npj | Breast Cancer

www.nature.com/npjbcancer

ARTICLE OPEN

Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes

James C. Mathews[1], Saad Nadeem[1], Arnold J. Levine[2], Maryam Pouryahya[1], Joseph O. Deasy[1] and Allen Tannenbaum[3]

We introduce a classification of breast tumors into seven classes which are more clearly defined by interpretable mRNA signatures along the PAM50 gene set than the five traditional PAM50 intrinsic subtypes. Each intrinsic subtype is partially concordant with one of our classes, and the two additional classes correspond to division of the classes concordant with the Luminal B and the Normal intrinsic subtypes along expression of the Her2 gene group. Our Normal class shows similarity with the myoepithelial mammary cell phenotype, including TP63 expression (specificity: 80.8% and sensitivity: 82.8%), and exhibits the best overall survival (89.6% at 5 years). Though Luminal A tumors are traditionally considered the least aggressive, our analysis shows that only the Luminal A tumors which are now classified as myoepithelial have this phenotype, while tumors in our luminal class (concordant with Luminal A) may be more aggressive than previously thought. We also find that patients with basal tumors surviving to 48 months exhibit favorable continued survival rates when certain markers for B lymphocytes are present and poor survival rates when they are absent, which is consistent with recent findings.

npj Breast Cancer (2019)5:30 ; https://doi.org/10.1038/s41523-019-0124-8

INTRODUCTION

Multiparametric genetic tests such as the PAM50/Prosigna Risk of Recurrence (ROR) for breast cancer prognostication are becoming commonplace.[1,2] However, due to limited accuracy and poor concordance with biological phenotypes, their clinical utility is still under investigation.[3] In this paper, we address these issues in the context of one of the most prevalent assays, the PAM50 ROR, which is mainly driven by an intrinsic subtype classification along a 50-gene mRNA expression profile. We reclassify these profiles using topological data analysis, incorporating prior knowledge of biological phenotype (basal/luminal stratification). Unlike the five traditional PAM50 intrinsic subtypes, our seven classes are accurately defined by clear patterns of activation and inactivation of gene groups directly interpretable in terms of specific normal mammary cell types: basal, luminal/ER, myoepithelial, and Her2-related gene groups.

The basal/luminal terminology refers to mammary cell differentiation from basal–epithelial cells near the basement membrane to the more differentiated luminal–epithelial cells near the lumen or ducts. It was the basis for the systematic molecular classification of breast cancer initiated by Perou et al.[4] Myoepithelial refers to a mammary cell type playing a key role in breast duct secretion.[5,6] Overexpression of Her2 (ERBB2) and a group of related genes marks the Her2 + cohort well-known since the 1990s for highly favorable response to the drug trastuzumab (herceptin). Figure 1 summarizes the history of the molecular classification and our contribution. Table 1 lists the new classes.

RESULTS

Clearly defined 50-gene signatures

The signature classes we defined show partial concordance with the PAM50 subtypes, with a Normalized Mutual Information (NMI) of 0.19 (29.1 times the maximum NMI found in 10,000 random permutation bootstrapping trials) (Fig. 2). However, our classes show tighter clustering along the 50-gene profile: the k-mean for the PAM50 subtypes is 87.9% of the total variance, and for our classification is only 82.7% (both using the L1 norm). To assess the quality of the signatures themselves, we consider the average silhouette width[7] of each class. The silhouette width is the average distance $a(i)$ between a sample i and the cluster to which it belongs subtracted from the smallest average distance $b(i)$ between i and the other clusters, normalized by max $(a(i), b(i))$. The average silhouette width over a given cluster (abbreviated SW) measures the tightness of the cluster with respect to the clustering scheme, with larger SW (closer to 1) indicating a good cluster and smaller SW (closer to −1) indicating a poor cluster.

Our Luminal class SW = 0.151 is greater than the PAM50 Luminal A SW by 0.107; Luminal/Basal SW = 0.131 is greater than the PAM50 Luminal B SW by 0.112; Myo/Luminal SW = 0.0422 is greater than the PAM50 Normal SW by 0.0432 (silhouette widths range from −1 to 1). The SWs of our Her2 and Basal/Myo SWs are very close to the SW of the PAM50 Her2 and Basal subtypes.

As shown in Fig. 2, the main example of a clear new signature is the heterogeneous expression of the myoepithelial gene group in the PAM50 Luminal A subtype, resolved by division into Luminal and Myo/Luminal classes. One exception is that the Basal/Her2

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; [2]Institute for Advanced Study, School of Natural Sciences, Princeton, NJ 08540, USA and [3]Departments of Computer Science & Applied Mathematics, Stony Brook University, Stony Brook, NY 11794, USA
Correspondence: Allen Tannenbaum (allen.tannenbaum@stonybrook.edu)

Received: 17 April 2019 Accepted: 5 August 2019
Published online: 09 September 2019

