together with their deviance values and predictive accuracy are listed in **Figure S3.** The top predictors from each of the three datasets define a logistic regression (LR)-specific IBC signature by taking the mutual intersection of the top 2000 scoring transcripts from each dataset (Figure 3).

References

1. Lim, B.; Woodward, W.A.; Wang, X.; Reuben, J.M.; Ueno, N.T. Inflammatory breast cancer biology: the tumour microenvironment is key. *Nat. Rev. Cancer* **2018**, *18*, 485–499.
2. Woodward, W.A. Inflammatory breast cancer: Unique biological and therapeutic considerations. *Lancet Oncol.* **2015**, *16*, e568–e576.
3. Rosenbluth, J.M.; Overmoyer, B.A. Inflammatory Breast Cancer: a separate entity. *Curr. Oncol. Rep.* **2019**, *21*, 86.
4. Mu, Z.; Wang, C.; Ye, Z.; Austin, L.; Civan, J.; Hyslop, T.; Palazzo, J.P.; Jaslow, R.; Li, B.; Myers, R.E.; et al. Prospective assessment of the prognostic value of circulating tumor cells and their clusters in patients with advanced-stage breast cancer. *Breast Cancer Res. Treat.* **2015**, *154*, 563–571.
5. Jolly, M.K.; Mani, S.A.; Levine, H. Hybrid epithelial/mesenchymal phenotype(s): The 'fittest' for metastasis? *Biochim. Biophys. Acta - Rev. Cancer* **2018**, *1870*, 151–157.
6. Laere, S.J. Van; Ueno, N.T.; Finetti, P.; Vermeulen, P.; Lucci, A.; Robertson, F.M.; Marsan, M.; Iwamoto, T.; Krishnamurthy, S.; Masuda, H.; et al. Uncovering the Molecular Secrets of Inflammatory Breast Cancer Biology: An Integrated Analysis of Three Distinct Affymetrix Gene Expression Datasets. *Clin. Cancer Res.* **2013**, *19*, 4685–4696.
7. Matise, L.A.; Palmer, T.D.; Ashby, W.J.; Nashabi, A.; Chytil, A.; Aakre, M.; Pickup, M.W.; Gorska, A.E.; Zijlstra, A.; Moses, H.L. Lack of transforming growth factor-β signaling promotes collective cancer cell invasion through tumor-stromal crosstalk. *Breast Cancer Res.* **2012**, *14*, R98.
8. Giampieri, S.; Manning, C.; Hooper, S.; Jones, L.; HIll, C.S.; Sahai, E. Localized and reversible TGFβ signalling switches breast cancer cells from cohesive to single cell motility. *Nat Cell Biol* **2009**, *11*, 1287–1296.
9. Jia, W.; Deshmukh, A.; Mani, S.A.; Jolly, M.K.; Levine, H. A possible role for epigenetic feedback regulation in the dynamics of the Epithelial-Mesenchymal Transition (EMT). *Phys. Biol.* **2019**, *16*, 066004.
10. Woodward, W.A.; Krishnamurthy, S.; Yamauchi, H.; El-Zein, R.; Ogura, D.; Kitadai, E.; Niwa, S.; Cristofanilli, M.; Vermeulen, P.; Dirix, L.; et al. Genomic and expression

analysis of microdissected inflammatory breast cancer. *Breast Cancer Res. Treat.* **2013**, *138*, 761–772.

