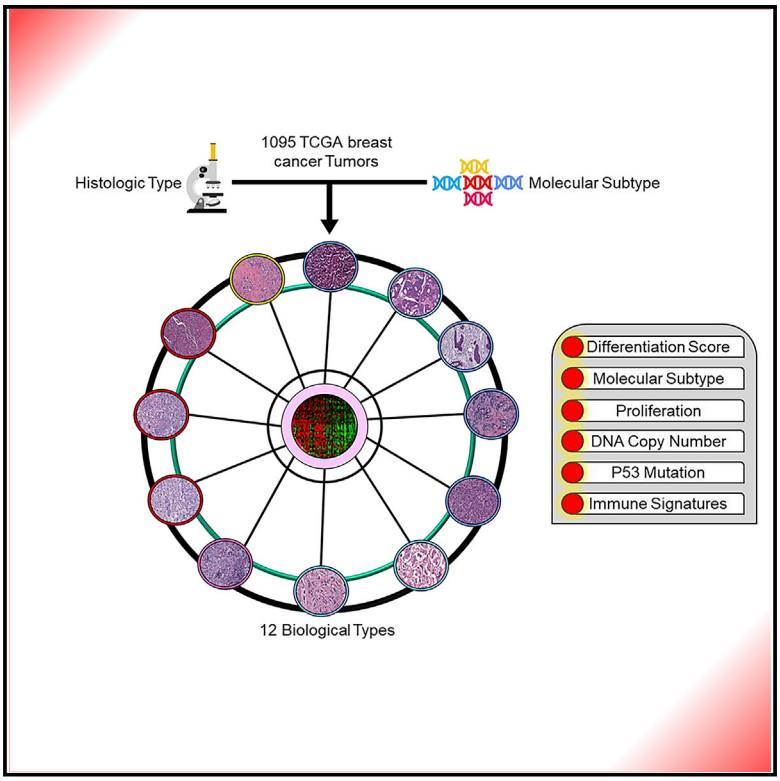
Cell Genomics

Resource

Molecular analysis of TCGA breast cancer histologic types

Graphical abstract



Authors

Aatish Thennavan, Francisco Beca, Youli Xia, ..., Katherine A. Hoadley, Andrew Beck, Charles M. Perou

Correspondence

cperou@med.unc.edu

In brief

Thennavan et al. describe the histologic type diagnostic annotations for the complete TCGA breast cancer cohort and report gene features of six rare histological types of breast cancer. They present an integrated model of breast cancer using histology and clinically relevant breast cancer genomic features.

Highlights

- The complete histologic annotation for all 1,095 TCGA primary breast cancer patient samples
- Genomic and molecular signatures of six rare breast cancer histologic types
- Utility of constructed mucinous gene signature in pan-cancer mucinous cancers
- Genomic- and histology-integrated model classifies 12 consensus groups

Thennavan et al., 2021, Cell Genomics *1*, 100067
December 8, 2021
https://doi.org/10.1016/j.xgen.2021.100067

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Resource

Molecular analysis of TCGA breast cancer histologic types

Aatish Thennavan,[1,2] Francisco Beca,[3] Youli Xia,[2,4] Susana Garcia-Recio,[2,4] Kimberly Allison,[3] Laura C. Collins,[5] Gary M. Tse,[6] Yunn-Yi Chen,[7] Stuart J. Schnitt,[8] Katherine A. Hoadley,[2,4] Andrew Beck,[9] and Charles M. Perou[2,4,10,11,*]

[1]Oral and Craniofacial Biomedicine Program, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[5]Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA
[6]Department of Anatomical and Cellular Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
[7]Department of Pathology and Laboratory Medicine, University of California, San Francisco, San Francisco, CA 94143, USA
[8]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School and Breast Oncology Program, Dana-Farber/Brigham and Women's Cancer Center, Boston, MA 02115, USA
[9]PathAI, Boston, MA, USA
[10]Department of Pathology & Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[11]Lead contact
*Correspondence: cperou@med.unc.edu
https://doi.org/10.1016/j.xgen.2021.100067