Published in partnership with the Breast Cancer Research Foundation

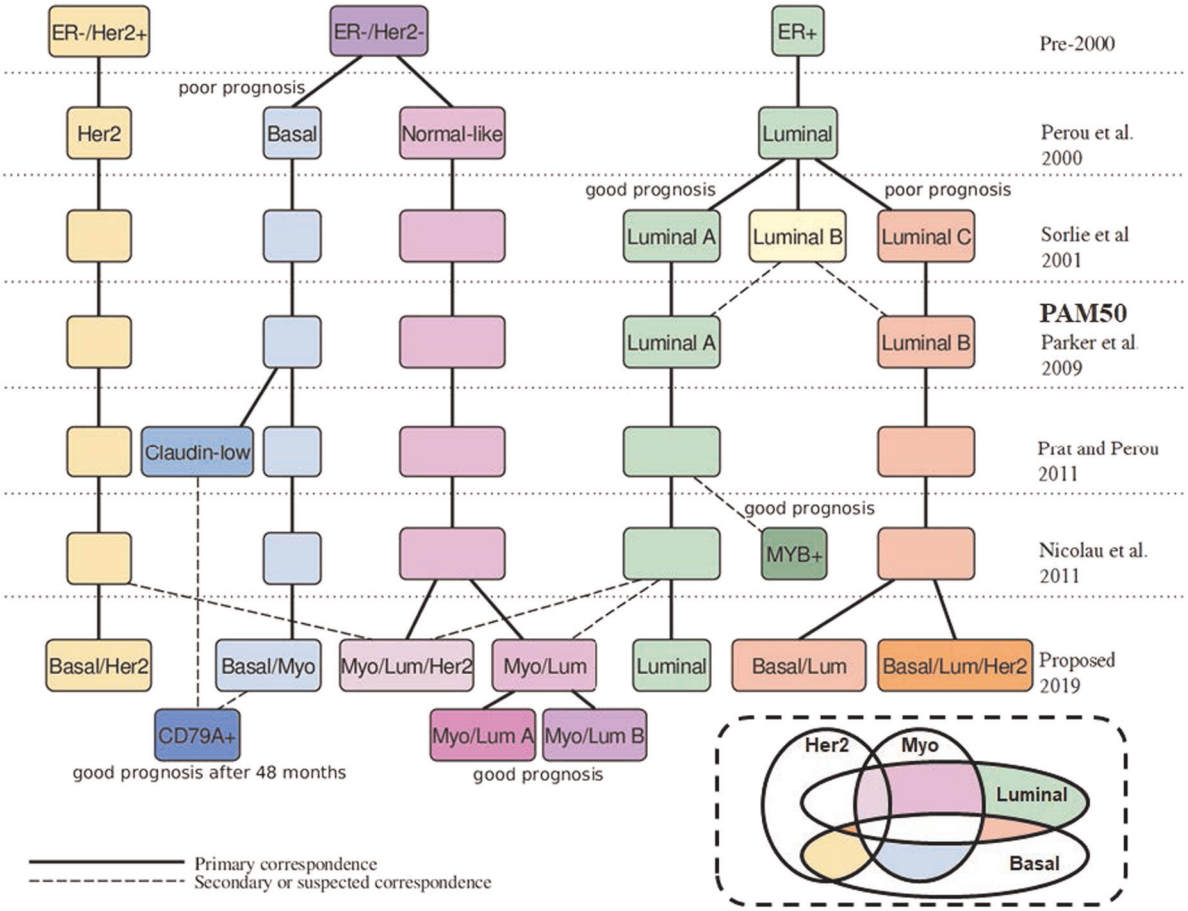


Fig. 1 History of the molecular classification of breast cancer. Names are shown at the chronological level at which they were introduced. The Her2+ breast tumors were already well-known in the 1990s for highly favorable response to the drug trastuzumab (herceptin), which was approved by the FDA for metastatic Her2+ breast cancer in 1998. The hierarchical clustering of Perou et al.[4] used genes whose expression differentiates between samples from different tumors better than between samples from the same tumor, finding four main classes: ERBB2+ (or Her2+), Basal, Luminal, and Normal breast-like. Sorlie et al.[26] explicitly incorporated clinically relevant outcome data such as overall survival, uncovering three Luminal subtypes, Luminal A, B, and C. Luminal A has higher overall survival than Luminal B, and Luminal B has higher overall survival than Luminal C. Later investigators found only two Luminal subtypes to be sufficiently robust. Parker et al.[27] introduced the 50 gene set that became known as the PAM50 (Prediction Analysis of Microarray) and introduced a straightforward centroid-based classifier for breast tumor RNA expression patterns along the PAM50 with five classes: Basal, Her2, Luminal A, Luminal B, and Normal. The authors used this classification as a key component in the model that became the Prosigna predictor of Risk Of Relapse (ROR). Prat and Perou[9] introduced the Claudin-low subtype carved largely out of the Basal group. The authors find that the Claudin-low subtype has poor prognosis compared to Luminal A, but no worse than the other subtypes. The Topological Data Analysis of Nicolau et al.[17] confirmed the distinction between more luminal, more basal, and more normal-like subtypes along branches of a graph structure modeling the distribution of breast tumor samples. They found a subgroup of patients exhibiting a very high survival rate, largely characterized by expression of MYB. Our proposed classification uses the method of Nicolau et al.[17] and incorporates gene sets and priors (e.g., the basal-to-luminal stratification) known to be relevant to breast cancer biology. (Below right) Our proposed system with seven classes defined by four elementary phenotypes (see also Figs 2, 7)

Table 1. Class signatures for reclassification of PAM50 subtypes of breast tumors

Basal genes	+	+		+	+		
Myoepithelial genes	+					+	+
Luminal genes			+	+	+	+	+
Her2 genes		+			+		+
Primary overlapping PAM50 subtypes	Basal	Her2	LumA	LumB	LumB	Normal	Normal

The genes in each gene group are shown in Fig. 2

class binds together the PAM50 Her2 with several PAM50 Luminal B samples. However, the Luminal B here clearly differ from the Her2 by the presence of Luminal markers, so to address this we divide this class into Basal/Her2 and Basal/Her2/Luminal. Also, the two myoepithelial gene groups are small and closely related, so we merge them together into a single myoepithelial group and accordingly merge the classes denoted Myo/Luminal A and Myo/Luminal B. The seven resulting signatures are shown in Table 1.