11. Boersma, B.J.; Reimers, M.; Yi, M.; Ludwig, J.A.; Luke, B.T.; Stephens, R.M.; Yfantis, H.G.; Lee, D.H.; Weinstein, J.N.; Ambs, S. A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer* **2008**, *122*, 1324–1332.
12. Bertucci, F.; Finetti, P.; Rougemont, J.; Charafe-Jauffret, E.; Nasser, V.; Loriod, B.; Camerlo, J.; Tagett, R.; Tarpin, C.; Houvenaeghel, G.; et al. Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer Res.* **2004**, *64*, 8558–65.
13. Laere, S. Van; Auwera, I. Van Der; Eynden, G.G. Van Den; Fox, S.B.; Bianchi, F.; Harris, A.L.; Dam, P. Van; Marck, E.A. Van; Vermeulen, B.; Dirix, L.Y. Distinct molecular signature of inflammatory breast cancer by cDNA microarray analysis. **2005**, 237–246.
14. Barbie, D.A.; Tamayo, P.; Boehm, J.S.; Kim, S.Y.; Moody, S.E.; Dunn, I.F.; Schinzel, A.C.; Sandy, P.; Meylan, E.; Scholl, C.; et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**, *462*, 108–112.
15. George, J.T.; Jolly, M.K.; Xu, S.; Somarelli, J.A.; Levine, H. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* **2017**, *77*, 6415–6428.
16. Jolly, M.K.; Boareto, M.; Debeb, B.G.; Aceto, N.; Farach-Carson, M.C.; Woodward, W.A.; Levine, H. Inflammatory Breast Cancer: a model for investigating cluster-based dissemination. *NPJ Breast Cancer* **2017**, *3*, 21.
17. Chakraborty, P.; George, J.T.; Tripathi, S.; Levine, H.; Jolly, M.K. Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Front. Bioeng. Biotechnol.* **2020**, *8*.
18. Robertson, F.M.; Bondy, M.; Yang, W.; Yamauchi, H.; Wiggins, S.; Kamrudin, S.; Krishnamurthy, S.; Le-Petross, H.; Bidaut, L.; Player, A.N.; et al. Inflammatory Breast Cancer: The Disease, the Biology, the Treatment. *CA. Cancer J. Clin.* **2010**, *60*, 351–375.
19. Bertucci, F.; Finetti, P.; Rougemont, J.; Charafe-Jauffret, E.; Cervera, N.; Tarpin, C.; Nguyen, C.; Xerri, L.; Houlgatte, R.; Jacquemier, J.; et al. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res.* **2005**, *65*, 2170–2178.
20. Hamida, A.B.; Labidi, I.S.; Mrad, K.; Charafe-Jauffret, E.; Arab, S. Ben; Esterni, B.; Xerri, L.; Viens, P.; Bertucci, F.; Birnbaum, D.; et al. Markers of subtypes in inflammatory breast cancer studied by immunohistochemistry: Prominent expression of P-cadherin. *BMC Cancer* **2008**, *8*, 28.
21. Ribeiro, A.S.; Paredes, J. P-Cadherin Linking Breast Cancer Stem Cells and Invasion: A Promising Marker to Identify an “Intermediate/Metastable” EMT State. *Front. Oncol.* **2015**, *4*, 371.
22. Plutoni, C.; Bazellieres, E.; Le Borgne-Rochet, M.; Comunale, F.; Brugues, A.; Séveno, M.; Planchon, D.; Thuault, S.; Morin, N.; Bodin, S.; et al. P-cadherin promotes collective cell migration via a Cdc42-mediated increase in mechanical forces. *J. Cell Biol.* **2016**, *212*, 199–217.
23. Ribeiro, A.S.; Albergaria, A.; Sousa, B.; Correia, A.L.; Bracke, M.; Seruca, R.; Schmitt, F.C.; Paredes, J. Extracellular cleavage and shedding of P-cadherin: A mechanism underlying the invasive behaviour of breast cancer cells. *Oncogene* **2010**, *29*, 392–402.
24. Vieira, A.F.; Ricardo, S.; Ablett, M.P.; Dionísio, M.R.; Mendes, N.; Albergaria, A.; Farnie, G.; Gerhard, R.; Cameselle-Teijeiro, J.F.; Seruca, R.; et al. P-cadherin is coexpressed with CD44 and CD49f and mediates stem cell properties in basal-like breast cancer. *Stem Cells* **2012**, *30*, 854–864.
25. Jolly, M.K.; Mani, S.A.; Levine, H. Hybrid epithelial/mesenchymal phenotype(s): the “fittest” for metastasis? *BBA - Rev. Cancer* **2018**, *1870*, 151–157.
26. Shimomura, Y.; Wajid, M.; Shapiro, L.; Christiano, A.M. P-cadherin is a p63 target

gene with a crucial role in the developing human limb bud and hair follicle. *Development* **2008**, *135*, 743–753.

27. Bocci, F.; Tripathi, S.C.; Vilchez, M.S.A.; George, J.T.; Casabar, J.; Wong, P.; Hanash, S.; Levine, H.; Onuchic, J.; Jolly, M. NRF2 activates a partial Epithelial-Mesenchymal Transition and is maximally present in a hybrid Epithelial/Mesenchymal phenotype. *Integr. Biol.* **2019**, *11*, 251–263.
28. Bocci, F.; Jolly, M.K.; Tripathi, S.C.; Aguilar, M.; Hanash, S.M.; Levine, H.; Onuchic, J.N. Numb prevents a complete epithelial-mesenchymal transition by modulating Notch signaling. *J. R. Soc. Interface* **2017**, *14*.
29. Subbalakshmi, A.R.; Kundnani, D.; Biswas, K.; Ghosh, A.; Hanash, S.M.; Tripathi, S.C.; Jolly, M.K. NFATc acts as a non-canonical phenotypic stability fatcor for a hybrid epithelial/mesenchymal phenotype. *bioRxiv* **2020**, 047803.
30. Paredes, J.; Albergaria, A.; Oliveira, J.T.; Jeronimo, C.; Milanezi, F.; Schmitt, F.C. P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. *Clin. Cancer Res.* **2005**, *11*, 5869–5877.
31. Bocci, F.; Gearhart-Serna, L.; Boareto, M.; Ribeiro, M.; Ben-Jacob, E.; Devi, G.R.; Levine, H.; Onuchic, J.N.; Jolly, M.K. Toward understanding cancer stem cell heterogeneity in the tumor microenvironment. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 148–157.
32. Cheung, K.J.; Padmanaban, V.; Silvestri, V.; Schipper, K.; Cohen, J.D.; Fairchild, A.N.; Gorin, M.A.; Verdone, J.E.; Pienta, K.J.; Bader, J.S.; et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc. Natl. Acad. Sci.* **2016**, *113*, E854–E863.
33. Wolfe, A.R.; Trenton, N.J.; Debeb, B.G.; Larson, R.; Ruffell, B.; Chu, K.; Hittelman, W.; Diehl, M.; Reuben, J.M.; Naoto, T.; et al. Mesenchymal stem cells and macrophages interact through IL-6 to promote inflammatory breast cancer in pre-clinical models. *Oncotarget* **2016**, *7*, 82482–82492.
34. Aceto, N.; Bardia, A.; Miyamoto, D.T.; Donaldson, M.C.; Wittner, B.S.; Spencer, J.A.; Yu, M.; Pely, A.; Engstrom, A.; Zhu, H.; et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **2014**, *158*, 1110–1122.
35. Jolly, M.K.; Ware, K.E.; Gilja, S.; Somarelli, J.A.; Levine, H. EMT and MET : necessary or permissive for metastasis ? *Mol. Oncol.* **2017**, *11*, 755–769.
36. Thangavel, H.; Angelis, C. De; Vasaikar, S.; Bhat, R.; Jolly, M.K.; Nagi, C.; Creighton, C.J.; Chen, F.; Dobrolecki, L.E.; George, J.T.; et al. A CTC-Cluster-Specific Signature Derived from OMICS Analysis of Patient-Derived Xenograft Tumors Predicts Outcomes in Basal-Like Breast Cancer. *J. Clin. Med.* **2019**, *8*, 1772.
37. Cohen, E.N.; Gao, H.; Anfossi, S.; Mego, M.; Reddy, N.G.; Debeb, B.; Giordano, A.; Tin, S.; Wu, Q.; Garza, R.J.; et al. Inflammation mediated metastasis: Immune induced epithelial-to-mesenchymal transition in inflammatory breast cancer cells. *PLoS One* **2015**, *10*.
38. Li, X.; Jolly, M.K.; George, J.T.; Pienta, K.J.; Levine, H. Computational Modeling of the Crosstalk Between Macrophage Polarization and Tumor Cell Plasticity in the Tumor Microenvironment. *Front. Oncol.* **2019**, *9*, 1–12.
39. Tripathi, S.; Jolly, M.K.; Woodward, W.A.; Levine, H.; Deem, M.W. Analysis of hierarchical organization in gene expression networks reveals underlying principles of collective tumor cell dissemination and metastatic aggressiveness of inflammatory breast cancer. *Front. Oncol.* **2018**, *8*, 244.
40. Jolly, M.K.; Celia-Terrassa, T. Dynamics of Phenotypic Heterogeneity Associated with EMT and Stemness during Cancer Progression. *J Clin Med* **2019**, *8*, 1542.
41. Meyer, A.S.; Heiser, L.M. Systems biology approaches to measure and model phenotypic heterogeneity in cancer. *Curr. Opin. Syst. Biol.* **2019**, *17*, 35–40.
42. Sharma, A.; Merritt, E.; Hu, X.; Cruz, A.; Jiang, C.; Sarkodie, H.; Zhou, Z.; Malhotra, J.; Riedlinger, G.M.; De, S. Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung

Tumors. *Cell Rep.* **2019**, *29*, 2164–2174.

43. Tripathi, S.; Chakraborty, P.; Levine, H.; Jolly, M.K. A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLOS Comput. Biol.* **2020**, *16*, e1007619.
44. Shin, Y.; Han, S.; Chung, E.; Chung, S. Intratumoral phenotypic heterogeneity as an encourager of cancer invasion. *Integr. Biol. (United Kingdom)* **2014**, *6*, 654–661.
45. Farquhar, K.S.; Charlebois, D.A.; Szenk, M.; Cohen, J.; Nevozhay, D.; Balázsi, G. Role of network-mediated stochasticity in mammalian drug resistance. *Nat. Commun.* **2019**, *10*, 2766.
46. Hari, K.; Sabuwala, B.; Subramani, B.V.; Porta, C. La; Zapperi, S.; Font-Clos, F.; Jolly, M.K. Identifying inhibitors of epithelial-mesenchymal plasticity using a network topology based approach. *npj Syst. Biol. Appl.* **2020**, *6*, 15.
47. Lee, J.; Lee, J.; Farquhar, K.S.; Yun, J.; Frankenberger, C.A.; Bevilacqua, E.; Yeung, K.; Kim, E.-J.; Balázsi, G.; Rosner, M.R. Network of mutually repressive metastasis regulators can promote cell heterogeneity and metastatic transitions. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E364-373.
48. Saha, M.; Kumar, S.; Bukhari, S.; Balaji, S.A.; Kumar, P.; Hindupur, S.K.; Rangarajan, A. AMPK–Akt double-negative feedback loop in breast cancer cells regulates their adaptation to matrix deprivation. *Cancer Res.* **2018**, *78*, 1497–1510.