SUMMARY

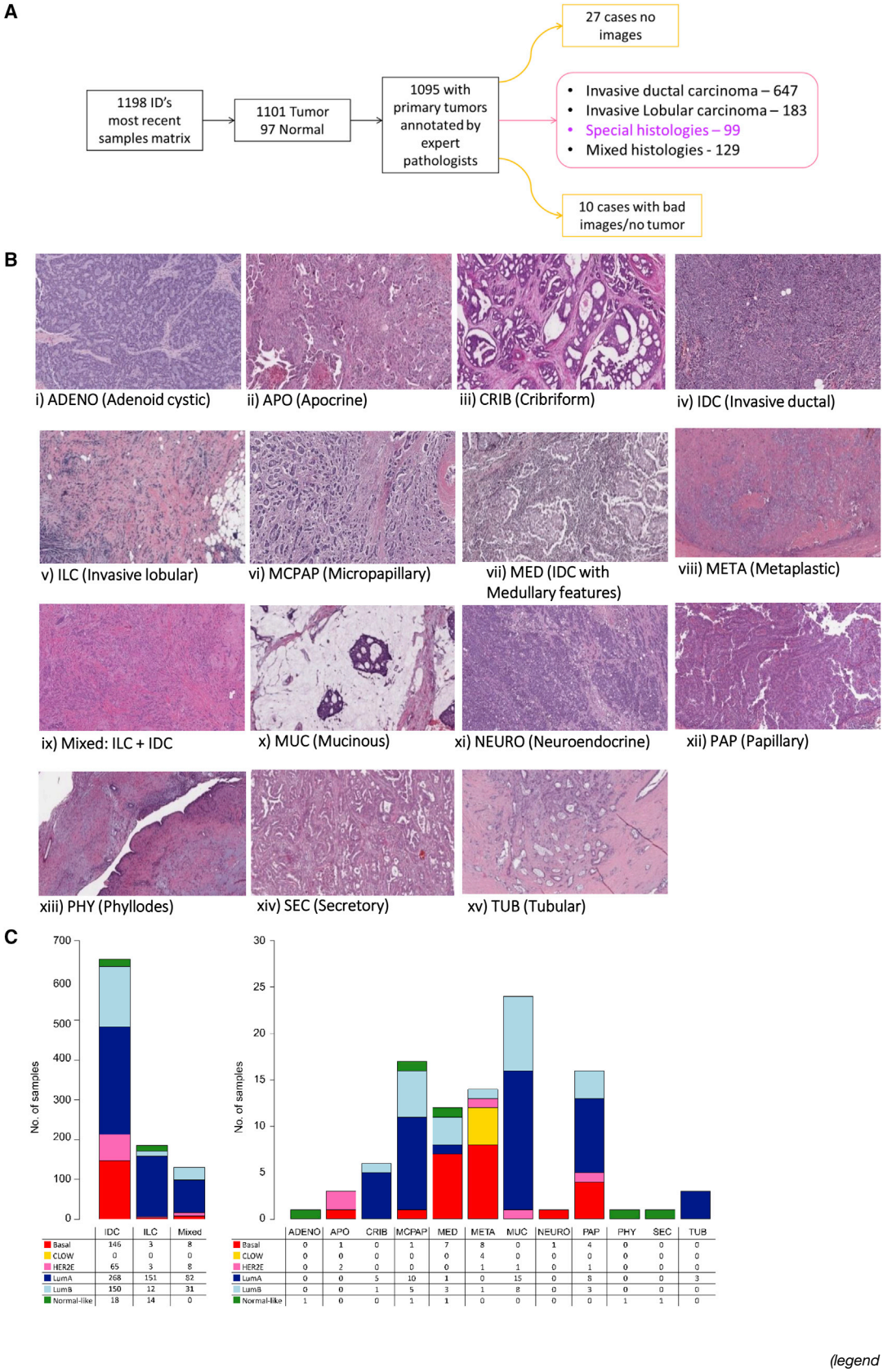
Breast cancer is classified into multiple distinct histologic types, and many of the rarer types have limited characterization. Here, we extend The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) dataset with additional histologic type annotations in a total of 1,063 breast cancers. We analyze this extended dataset to define transcriptomic and genomic profiles of six rare, special histologic types: cribriform, micropapillary, mucinous, papillary, metaplastic, and invasive carcinoma with medullary pattern. We show the broader applicability of our constructed special histologic type gene signatures in the TCGA Pan-Cancer Atlas dataset with a predictive model that detects mucinous histologic type across cancers of other organ systems. Using a normal mammary cell differentiation score analysis, we order histologic types into a continuum from stem cell-like to luminal progenitor-like to mature luminal-like. Finally, we classify TCGA-BRCA into 12 consensus groups based on integrated genomic and histological features. We present a rich, openly accessible resource of genomic, molecular, and histologic characterization of TCGA-BRCA to enable studies across the range of breast cancers.

INTRODUCTION

The World Health Organization classifies breast epithelial tumors into multiple histologic types,[1–3] which were originally defined by their unique cytologic and architectural features. Among these, the most common type of breast cancer histology is invasive breast carcinoma no special type (NST), also known as invasive ductal carcinoma (IDC), which accounts for 70%–80% of all breast cancers.[4,5] According to the 5th edition of the WHO classification of breast tumors, breast cancers with a special histologic pattern in > 90% of the cancer are designated as pure special histologic types, but cancers lacking such specific features are designated as IDC (with cancers containing between 10% and 90% a special type considered "mixed IDC-special subtype"). The pure special histologic types together make up the remaining 20%–30% of breast cancers.[3,5]

IDC exhibits heterogeneity in terms of clinical characteristics, treatment, and prognosis, which are heavily determined by its intrinsic molecular subtype as defined by gene expression.[6–10] Gene expression-defined intrinsic subtypes are also available for many of the other special histologic types;[4,11,12] however, there are specific phenotypic features that characterize these distinct rarer histologic types. Among the special types, invasive lobular carcinoma (ILC) is the second most frequent histologic form of breast cancer after IDC and was comprehensively characterized in a TCGA study.[13] ILC is histologically characterized by an absence of gland and nest formation and shows a unique appearance of dyshesive cancer cells in slender strands known as a "single file" pattern invading into the stroma.[1,2] We and many other groups before us have reported that ILC histologic type is associated with a mutation and/or low gene expression of the *CDH1*/E-cadherin gene.[13–15] Furthermore, intrinsic subtyping of ILC identifies it as a relatively homogeneous group predominantly falling under the estrogen receptor positive (ER+) luminal A (LumA) molecular subtype, but it is associated with a worse prognosis than LumA IDC.[16,17]

The remaining special histologic types are low in frequency, with their individual prevalence ranging from 0.1% to 6%.[18] These rare histologies typically show homogeneous intrinsic subtypes divided broadly into two groups: (1) ER+ luminal subtype group



(legend on next page)

composed of mucinous (MUC), tubular, and papillary (PAP) carcinomas[4,19–21] and (2) ER− basal-like subtype group including invasive carcinoma with medullary pattern (MED), adenoid cystic, and metaplastic (META).[11,22,23] However, distinct features likely exist within each rare histologic type, and the aforementioned findings suggest that each type is a unique entity with specific molecular features extending beyond intrinsic subtypes. Only a few studies have characterized together multiple rare histologic types.[4,18]

In this work, we present the final dataset for genomic, molecular, and histological analyses of 1,095 primary breast cancers in TCGA-BRCA, including for the first time complete histologic type annotations for 1,058 samples derived from the international breast cancer pathology expert committee review. With this expanded dataset, we characterize the transcriptomic and genomic profiles of six rare special histologic types in the TCGA-BRCA. We identify differentially expressed (DE) genes by histology and independent of intrinsic subtype, construct histologic group-specific gene lists that cluster the rare histologic types, and study inter-relationships relative to ILCs, IDCs, and other known molecular subtype features.