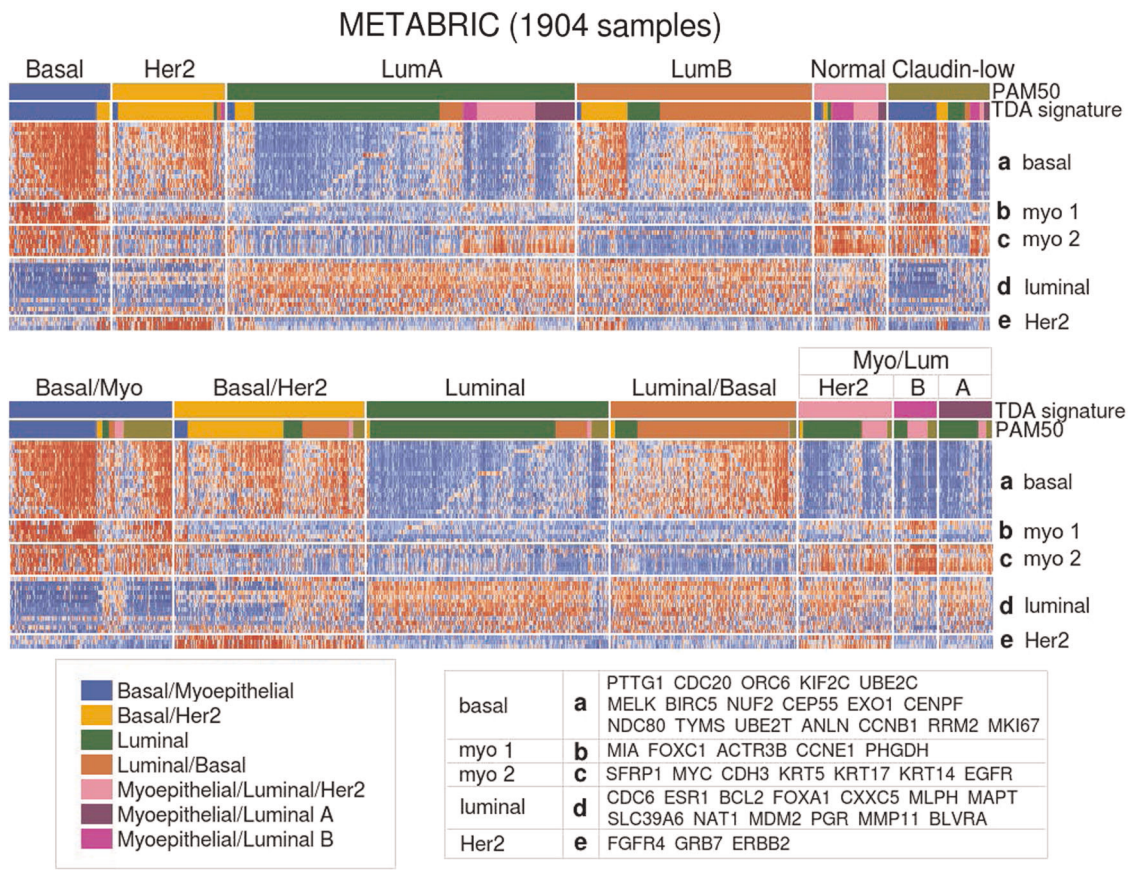


Fig. 2 RNA expression heatmap of the 1904 METABRIC breast tumor samples. (Above) Organized first by PAM50 subtype and then by the TDA signatures classes assigned by the Mapper-derived classifier along the PAM50 gene set (BAG1, MYBL2, GPR160, and TMEM45B omitted due to missing values). (Below) Organized first by TDA signature class then by PAM50 subtype

Myo/Luminal class with good survival

The Kaplan–Meier survival analysis of the new classes is shown in Fig. 3 for both 1904 METABRIC and 1082 TCGA samples (Figs 3–5). The plots show that the Myo/Luminal class exhibits the greatest survival rate, even greater than PAM50 Luminal A (the log-rank test for statistically significant difference between Normal and Myo/Luminal survival curves yields $p = 0.003$). Many of the Myo/Luminal tumors are designated PAM50 Luminal A, and since the Luminal A subtype is already the one with the best prognosis in the PAM50 scheme, we conclude that the Myo/Luminal class preferentially selects from Luminal A subtype the patients with especially good prognosis even among Luminal A.

The Myo/Luminal and Myo/Luminal/Her2 subtypes have signatures with the most new features. Kaplan–Meier analysis shows that the Myo/Luminal A (FOXC1-/MIA-/PHGDH-) phenotype has the best prognosis of all, with 93% survival at 5 years (Fig. 4).

To investigate the Myo/Luminal class further, we drew upon the classification of normal mammary cell types of Santagata et al.[5] in terms of marker genes/proteins ESR1, AR, VDR, KRT5, MKI67, KRT18, MME, SMN1, and TP63. Figure 5 shows the Mapper analysis of the 290 normal breast tissue samples of the GTEx RNA expression database.[8] We found normal tissue expression patterns were similar to one of our class' signatures along the PAM50, and also similar to one of the cell type patterns of Santagata et al.[5] along their marker genes. One of the clearest patterns was activation of only the basal gene group along the normal cell type denoted L1, characterized by expression of the proliferation marker MKI67. In addition, a clear subset of samples, displaying a superposition of the pattern of normal myoepithelial cell-type M2 and normal cell-type L7 (KRT5+/VDR+), also displayed the signature Myo/Luminal/Her2. The main characteristic of M2 is expression of TP63. We found that TP63 expression can be used as a single marker for the Myo/Luminal class (specificity: 80.8%, sensitivity: 82.8%).

Basal/myoepithelial (triple-negative) subclass with immune-related survival advantage

Since the Myo/Luminal class is heterogeneous with respect to FOXC1, MIA, and PHGDH expression, we expected that FOXC1+/MIA+/PHGDH+ would be associated with a more aggressive phenotype (Fig. 6). After all, these genes are highly expressed in the PAM50 Basal subtype (Basal/Myo). We found that while this is true for the first 48 months after diagnosis, the FOXC1+, MIA+, and PHGDH+ phenotypes all showed very favorable survival rates contingent on survival to 48 months (Fig. 6). We hypothesized that this phenomenon might generalize to the PAM50 Basal subtype. To test this, we sought genes from the set of 18,543 genes available for the METABRIC cohort which would separate the long-term and short-term survivors in the FOXC1+/MIA+/PHGDH+ group. The 100 most significant genes with respect to the *t* test for difference of mean expression ($-\log_{10}(p)$ value > 6.7) included the genes coding for the B-cell antigen receptor complex-associated protein alpha and beta chains, the B-cell-specific coactivator OBF-1, the pre-B-lymphocyte-specific protein-2, and B-cell maturation factor (CD79A, CD79B, POU2AF1, IGLL1, and TNFRSF17), as well as CD38, expressed by many immune cells. (In fact, CD79A is one of the major positive expression markers for the Claudin-low subtype introduced by Prat and Perou.[9] The Claudin-low subtype and our CD79A+/CD38+/IGLL1+ type are both subgroups of the Basal group).