Supplementary figures

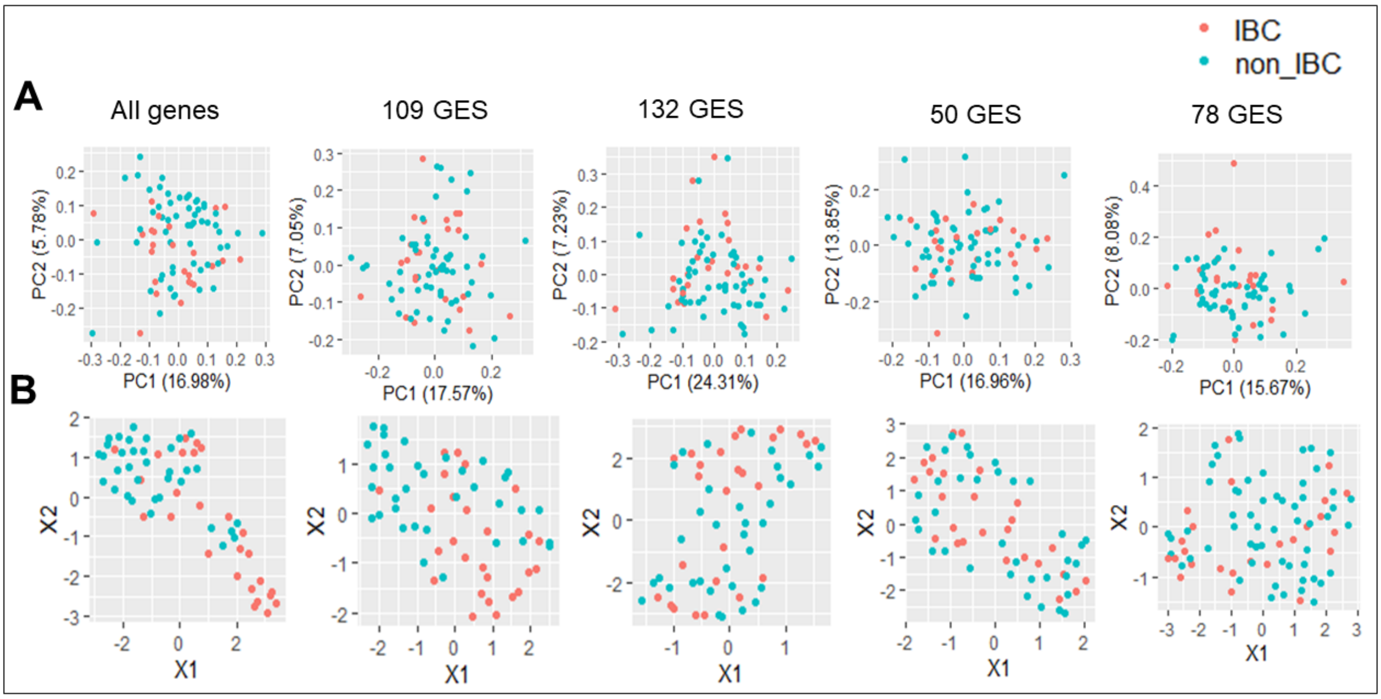


Fig S1: IBC gene signature expression in GSE22597. A) PCA **B)** uMAP

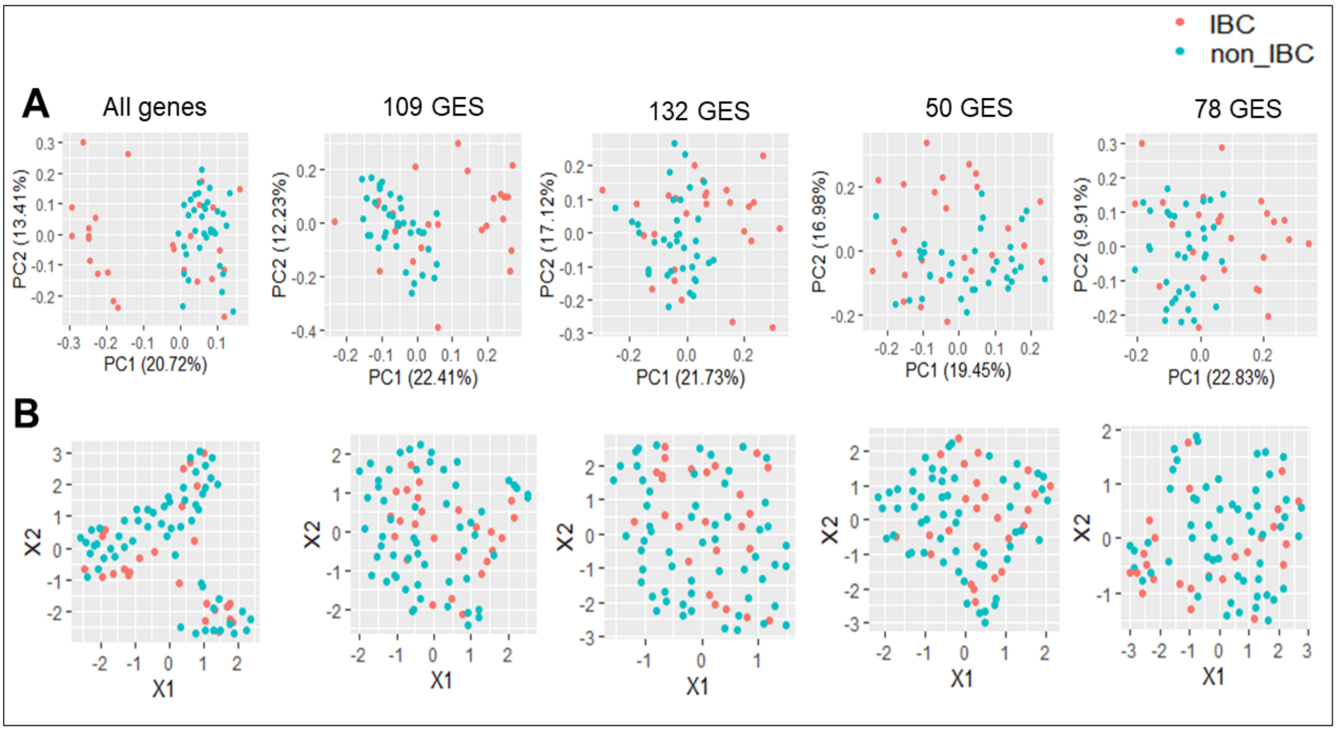
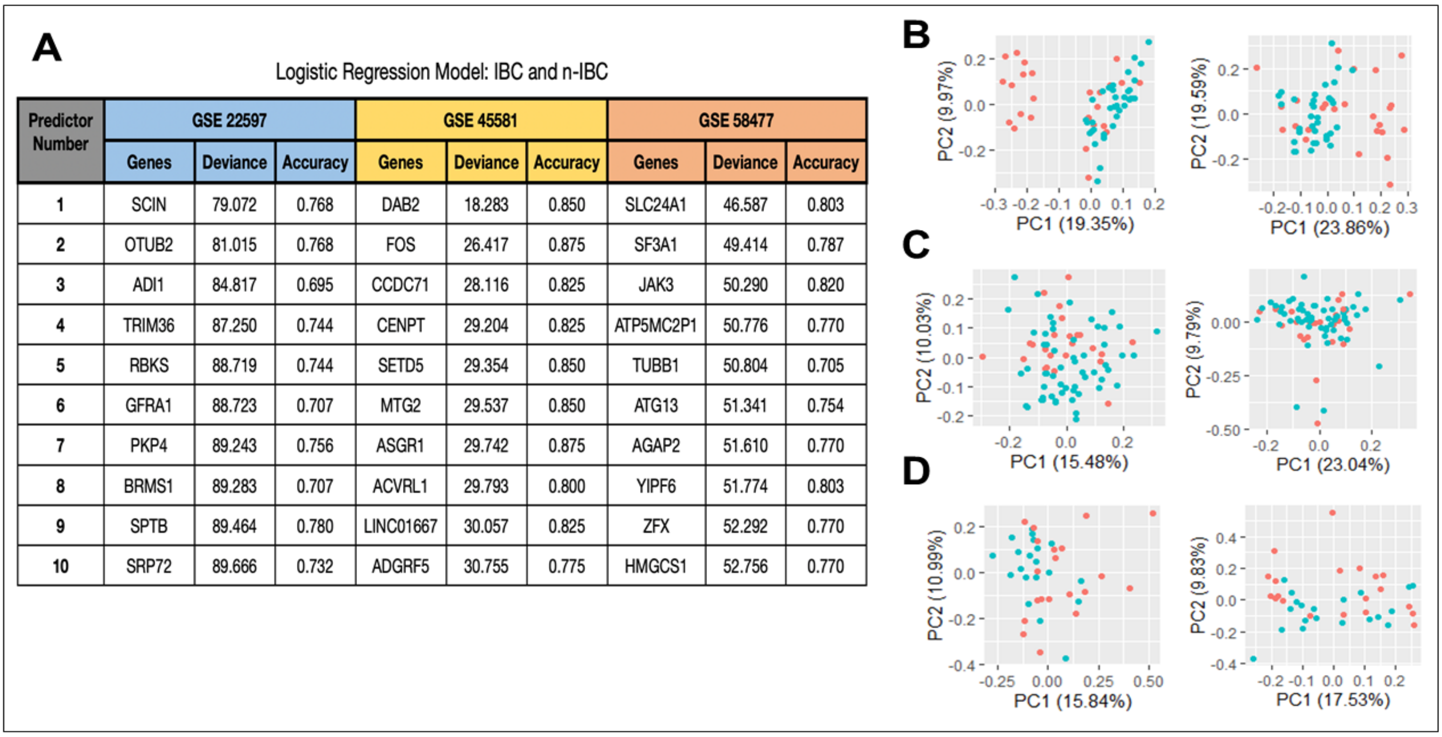