RESULTS

Gene expression analysis identifies unique features in rare histologic types

To identify rare histologic types in the TCGA database, pathologists examined the remaining 245 cases that were previously designated as "others" in their histologic type using the same criteria and methods as described for the previously published 850 TCGA breast tumors,[24] all of which were based on using the TCGA hematoxylin and eosin (H&E) slide database (Figure 1). The classification of previously designated "others" into special histologic types was performed by review of each case by at least two pathologists independently in the year 2016 and yielded 1,058 cases with interpretable H&E virtual slides, out of which 99 cases were diagnosed as rare histologic types (i.e., not IDC, ILC, or mixed IDL/ILC; Figures 1A and 1B). The pathologists used the "Breast_EPC_TCGA scoring sheet" for defining all histologic types and all histologic features; an example of this sheet can be found in the published work by Heng et al.[24] We performed intrinsic subtyping (a methodology to classify breast cancer based on gene expression) of the entire set of 1,095 breast cancers in the TCGA-BRCA dataset and noted that the TCGA breast cancer rare histologic types had a predominance of ER+ LumA and luminal B (LumB) subtypes, except for the META and invasive carcinoma with medullary pattern (MED) tumors. These two were mostly ER negative, progesterone receptor (PR) negative, and human epidermal growth factor receptor (HER2) negative, thus constituting the triple-negative breast cancer (TNBC) clinical subtype and showing mostly basal-like gene expression subtype features (Figure 1C; Data S1). Five claudin-low (CLOW) molecular subtype samples were identified in this dataset using hierarchical clustering analysis and the cell line predictor of Prat et al.[25] to identify those tumors that are called CLOW by both methods, out of which four showed the META histologic type (Figure 1C); the remaining META sample did not have representative images to be re-evaluated into a histologic type for the purpose of the present work, but it should be noted that it was denoted as a META carcinoma in the original histopathology report (Data S1). Overall, these molecular subtyping results were consistent with previous reports.[20,23,26] Interestingly, the PAP carcinomas showed intrinsic subtype heterogeneity like IDC, with the presence of LumA and basal-like subtype samples within this single histologic type (Figure 1C; Figures S1A and 1B).

The identification of basal-like gene expression features in a few PAP carcinomas was unexpected considering current knowledge of PAP neoplasm classification, which classically exhibits ER positivity and is considered to be a LumA molecular subtype.[21,27] Therefore, three experienced breast cancer pathologists (K.A., L.C.C., and S.J.S.) from the TCGA Breast Cancer Pathology Group specifically re-evaluated the digital slides of the PAP carcinoma samples in the TCGA-BRCA digital slide database and categorized them into (1) encapsulated, (2) solid, and (3) invasive PAP carcinomas. We found that PAP carcinomas of LumA and LumB molecular subtypes (PAP-Luminal) were mostly solid PAP carcinomas: invasive type (5/11), *in situ* (1/11), followed by IDC with foci of solid PAP carcinoma (2/11), IDC alone (2/11), and one case of encapsulated PAP carcinoma (Table S1). The PAP carcinomas of basal-like molecular subtype (PAP-Basal) were re-classified as IDC with pseudo-PAP features (3/4) and one solid PAP carcinoma (Figures S1C–S1F and Table S1). Lastly, one PAP carcinoma of HER2-enriched (HER2E) molecular subtype was classified as a high-grade encapsulated PAP carcinoma. It was thus decided that the utilization of PAP-Basal as a unique group would not be appropriate, as most of these cases upon re-review resembled IDC with "pseudo" papillae formation; we provide the representative H&E images of these four PAP-Basal cases and what we considered as pseudo-PAP (Figures S1C–S1E) with one true solid PAP carcinoma diagnosis (Figure S1F). All of these PAP-Basal cases had a *TP53* mutation and high histologic grade (including more cellular proliferation with < 10% tubule formation [n = 4/4], higher mitotic figures/10 high-power field [HPF; n = 4/4], and marked variation in nuclear pleomorphism [n = 3/4]).

Figure 1. Histopathologic annotation schema and overall distribution of all histopathologic types of breast cancer according to molecular subtype in TCGA

(A) Schematic of TCGA-BRCA histopathological annotation schema, with 99 samples classified into special histologic types and their representative hematoxylin and eosin (H&E)-stained photomicrographs of all the annotated TCGA-BRCA histologic types (n = 1,058; magnification 20×).

(B) Distribution of all 1,095 TCGA-BRCA primary breast cancer patients according to re-annotated histologic type and the PAM50 molecular subtype (Basal, LumA, LumB, HER2E, and normal-like), including claudin low (CLOW).

(C) ADENO, adenoid cystic carcinoma (n = 1); APO, apocrine carcinoma (n = 3); CRIB, cribriform carcinoma (n = 6); IDC, invasive ductal carcinoma not otherwise specified (n = 647); ILC, invasive lobular carcinoma (n = 183); MCPAP, micropapillary carcinoma (n = 17); MED, invasive carcinoma with medullary features (n = 12); META, metaplastic carcinoma (n = 14), MUC, mucinous carcinoma (n = 24), NEURO, neuroendocrine carcinoma (n = 1); PAP, papillary carcinoma (n = 16); PHY, phyllodes tumor (n = 1); SEC, secretory carcinoma (n = 1); TUB, tubular carcinoma (n = 3); mixed, a combination of more than one histologic type (n = 129); LumA, luminal A subtype; LumB, luminal B subtype; HER2E, HER2-enriched subtype.

To identify genes uniquely associated with each rare histologic type, we analyzed only those histologic types with a minimum of five samples from the 2016 histology annotation schema (Data S1), which left us with invasive cribriform (CRIB, n = 6), invasive micropapillary (MCPAP, n = 17), MED (n = 12), MUC (n = 24), META (n = 14), and PAP (n = 16). Differential gene expression analysis comparing each histologic type against the rest of the TCGA-BRCA samples yielded "raw" gene expression signatures containing significant upregulated genes associated with these six rare subtypes (Figure 2A; Data S3). For the sake of completeness, similar analyses were performed for the four intrinsic subtypes of IDC (LumA, LumB, basal-like, HER2E) and for ILC as a whole; note that for most comparative analyses, tumors with "mixed" histologic type annotation were excluded. These significant gene sets from the "raw" gene signatures showed an enrichment of the PAM50 intrinsic subtype-determining genes, due to the typical enrichment of one or a few intrinsic subtypes within a given rare histologic type.