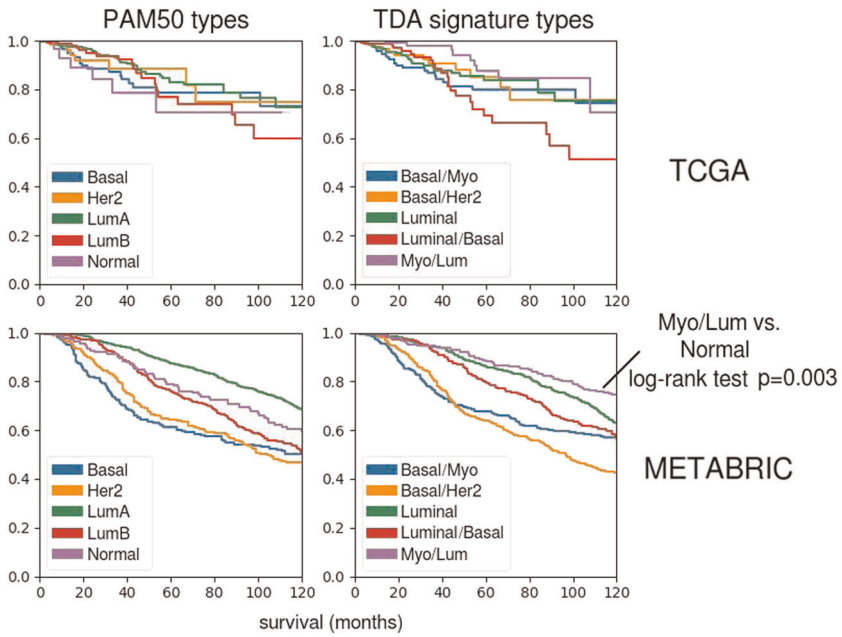


Fig. 3 Kaplan–Meier survival analysis of the subgroups of the TCGA and METABRIC cohorts defined by PAM50 subtypes and the major corresponding TDA signature classes. The Myo/Luminal class has the highest survival rate, statistically significantly greater than the primary corresponding PAM50 subtype, the Normal subtype. In the TCGA data set, the log-rank test for PAM50 Normal versus Myo/Luminal yields $p = 0.023$, while in the METABRIC data set (with approximately twice as many samples) the test yields $p = 0.003$

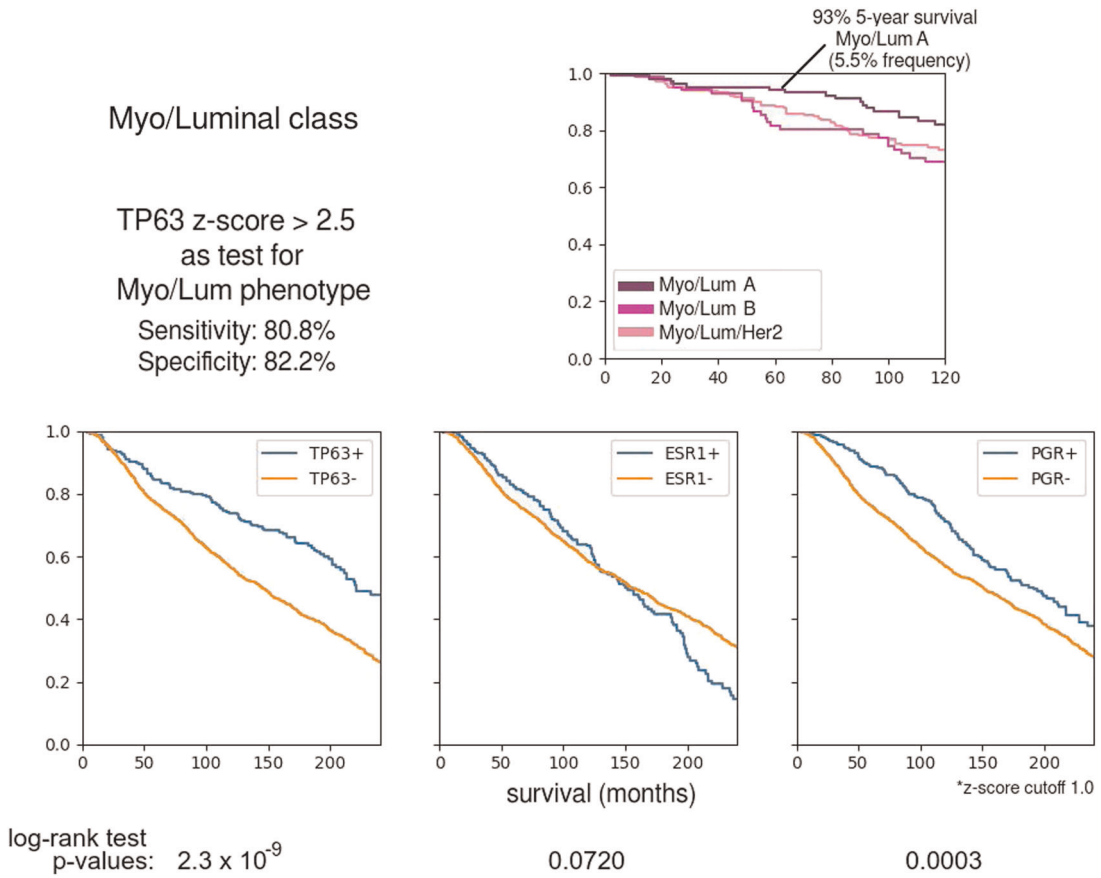


Fig. 4 Stratification of the Myo/Luminal class by three TDA signature classes. The Kaplan–Meier plot shows slightly different survival rates, with Myo/Luminal A having the best prognosis; better than PAM50 Luminal A. TP63 expression (a known myoepithelial marker; see Fig. 5) somewhat robustly defines the Myo/Luminal class. Kaplan–Meier survival analysis plots are shown comparing the survival probabilities between TP63+ and TP63− phenotypes across the whole METABRIC cohort. TP63+ confers a survival advantage comparable to that of PGR+