Fig S2: IBC gene signature expression in GSE5847. **A)** PCA **B)** uMAP



A

Logistic Regression Model: IBC and n-IBC

Predictor Number	GSE 22597			GSE 45581			GSE 58477		
	Genes	Deviance	Accuracy	Genes	Deviance	Accuracy	Genes	Deviance	Accuracy
1	SCIN	79.072	0.768	DAB2	18.283	0.850	SLC24A1	46.587	0.803
2	OTUB2	81.015	0.768	FOS	26.417	0.875	SF3A1	49.414	0.787
3	ADI1	84.817	0.695	CCDC71	28.116	0.825	JAK3	50.290	0.820
4	TRIM36	87.250	0.744	CENPT	29.204	0.825	ATP5MC2P1	50.776	0.770
5	RBKS	88.719	0.744	SETD5	29.354	0.850	TUBB1	50.804	0.705
6	GFRA1	88.723	0.707	MTG2	29.537	0.850	ATG13	51.341	0.754
7	PKP4	89.243	0.756	ASGR1	29.742	0.875	AGAP2	51.610	0.770
8	BRMS1	89.283	0.707	ACVRL1	29.793	0.800	YIPF6	51.774	0.803
9	SPTB	89.464	0.780	LINC01667	30.057	0.825	ZFX	52.292	0.770
10	SRP72	89.666	0.732	ADGRF5	30.755	0.775	HMGCS1	52.756	0.770

Figure S3: Features of the top LR predictors. A) The top 10 predictors for the GSE 22597 (blue), GSE 45581 (yellow), and GSE 58477 (red) datasets were identified based on their minimal deviance values. Leave-one out predictive accuracy is reported for each top transcript. **B)** GSE5847 PCA using top 100 LR predictors based on GSE22597 and GSE45581**. C)** GSE22597 PCA using top 100 LR predictors based on GSE5847 and GSE45581. **(D)** GSE45581 PCA using top 100 LR predictors based on GSE22597 and GSE5847.