To identify further transcriptomic features, we conducted gene set enrichment analysis (GSEA) that identified similar overlap between histologic types. The LumA/LumB subtype predominant histologies (i.e., CRIB, MCPAP, MUC, PAP-Luminal) shared GSEA hallmarks including early estrogen response, late estrogen response, and protein secretion (FDR = 0, Data S3), whereas the basal-like group histologies (MED, META) shared immune-related GSEAs (inflammatory response, TNF-α signaling, IFN-γ signaling), mitosis-related GSEAs (G2M checkpoint, mitotic spindle, E2F targets), and signaling pathways like MYC, KRAS, and WNT-β-catenin (FDR = 0, Data S3). Overall, these results demonstrate that although the upregulated "raw" genes were unique to each histologic type, the intrinsic subtype genes were still dominating the biological significance of this comparison.

To avoid the confounding factor of intrinsic subtype, we performed differential expression analysis of each rare histologic type against the rest of the breast cancers in TCGA-BRCA (including IDC and ILC in the analysis) falling only within the same molecular subtype of the respective rare histologic type. We refer to this differential gene expression as "mol-sub" (molecular-subtype). In this manner, we made the following comparison groups: (1) CRIB versus LumA BRCA, (2) MCPAP versus LumA/LumB BRCA, (3) MUC versus LumA/LumB BRCA, (4) MED versus LumB/basal-like BRCA, (5) META versus basal-like BRCA, and (6) PAP-Luminal versus LumA/LumB BRCA. For the META comparison group, we excluded the META samples with a CLOW gene signature (META-CLOW, 4/14 of the META group) as these had no complimentary CLOW samples in any other histologic type in the TCGA-BRCA dataset. The "mol-sub" differentially expressed genes showed reduction of PAM50 genes in each rare histologic type when compared to the "raw" lists. Overall, comparing "raw" gene expression signatures with "mol-sub" signatures showed an increase of gene sets in comparison groups for CRIB, MCPAP, MED, and META (Figure 2B; Data S3). Further, while the "raw" differential gene expression showed GSEA enrichment of estrogen-related and immune-related genes, "mol-sub" also showed enrichment in additional gene ontology (GO) and hallmark gene sets with histologic relevance.

CRIB "mol-sub" GSEAs were enriched for genes associated with ribosome activation and protein transportation (family-wise error rate [FWER] p value = 0, FDR = 0, Data S3). MCPAP "mol-sub" GSEA was enriched for significant amounts of angiogenic (FWER p value = 0, FDR = 0), endothelial (FWER p value = 0.005, FDR = 0.0004), and vasculature development (FWER p value = 0, FDR = 0) pathways in comparison to MCPAP "raw" GSEA (Data S3); MCPAP histology is characterized by tumor cell clusters present in spaces resembling vascular spaces and associated with increased lymphovascular invasion (LVI),[28] and we found LVI present in 9/17 (53%) TCGA-BRCA MCPAP samples as well (Data S2 and Table S2). MED "mol-sub" GSEA was enriched for cell-cell adhesion GO (FWER p value = 0, FDR = 0) in comparison to MED "raw" GSEA (Data S3). Finally, META "mol-sub" GSEA was enriched for many relevant pathways, like hallmark KRAS signaling pathway (FWER p value = 0.01, FDR < 0.0001), p53 pathway (FWER p value = 0.01, FDR < 0.0001), and keratinocyte differentiation (FWER p value = 0, FDR = 0). Interestingly, similar GSEA analysis on PAP-Luminal identified a high enrichment of axonemal assembly (FWER p value = 0.007, FDR = 0.001) (Data S3).

Supervised clustering of breast cancer rare histologies across multiple datasets

To analyze additional datasets for validation, we constructed combined gene lists from individual "mol-sub" gene lists derived from each histologic type (Data S3). Upregulated genes from each histology "mol-sub" gene list consistently performed well; the top upregulated significant genes (FDR < 0.05) with a log2 fold change > 2 from the six "mol-sub" gene lists were combined. From this list, duplicated genes were removed and counted only once. We designate this constructed gene list as the "upregulated mol-sub" intrinisic histology gene list. Supervised hierarchical clustering utilizing the "upregulated mol-sub" gene list separately clustered the samples largely according to the histologies (Figure 3A) when using TCGA training set data. Using consensus clustering to assess robustness, we found eight clusters that largely tracked with histologic type (Figures S3A–3C).

To validate its performance, we utilized the "upregulated mol-sub" gene list in supervised hierarchical clustering analysis on breast cancer datasets, including rare histologic types from the METABRIC dataset (Figure 3B) and the NKI special histologies dataset (Figure 3C), which effectively clustered samples based on their histologic type rather than based on the molecular subtype. Using METABRIC, which had 32 MED and 46 MUC samples, 27/32 MED samples clustered together and 43/46 MUC samples were grouped together. For the NKI dataset, 15/19 MUC samples were grouped together, as were 7/8 MCPAP, 14/20 META, and 8/10 MED samples.

Breast cancer MUC gene signature predicts MUC histologic type in other cancers

We next considered the applicability of these breast cancer special histologic types to the 9,000 other samples in the TCGA Pan-Cancer Atlas dataset.[29] We searched for similar histologic types first by name/term, which were present in other tissue systems with available histologic data. Thyroid and kidney cancer