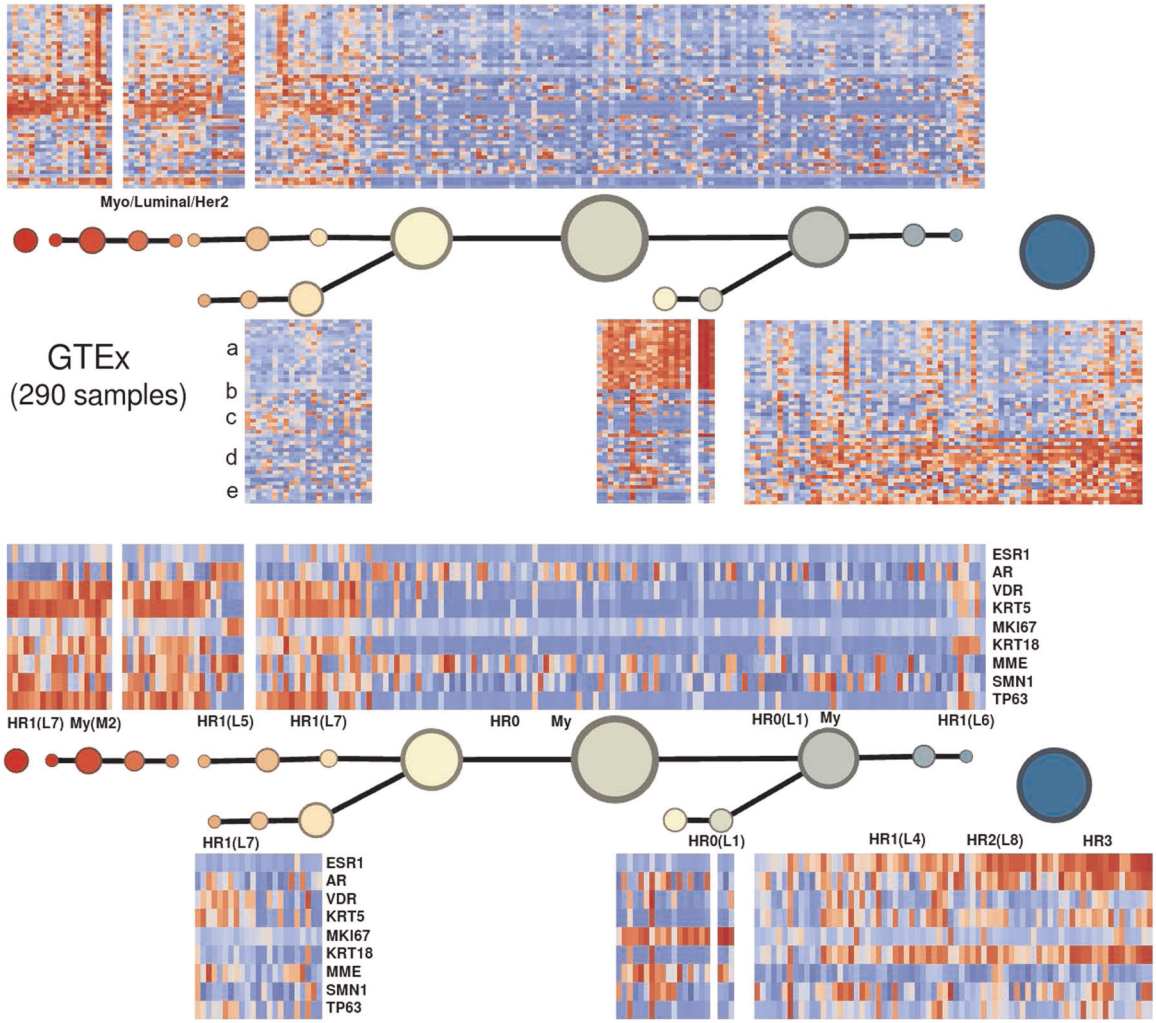


Fig. 5 Mapper analysis of the 290 GTEx normal mammary tissue samples using the basal–luminal score as filter function. (Above) Along the PAM50 gene set, using the basal–luminal score as filter function. (Below) The same sample set, in the same order, showing the expression of the marker genes of Santagata et al.,[5] which define the normal mammary cell-type classification proposed by those authors. A substantial group displays the Myo/Luminal/Her2 phenotype. According to the Santagata et al. classification, these samples are primarily a combination of the myoepithelial type M2 (TP63+/KRT5+) and the luminal–epithelial type L7 (VDR+/KRT5+)

Figure 6 shows that expression of each of CD79A, CD38, and IGLL1 strongly stratifies the Basal tumors into a poor prognosis group and another group with much better prognosis after 48 months.

DISCUSSION

Only certain combinations of the elementary phenotypes we identified, Basal, Luminal, Myoepithelial, and Her2, are observed in breast tumors. For example, the Luminal/Basal, Basal/Myo, and Myo/Luminal are all observed, but the combination Luminal/Basal/Myo is not. We conclude that in the tumor development process, the activation of any two of the Luminal, Basal, and Myoepithelial gene groups precludes the further activation of the third.