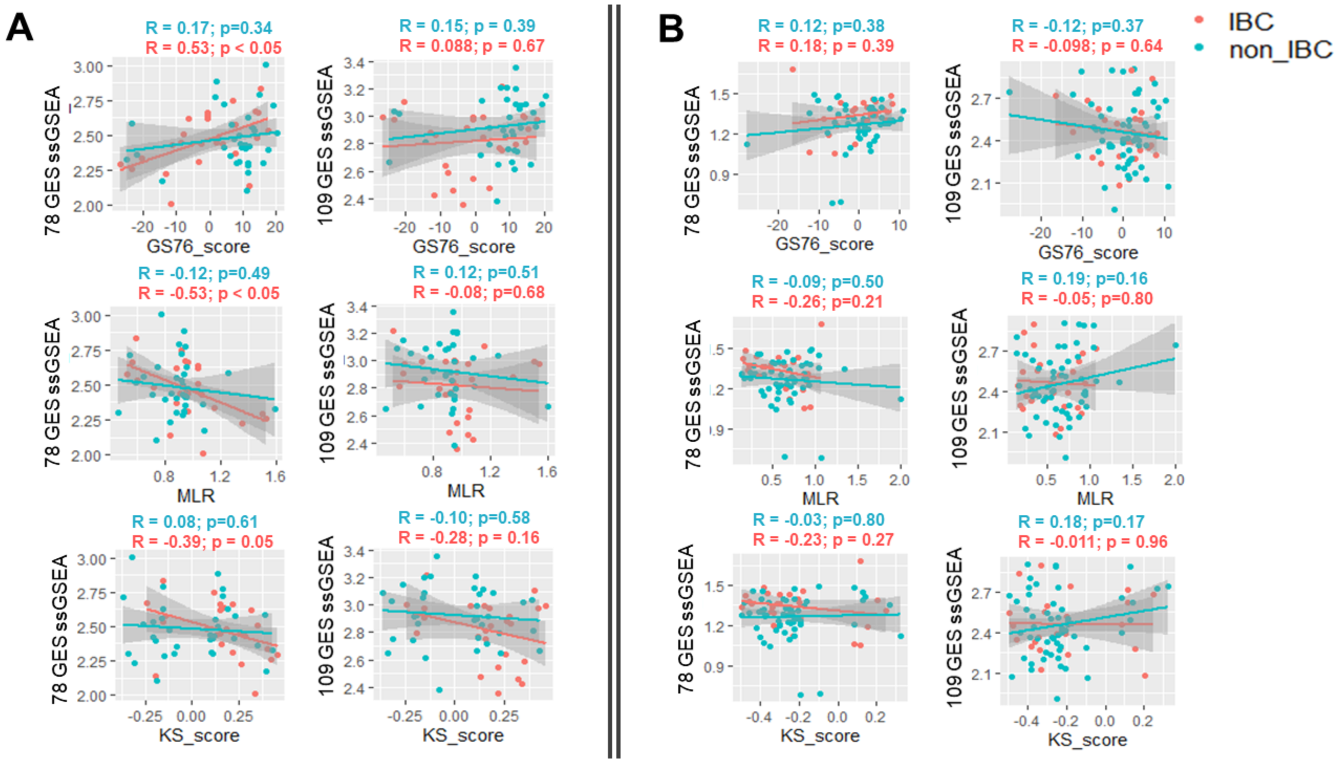


Fig S4: Correlation between ssGSEA score (78 GES and 109 GES) and EMT scoring methods. A) GSE5847 **B)** GSE22597. Pearson's correlation R and p-values high-lighted above each scatter plot (blue – nIBC , red – IBC).

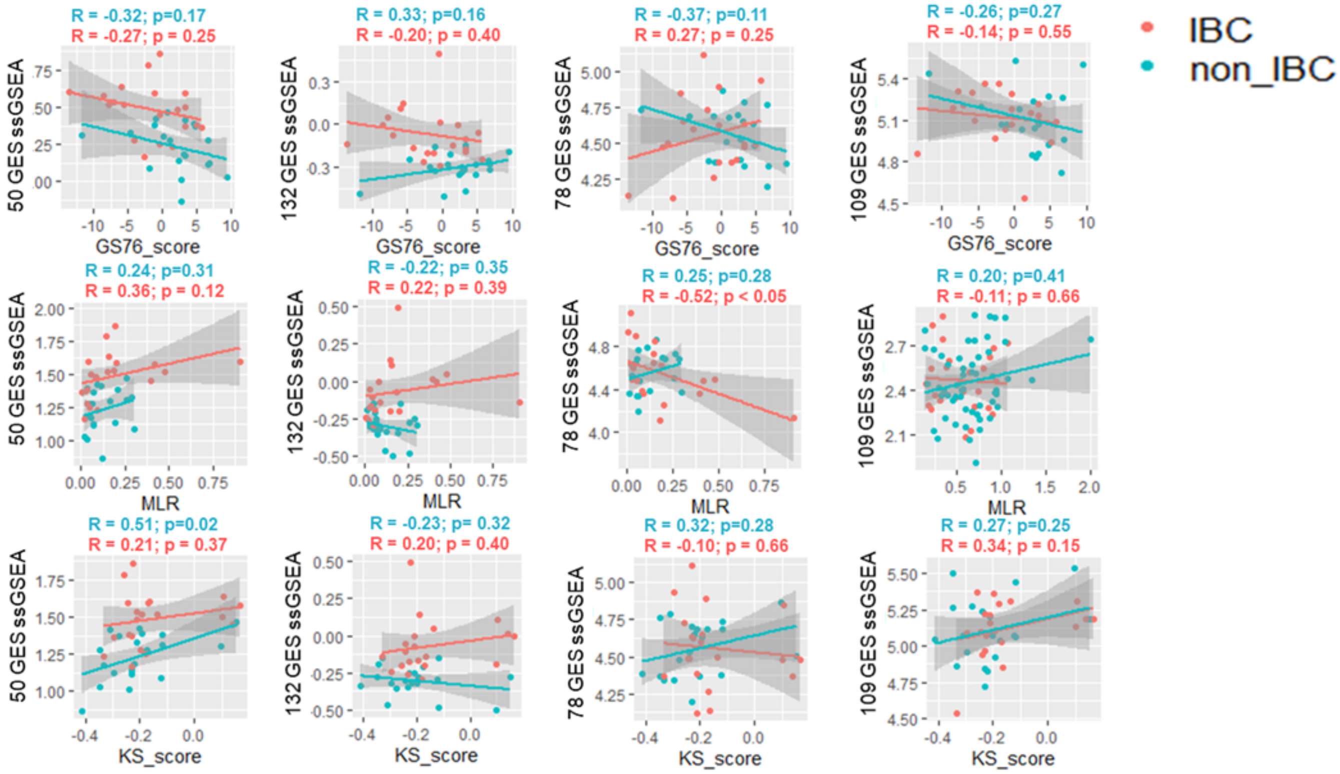


Fig S5: Correlation between ssGSEA score and EMT scoring methods in GSE45581. Pearson's correlation R and p-values high-lighted above each scatter plot (blue – nIBC , red – IBC).