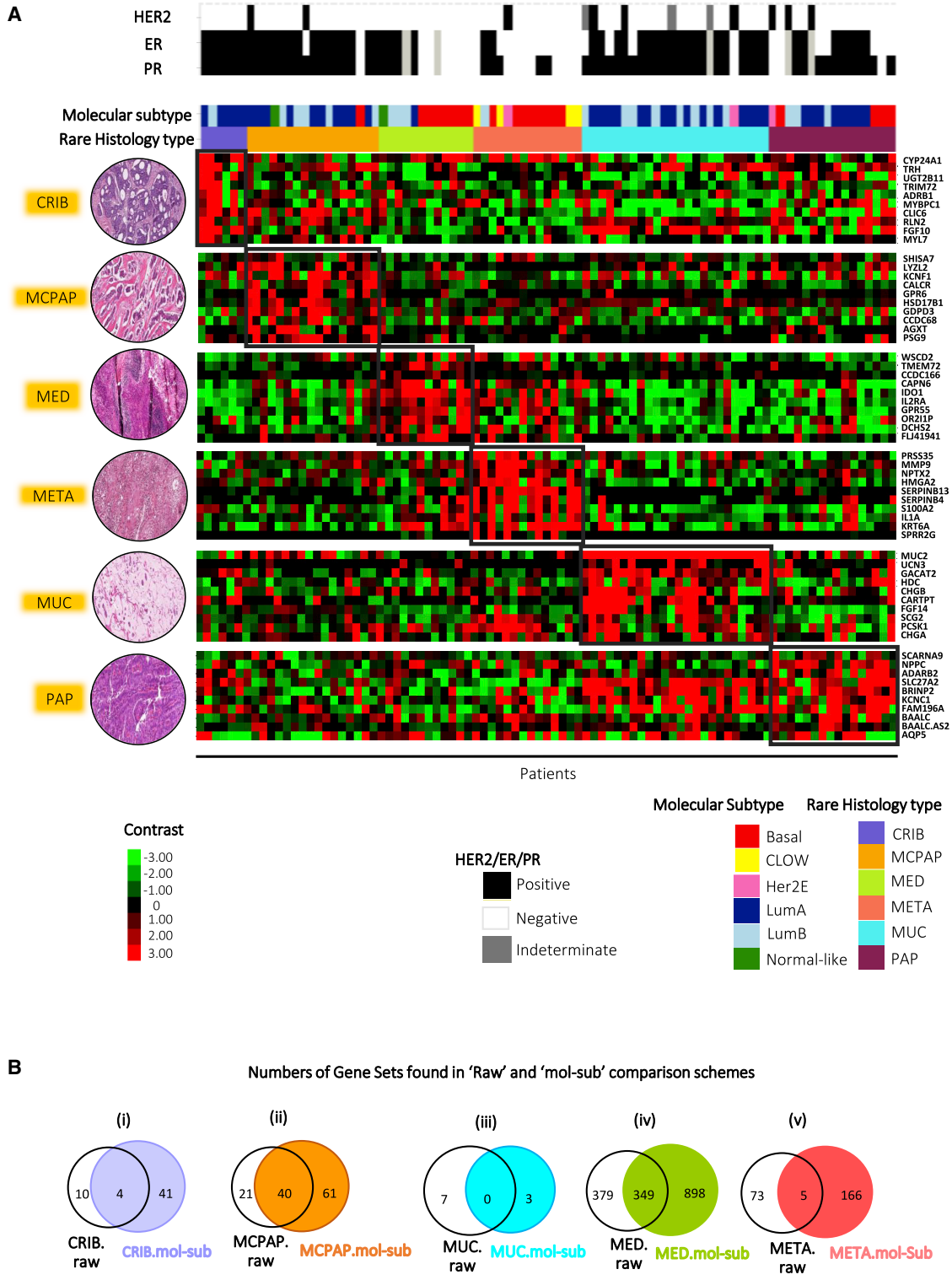
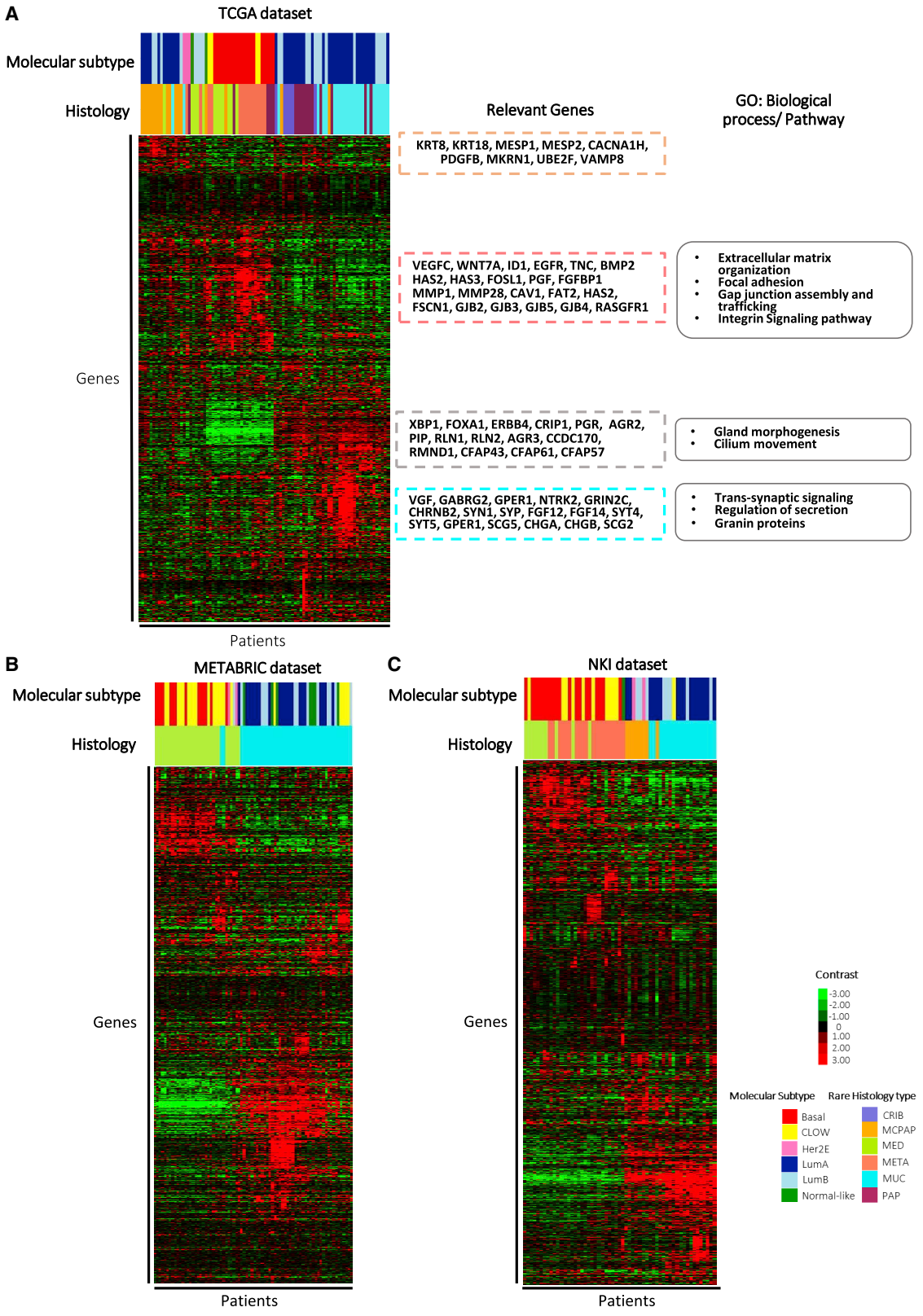