Some of the genes in the new Myoepithelial gene group (denoted b in Fig. 2) concurrently stratify the Myo/Lum class. These include FOXC1, MIA, and PHGDH. The protein product of PHGDH, phosphoglycerate dehydrogenase, is a key enzyme participating in biosynthesis of serine. Maddocks et al.[10] find that functioning p53 is required for complete activation of the serine synthesis pathway in human cancer cells. Since the Myo/Luminal tumors have a very low TP53 mutant rate of only 15.6% in comparison to 78% for Basal/Myo, the Myo/Luminal tumors, with functioning p53, are probably capable of synthesis of serine in response to serine starvation. Only the Myo/Luminal B subclass of Myo/Luminal actually expresses PHGDH, suggesting serine synthesis and metabolism. Since Myo/Luminal A exhibits better survival rates than Myo/Luminal B, our findings are consistent with the results of Labuschagne et al.[11] and Amelio et al.[12] implicating serine metabolism in promoting tumor growth.

TP63 is one of the myoepithelial markers in the work of Santagata et al.[5] and also a key marker for our Myo/Luminal class. From the Kaplan–Meier analysis in Fig. 4, we conclude that TP63 expression confers a survival advantage even greater than the well-known survival advantage conferred by PGR expression across the whole METABRIC cohort.

The PAM50 subtype most resembling the Myo/Luminal class is Normal-like. The status of the Normal-like subtype has been uncertain since its introduction by Perou et al.[4] It is often thought to represent non-cancer tissue which is incidentally present in bulk tissue samples. For example, the PAM50 classifier uses actual normal tissue samples to train the centroid of the Normal class.

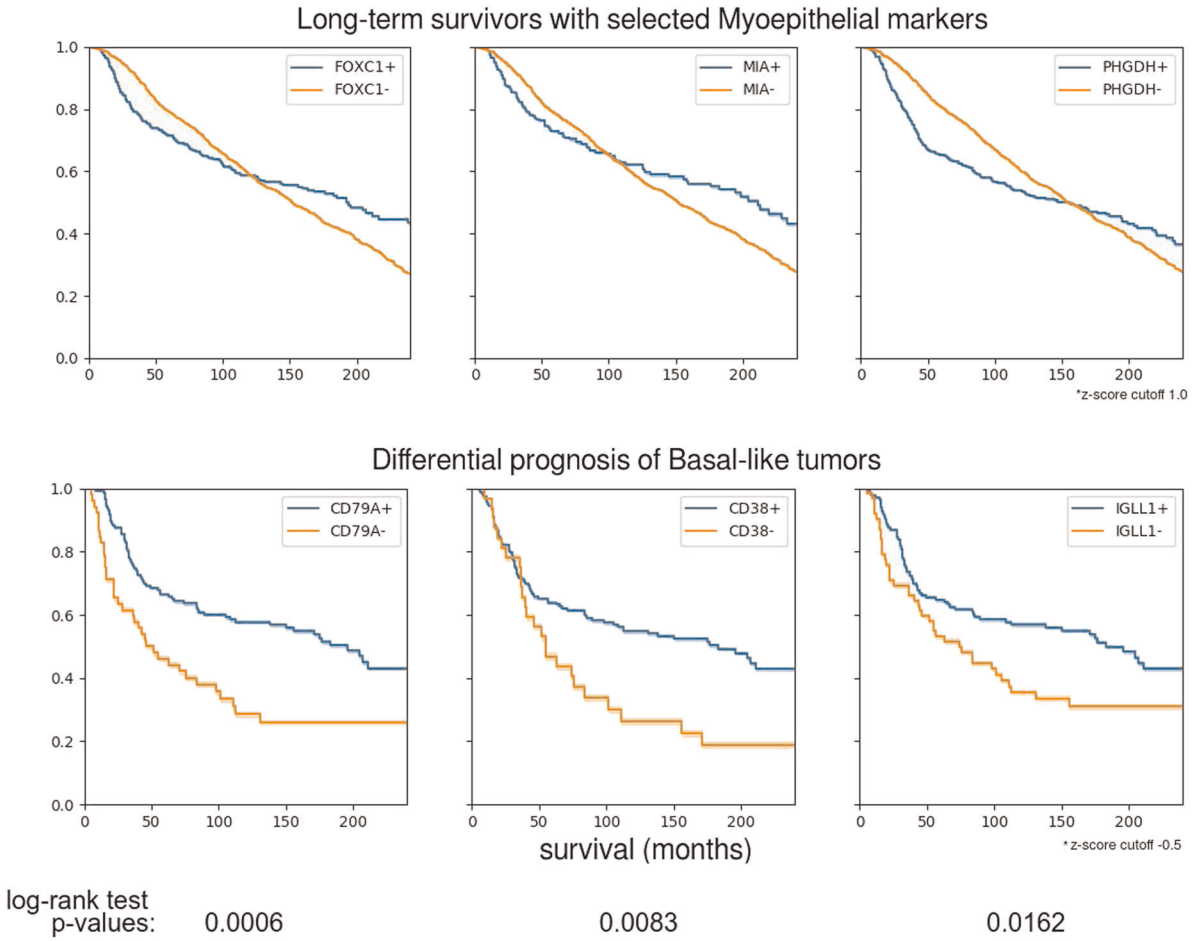


Fig. 6 Survival analysis of FOXC1+/MIA+/PHGDH+ and CD79A+/CD38+/IGLL1+ phenotypes. (Above) The FOXC1+/MIA+/PHGDH+ phenotype, observed in the Myo/Luminal B class but not the Myo/Luminal A class, confers a survival disadvantage for approximately the first 48 months after diagnosis, and a survival advantage afterwards. (Below) Of the top 100 genes out of 18,543 exhibiting statistically significant mean differences between the FOXC1+/MIA+/PHGDH+ short-term and long-term survivors, several are B-lymphocyte-related, including: CD79A (immunoglobulin-alpha), CD38, and IGLL1 (immunoglobulin lambda-like polypeptide 1). FOXC1+/MIA+/PHGDH+ is also observed in the PAM50 Basal subtype. Within the Basal subtype, CD79A+, CD38+, and IGLL1+ confer a significant survival advantage after 48 months

However, in our analysis all of the classes of breast cancer show similarity to some combination of normal mammary cell types.