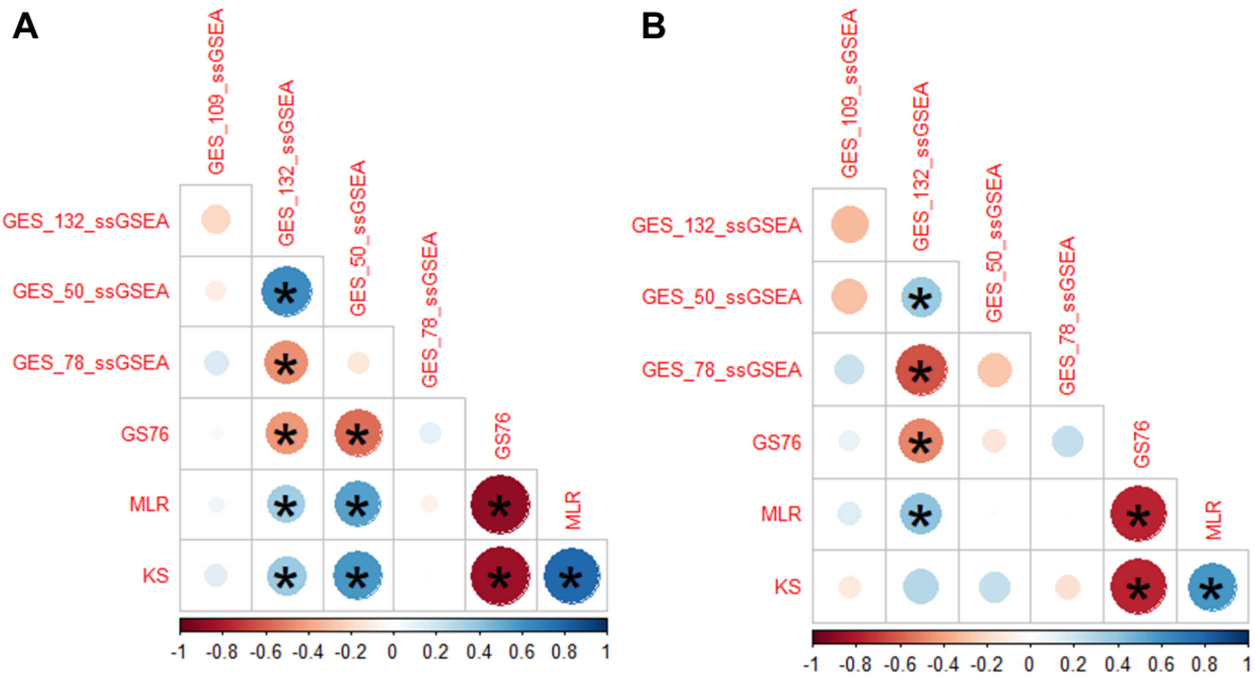


Fig S6: ssGSEA scores and EMT score Spearman's correlation in GSE22597. A) IBC **B)** nIBC

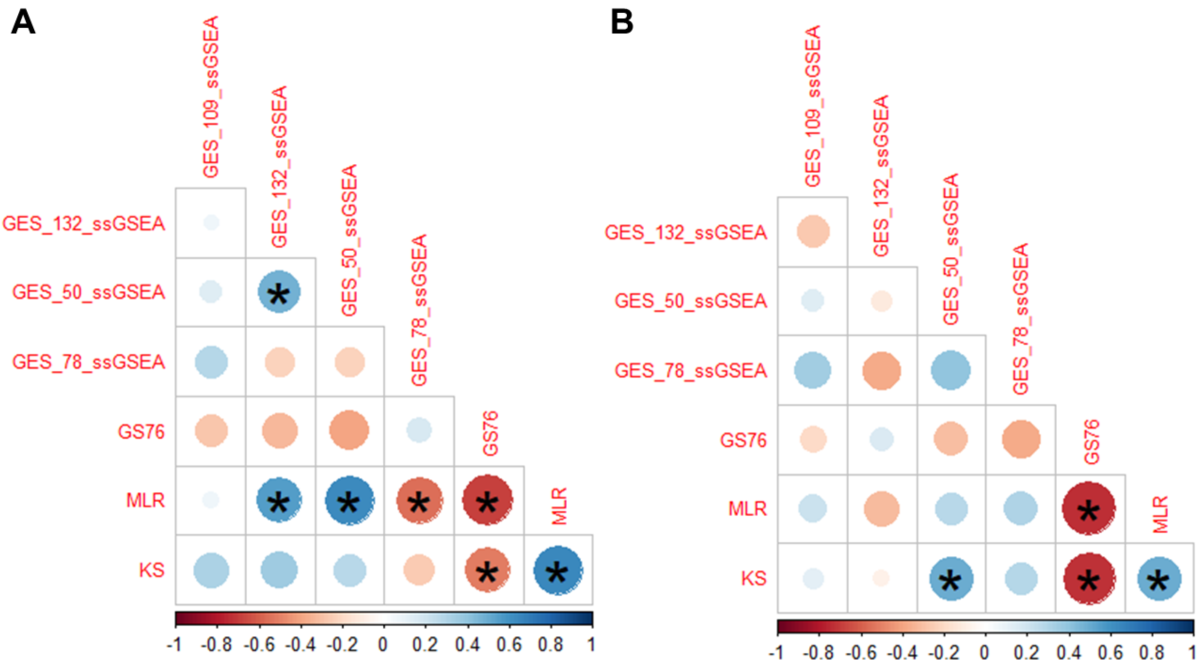


Fig S7: ssGSEA scores and EMT score Spearman's correlation in GSE45581.(A) IBC **(C)** nIBC