Figure 2. Upregulated genes from the "raw" and "mol-subtype" comparisons for each of the six special histologic types of breast cancer

(A) Heatmap representation of the top ten upregulated genes in the "raw" gene signatures constructed for 89 patients of CRIB (n = 6), MCPAP (n = 17), MED (n = 12), META (n = 14), MUC (n = 24), and PAP (n = 16) histologic subtypes (Bonferroni adjusted p value < 0.0001) along with important clinicopathological parameters like estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status, and American Joint Committee on Cancer (AJCC) stage for the respective histologies.

(B) Venn diagrams showing the gene sets enriched via GSEA in "raw" versus "mol-sub" comparisons for respective special histologic types (FDR < 0.05).



(legend on next page)

samples had "PAP carcinomas" histologic diagnosis; however, this is the major histologic type in these organs and, essentially, not rare cancers like the breast PAP carcinomas. Conversely, the other special histologic type relatively abundantly found in other organ systems was the MUC type, which was present in stomach adenocarcinoma (n = 20), rectal adenocarcinoma (n = 13), colon adenocarcinoma (n = 61), pancreas adenocarcinoma (n = 4), and cervical adenocarcinoma (n = 17).

We combined all of these cancer types together, along with all other epithelial cancer types, to form a dataset that had a mixture of all epithelial types of cancers, with 65 histologic types including cancers with MUC histology (Data S4). This dataset had 6,017 samples from 16 tumor types and included 132 samples diagnosed as a MUC histologic type (excluding breast cancers, Data S4). We distributed the pan-cancer 6,017-sample dataset into a training and testing dataset (70% cases in training and 30% in testing) with an equal distribution of MUC samples. We then trained an elastic net model on the TCGA-BRCA MUC "upregulated-nodal" gene list using the training dataset. The MUC "upregulated-nodal" gene list consists of the 44 genes that had high nodal correlation (> 0.80) and strongly influenced clustering of MUC samples in the TCGA BRCA dataset (Figure 3A). Our TCGA-BRCA-trained elastic net model predicted the non-BRCA MUC samples in the testing dataset with a high degree of accuracy, with an area under the curve (AUC) of 0.93 (Figure 4), thus highlighting the conserved nature of this distinct histologic type. Classical mucin gene *MUC2* was the highest positive predictive gene in this classifier (Figure 4; Table S3). Finally, GSEA analysis identified pan-cancer MUC samples to have an enrichment of O-glycan processing (FDR = 0.001, Data S4) and protein O-linked glycosylation (FDR = 0.002, Data S4) gene ontologies. As a note, we also applied the "upregulated-nodal" gene list from the breast "PAP-Luminal" histologic type to identify other PAP carcinomas in the same 6,017-sample dataset; however, the classifier failed to predict PAP carcinomas in other organs.

Breast cancer META samples and the CLOW molecular phenotype

In the context of TCGA Pan-Cancer Atlas samples, we also analyzed the distribution of breast cancers with special histology according to the clustering of cluster assignments (CoCA) pan-cancer grouping, which yields 24 major molecular subtypes.[29] Based on CoCA assignments founded upon 10,000 tumors, most of the BRCA samples fell into two major groups: 661/879 were classified in CoCA 23, and 157/879 were in CoCA 20; these two CoCA groups were largely composed of breast tumor samples and correspond to "luminal" (CoCA 23) and "basal-like" (CoCA 20). CoCA classification of the special histologic samples, regardless of the histologic type, placed these samples into two major CoCA groups; the LumA, LumB, and HER2E special histologic samples fell in CoCA 23 and the basal-like special histologic type samples in CoCA 20 (Figure 5A).

However, some META breast tumors clustered into non-breast-predominant CoCA groups, namely CoCA groups 7, 8, and 14, which prompted a more in-depth analysis of the META histologic samples. To examine the biology of the META samples, an important expression subtype not contained within the PAM50 assay needs to be considered—the CLOW subtype, which shows many stem cell-like and mesenchymal-like features.[25,30,31] Identifying CLOW samples using the Prat et al.[25] cell line-based centroid predictor typically over-calls the number of CLOW samples relative to the number obtained from hierarchical clustering, where it was originally identified; thus, we used this centroid predictor together with hierarchical clustering analysis to identify CLOW samples in the TCGA pan-cancer 1,095-sample dataset; specifically, samples had to be called CLOW by the cell line-based predictor, cluster together in a hierarchical cluster, and show the distinctive low expression of the CLOW-defining gene set. Surprisingly, only five samples were identified as centroid predictor positive and clustered together by hierarchical clustering analysis. Of note is that 4/5 of these CLOW molecular subtype samples showed META histologic type, with the fifth lacking a tumor slide image for pathologic re-annotation (Data S1); however, also note that the original TCGA report classified it as a META sample (Data S1). All the META-CLOW samples (n = 4) had a predominant mesenchymal component upon histologic examination and clustered with CoCA group 7 or 14, both of which had sarcoma (SARC) and melanoma samples as the group predominant tumor types (Figure 5A).

Additionally, the three pathologists from the TCGA Breast Cancer Pathology Group also performed a re-review of all the META samples in this analysis to delineate the META component. META samples that had a predominant squamous histologic component without a spindle cell component (n = 3; Table S4) clustered with the pan-squamous cancers CoCA group 8 that was chiefly composed of head and neck, cervical, lung, and esophageal squamous carcinomas. In contrast, the non-CLOW META samples had more features of mixed META carcinoma (2/7) with admixture of META components (Table S4). All the META-CLOW samples had a predominant mesenchymal/spindle cell component (Table S4).