We found a significantly lower death rate after 4 years for patients with basal tumors expressing key B-lymphocyte-related markers CD79A, CD79B, POU2AF1, IGLL1, and TNFRSF17. This group is 80.3% of all patients with basal tumors surviving to 4 years. We conclude that the remaining 19.7% of these patients, with basal tumors lacking these markers, are still at high risk of mortality. This observation is consistent with the finding of Rueda et al.[13] that a certain subgroup of triple-negative breast cancers can be defined which rarely recurs after 5 years.

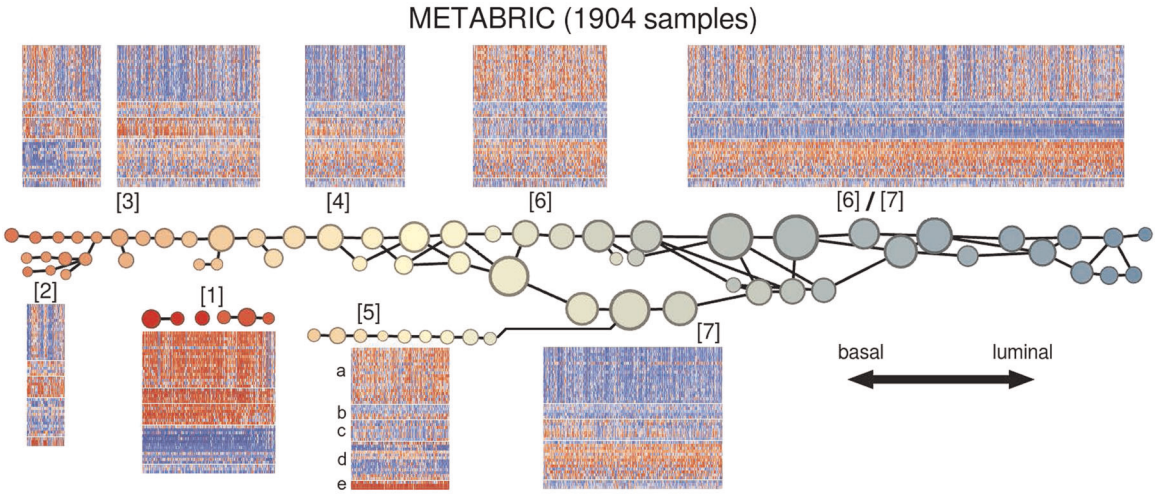
Regarding future work, responses to specific drugs or therapies should be investigated to decide whether some patients with Luminal but not Myo/Luminal tumors are under treated. Moreover, future work should address the question of why the four main gene groups appear. One possible explanation is that the four prototypical expression patterns Luminal, Basal, Myoepithelial, and Her2-related represent types of clones derived from an original transformation, and the combinations of these prototypes correspond to a certain clonal mixture. Another possibility is that the observed expression patterns are superpositions of actual tumor expression, expression of tumor microenvironmental normal cells with types related to the four prototypes, or expression patterns similar to original normal ancestor cells. New techniques of single-cell sequencing, potentially in conjunction with tumor-level spatial mapping, may provide answers to these questions.

Finally, the differential prognosis among triple-negative tumors observed with respect to the B-lymphocyte-related stratification suggests that the immune systems of ~51% of patients with triple-negative tumors can naturally and reliably mount a successful response to the tumor. If this hypothesis is correct, a longitudinal study monitoring the immune system of triple-negative patients should be able to discover exactly what response is mounted, which could lead to potential new therapies that induce this natural response.

METHODS

Topological data analysis

Topological data analysis (TDA) methods, employing ideas from the mathematical field of topology, have gained popularity in recent years. More precisely, discrete algorithmic counterparts of topological concepts have emerged in response to the availability of large data sets harboring hidden structures. Mapper,[14] a discrete analogue of a Morse-theoretic analysis of a manifold with respect to a height function, or "filter" function, has received particular attention with regards to both its theoretical foundations[15,16] and, following Nicolau et al.,[17] its application to cancer



	Unsupervised signature	gene group a	b	c	d	e
Basal/Myo	[1]	+	+	+	-	-
Myo/Luminal/Her2	[2]	-		+		+
Myo/Luminal B	[3]	-	+	+		
Myo/Luminal A	[4]	-	-	+		
Her2/Basal	[5]	+			-	+
Basal/Luminal	[6]	+	-	-	+	-
Luminal	[7]	-	-	-	+	-

Fig. 7 Mapper analysis of the 1904 METABRIC breast tumor samples, along the PAM50 gene set, using the basal–luminal score as filter function. The circular nodes represent clusters in the strata or bins defined by the filter function at the chosen level of resolution. For example, there are three clusters in the stratum shown in yellow; two clusters shown higher and labeled with unsupervised signature number [4], and one cluster shown lower labeled with unsupervised signature number [5]. All three have the same basal–luminal score range, indicated by color. In Table 1, the salient signatures are recorded. These signatures differ slightly, in two ways, from the seven classes we finally propose as in Fig. 1. First, for the sake of simplicity we merge the two myoepithelial–related gene groups (**b**, **c**) into a single-gene group, consequently merging Myo/Luminal A and Myo/Luminal B into Myo/Luminal. Second, on account of the salient signatures observed in the heatmaps in Fig. 2, we split Her2/Basal [5] into Her2/Basal and Luminal/Basal/Her2. Where blanks appear, the corresponding gene group (**a**–**e**) is neither uniformly positively nor uniformly negatively expressed

genomics.[18–20] Mapper builds a graphical summary of a given sample set with respect to a chosen stratification (filter) function.

We use three sample sets: TCGA, METABRIC,[21,22] and GTEx.[8] The 1082 TCGA and 1904 METABRIC mRNA expression *z*-score data sets along the PAM50 gene set were retrieved from cBioPortal.[23,24] The 290 GTEx normal breast data set was downloaded from the GTEx portal; metadata supporting data files may be found in.[25]

Due to the retrospective nature of this study using only publicly available data, ethics approval for the study was not required.