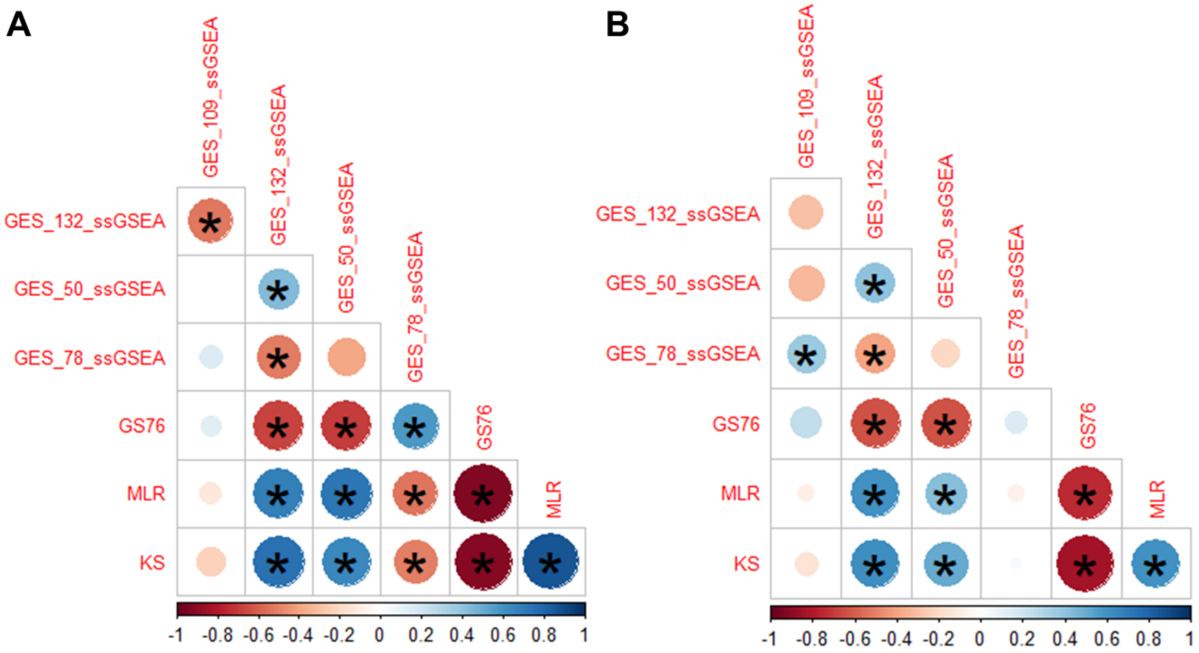


Fig S8: ssGSEA scores and EMT score Spearman's correlation in GSE5847. A) IBC **B)** nIBC

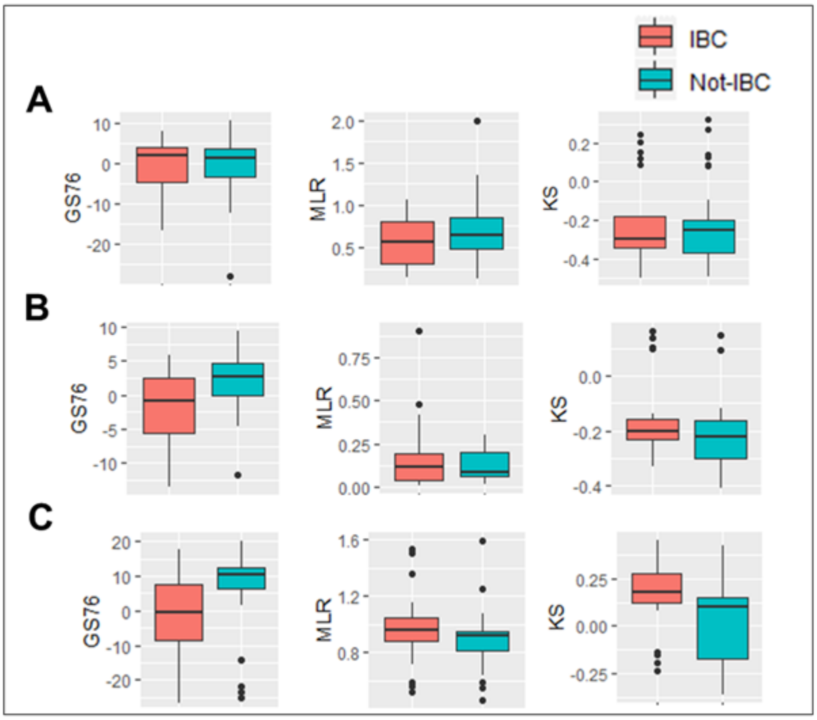


Fig S9: EMT scores across IBC and nIBC groups. A) GSE22597 **B)** GSE45581 **C)** GSE5847

Supplementary table legends
S1: LR top 2000 predictors
S2: EMT scores
S3: Clustering accuracy of different combination of variables