We next performed gene expression and ontology analysis on these specific META samples in CoCA group 8 (META-squamous) and CoCA groups 7 and 14 (META-CLOW) from overall META (non-squamous/non-CLOW) (n = 7) samples and found squamous cell carcinoma and sarcoma genes enriched in these groups, respectively (Figure 5B; Table S5). For the META-squamous samples with a predominant squamous META component, there were several keratinocyte differentiation and epithelial development genes upregulated (*WNT7A*, *WNT10A*, *JAG1*, *SFN*, *FOXN1*, *LCE1C*, *AQP3*, *EPHA2*, and *TP63*; FDR Benjamini & Hochberg [B&H] = 0.003; Data S5). For the META-CLOW

Figure 3. Construction of a special histologic type-specific gene list that groups breast cancer patient samples according to histologic type

(A) Supervised clustering utilizing the "upregulated mol-sub" intrinsic histologic gene list clusters samples predominately according to special histologic types with clustering driven by MCPAP-, META-, MED-, and MUC-associated genes and biological pathways (n = 89).

(B and C) Supervised clustering utilizing the "upregulated mol-sub" intrinsic histologic gene list clusters samples predominately according to special histologic types and not according to the intrinsic molecular subtypes in the METABRIC (n = 78) (B) and NKI datasets (n = 57) (C).

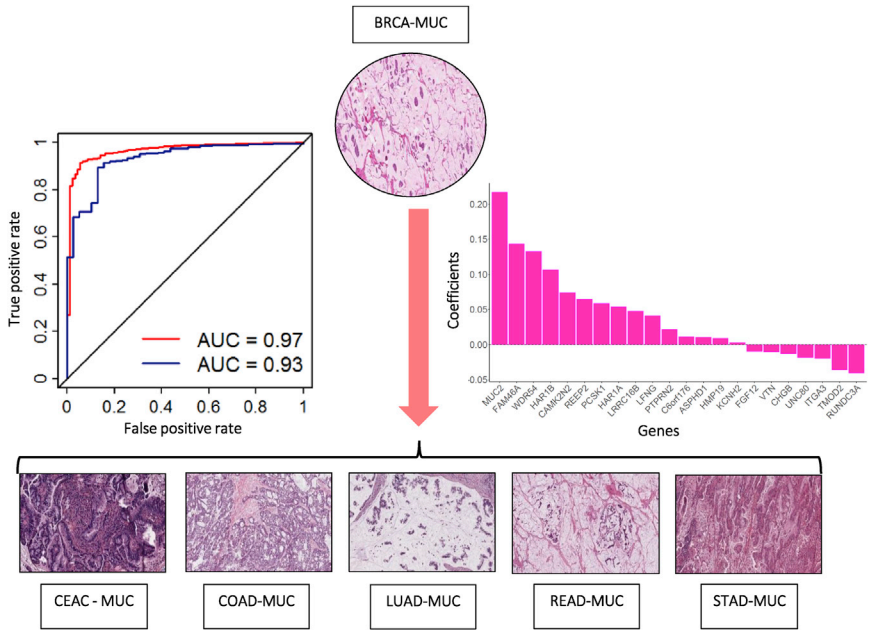


Figure 4. Elastic net modeling using breast mucinous carcinoma genes predicts mucinous carcinomas of other cancer types in the TCGA Pan-Cancer Atlas dataset

Schematic illustration of elastic net model for mucinous carcinoma histologic predictions in the TCGA Pan-Cancer Atlas dataset, containing 132 mucinous carcinomas from various organ systems. Solid red line, AUC for the training dataset; solid blue line, AUC for the testing dataset. At right, a bar plot of the coefficients of genes from the breast mucinous gene signature, in descending order, showing positive and negative contribution of genes in the mucinous histologic predictor. CEAC, cervical adenocarcinoma; COAD, colon adenocarcinoma; LUAD, lung adenocarcinoma; READ, rectal adenocarcinoma; STAD, stomach adenocarcinoma. All representative photomicrographs are of 10× magnification.

samples that had predominant mesenchymal sarcomatous features (chondrosarcoma/osteosarcoma), there were collagen-binding and extracellular matrix organization genes upregulated (*MMP9*, *MMP13*, *COL5A3*, *CTSK*, *SPARC*, *P3H1*, and *TGFBI*; FDR B&H < 0.0001; Data S5).

Consensus classifications based on genomics and histologic type

It is well appreciated that breast cancers can be classified based on histology and genomics; throughout this manuscript we noted many agreements between these two classification schemes, as well as a few disagreements. Thus, we strove to arrive at a classification schema that combines these two together into a framework for future use. For genomics-based classification, we define 5 groups (LumA, LumB, HER2E, basal-like, and CLOW), and from histology we define 8 groups (IDC, ILC, CRIB, MCPAP, MED, MUC, PAP, and META). We combined these together by creating groups based first on histology, and then according to the predominant molecular subtype (and using only classifications that have five or more representatives present within this TCGA 1,095-sample dataset); when doing so, we arrive at 12 tumor consensus groups: IDC-Basal, IDC-LumA, IDC-LumB, IDC-HER2E, ILC-Luminal, CRIB, MCPAP, PAP-Luminal, META-CLOW, META, MUC, and MED. For the sake of completeness, we show the MIXED group samples in TCGA-BRCA (any combination of two histologic types) but exclude the MIXED group from the 12 biological groups. MIXED histologic group is not a clear biologic group, but instead is a group of samples with complex histologic makeup and heterogeneous molecular subtype composition, and thus these MIXED samples do not form a single biologically homogeneous group; additional future studies are needed to identify unique properties of the MIXED group.