Filter function

The "filter function" or initial stratification is taken to be a basal–luminal epithelial differentiation score, calculated as the average expression *z*-score of luminal–epithelial markers (XBP1, FOXA1, GATA3, ESR1, and ANXA9) minus the average expression *z*-score of basal–epithelial markers (KRT17, KRT5, DST, ITGB4, LAMC2, CDH3, LAD1, and ITGA7). Selected largely on the basis of Perou et al.,[4] the basal markers are all associated with anchorage of epithelial cell layers to the basement membrane, while the luminal markers are all expressed in well-differentiated or mature luminal epithelial cells.

The Mapper graph and 50-gene signatures determined from the METABRIC breast tumor samples are shown in Fig. 7. Correlation-based clustering along small contiguous subsets with respect to the graph yielded the five main gene groups.

A simple classifier was constructed from the table of observed signatures (see Fig. 7) as follows: For a given sample and a given signature or profile, the average values for each gene group are calculated, then added together with the signature signs as weights. The resulting number is a similarity score between the sample and the signature. The sample is assigned to the highest-scoring signature.

Finally, the classes and gene groups shown in Fig. 2 were adjusted: The two myoepithelial gene groups were merged, the Myo/Luminal A and Myo/Luminal B classes were merged as a result, and Luminal expression was used to delineate classes Basal/Her2 and Basal/Luminal/Her2.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data generated and analysed during this study are publicly available in GitHub as described in the following data record: https://doi.org/10.6084/m9.figshare.9199289.[25] The study used three publicly available datasets and the raw data can be accessed from cBioPortal at https://identifiers.org/cbioportal:brca_metabric (METABRIC data) and at https://identifiers.org/cbioportal:brca_tcga_pan_can_atlas_2018 (TCGA data). The raw normal breast data can be accessed from the GTEx portal at https://gtexportal.org/home/datasets.[8,21,22]

CODE AVAILABILITY

The code written in Python and R is available upon request. The analysis methodology is described in detail in the Supplementary Information file.

ACKNOWLEDGEMENTS

This study was supported by AFOSR grant (FA9550-17-1-0435), NIA grant (R01-AG048769), MSK Cancer Center Support Grant/Core Grant (P30 CA008748), and a grant from Breast Cancer Research Foundation (grant BCRF-17-193). The article was previously published as a preprint: Ref. Mathews, J. C. et al. Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial andimmune phenotypes. Preprint at: https://doi.org/10.1101/480723v3 (2019).

AUTHOR CONTRIBUTIONS

J.C.M. performed research and analyzed data. J.C.M. and S.N. drafted the paper. A.L. contributed a number of key suggestions for the present version. S.N., A.L., M.P., J.O. D. and A.T. edited the paper. A.T. and J.O.D. directed the research.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Breast Cancer* website (https://doi.org/10.1038/s41523-019-0124-8).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Duffy, M. et al. Clinical use of biomarkers in breast cancer: updated guidelines from the european group on tumor markers (egtm). *Eur. J. Cancer* **75**, 284–298 (2017).
2. Coates, A. S. et al. Tailoring therapies–improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* **26**, 1533–1546 (2015).
3. Untch, M. et al. Primary therapy of patients with early breast cancer: evidence, controversies, consensus. *Geburtshilfe und Frauenheilkd.* **75**, 556–565 (2015).
4. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
5. Santagata, S. et al. Taxonomy of breast cancer based on normal cell phenotype predicts outcome. *J. Clin. Invest.* **124**, 859–870 (2014).
6. Gudjonsson, T., Adriance, M. C., Sternlicht, M. D., Petersen, O. W. & Bissell, M. J. Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia. *J. Mammary Gland Biol. Neoplasia* **10**, 261–272 (2005).
7. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
8. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
9. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* **5**, 5–23 (2011).
10. Maddocks, O. D. et al. Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. *Nature* **493**, 542–546 (2013).
11. Labuschagne, C. F., van den Broek, N. J., Mackay, G. M., Vousden, K. H. & Maddocks, O. D. Serine, but not glycine, supports one-carbon metabolism and proliferation of cancer cells. *Cell Rep.* **7**, 1248–1258 (2014).
12. Amelio, I., Cutruzzola, F., Antonov, A., Agostini, M. & Melino, G. Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* **39**, 191–198 (2014).
13. Rueda, O. M. et al. Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups. *Nature*. https://doi.org/10.1038/s41586-019-1007-8 (2019).
14. Singh, G., Memoli, F. & Carlsson, G. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics* (eds. Botsch, M., Pajarola, R., Chen, B. & Zwicker, M.) (The Eurographics Association, 1–11 2007).
15. Carrière, M. & Oudot, S. Structure and stability of the one-dimensional mapper. *Found. Comput. Math.* https://doi.org/10.1007/s10208-017-9370-z (2017).
16. Dey, T. K., Mémoli, F. & Wang, Y. Multiscale Mapper: Topological Summarization via Codomain Covers. In: Krautgamer, R. (ed.) *Proc. Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, 997–1013 (SIAM Publications, Philadelphia, PA, 2016).
17. Nicolau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl Acad. Sci. USA* **108**, 7265–7270 (2011).
18. Lum, P. Y. et al. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **3**, 1236 (2013).
19. Lockwood, S. & Krishnamoorthy, B. Topological features in cancer gene expression data. Preprint at: http://arxiv.org/abs/1410.3198v1 (2014).
20. Jeitziner, R. et al. Two-tier mapper: a user-independent clustering method for global gene expression analysis based on topology. Preprint at: https://arxiv.org/pdf/1801.01841.pdf (2017).
21. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
22. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
23. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).
24. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Disco.* **2**, 401–404 (2012).
25. Mathews, J. C. et al. Metadata supporting data files of the related article: robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes. *figshare*. https://doi.org/10.6084/m9.figshare.9199289 (2019).
26. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
27. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).