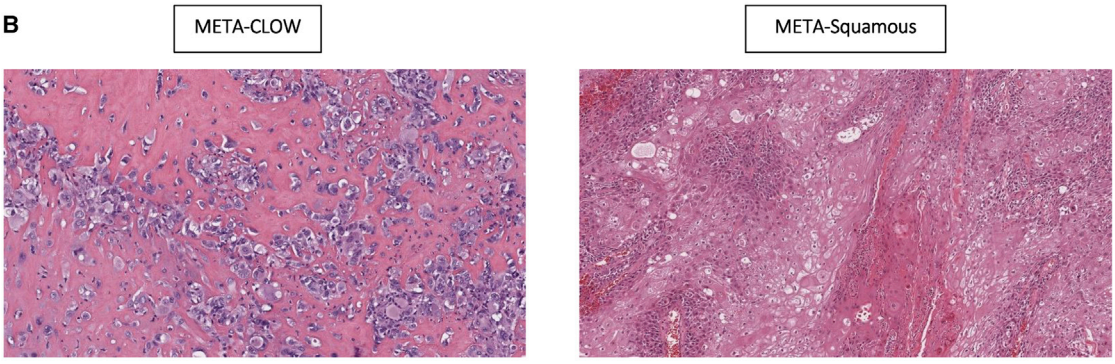
To initially characterize these 12 consensus groups, we analyzed them using our previously published breast epithelial "differentiation score" (D score)[25], which is a transcriptome-based score based on three FACS-purified normal breast epithelial cell populations (i.e., mammary stem cell, luminal progenitor, and mature luminal, each of which was gene expression profiled). Briefly, this method utilizes a distance-weighted discrimination (DWD) alignment of gene expression values of a test sample to a defined normal mammary cell type axis;[25] the D score ranks gene expression along this scale from the mammary stem cell signature at one end of the scale, the luminal progenitor in the middle, and the mature luminal signature at the other end. We organized the 12 consensus groups into a gradual continuum based on ascending order of the D score (Figure 6A).

Next, we were interested in performing genomic analyses of these 12 consensus groups, including DNA copy number (CNA) and mutational analyses; however, the small sample sizes of the groups were underpowered for routine methods of analyzing TCGA CNA and somatic mutation profiles. Instead, we calculated percentages of samples within a consensus group exhibiting broad CNA chromosome arm level (GISTIC2 calls) and somatic mutation events. Percentage counts of DNA CNA events were compared based on the transcriptomic D score; as such, the 12 consensus groups were organized into four broader groups: (1) basal group (IDC-Basal), (2) low differentiation group (MED, META; this group contained special histologic types with low D scores and histologically exhibited features of low-differentiation-like metaplasia), (3) luminal group (IDC-LumA/LumB/HER2E, ILC-Luminal, MCPAP), and (4) high differentiation group (MUC, CRIB, PAP-Luminal; this group had special histologies with high D scores and features of higher-differentiation-like mucus secretion and papillae formation). Upon comparison of all four groups together, the broad chromosomal arm events that were significantly associated with D scores were 5q loss, 3q gain, 4p loss, 8q gain, 13q loss, and 2p gain (ANOVA p value for all groups < 0.0001 and ANOVA p value for pairwise comparison < 0.005, Table S6). Among these events, however, 4p loss and 2p gain were the two events that showed a steady

A

Histology+MolecularSubtype	CoCA-6	CoCA-7	CoCA-8	CoCA-13	CoCA-14	CoCA-20	CoCA-22	CoCA-23	CoCA-24	Total
CRIB	0	0	0	0	0	0	0	6	0	6
IDC-Basal	0	0	0	7	0	136	0	1	3	147
IDC-HER2E	0	0	0	2	0	5	0	45	13	65
IDC-LumA	0	0	0	10	0	0	1	254	2	267
IDC-LumB	0	0	0	4	0	0	0	145	0	149
ILC-Luminal	0	0	0	10	0	0	0	152	0	162
MCPAP	1	0	0	0	0	1	0	15	0	17
MED	0	0	0	0	0	8	0	4	0	12
META	0	0	0	1	0	3	0	3	0	7
META-Squamous	0	0	3	0	0	0	0	0	0	3
META-CLOW	0	2	0	0	2	0	0	0	0	4
MUC	0	0	0	0	0	0	0	24	0	24
PAP-Basal	0	0	0	0	0	4	0	0	0	4
PAP-Luminal	0	0	0	0	0	0	0	12	0	12
CoCA Group Total	1	2	3	25	2	157	1	661	18	879
CoCA Group Predominant Cancer Type	PAAD	SARC	Pan-Squam	Multiple	SARC	BRCA	LAML	BRCA	BRCA	

B



Disease	p-value	q-value FDR B&H
Dermatofibrosarcoma	1.44E-05	3.94E-03
Osteosarcoma of bone	7.45E-05	0.009147

Name	p-value	q-value FDR B&H
Squamous cell carcinoma	4.5E-11	1.24E-07
Adenocarcinoma	1.63E-07	0.000225
Carcinoma breast stage IV	6.51E-07	0.0006

Figure 5. Metaplastic carcinomas of breast group away from other breast carcinomas into other subtypes dictated by their predominant type of metaplasia

(A) Tabular representation of TCGA-BRCA samples according to the pan-cancer cluster of cluster assignment (CoCA) groups and histologic types. Colored rows represent breast samples clustering out of the predominant breast CoCA group.

(B) Representative disease correlation using top upregulated genes obtained after differentially expressed (DE) gene analysis between META-squamous group versus META (right panel) and META-sarcoma group versus META (left panel). PAAD, TCGA pancreatic adenocarcinoma; SARC, TCGA sarcomas; Pan-Squam, TCGA pan-squamous group; LAML, TCGA acute myeloid leukemia; BRCA, TCGA breast cancer; multiple, no specific TCGA cancer study; B&H, Benjamini-Hochberg adjusted; FDR, false discovery rate.

decrease in percentage as the D score and degree of histologic differentiation increased (Figure 6B; Data S6). The second comparison was the presence of events specifically frequent within the low-differentiated histologies group (MED, META) and IDC-Basal. In this scenario, 14q loss (ANOVA p value < 0.0001 and ANOVA p value for pairwise comparison < 0.005, Table S6) were not frequent marks seen in the META histologic type, showcasing the uniqueness of the META histologic type within this group (not shown in the figure). The final broad CNA pairwise comparison was the presence of events frequent within the high-differentiation group (CRIB, MUC, and PAP-Luminal) and the luminal group (IDC-LumA, IDC-LumB, ILC-Luminal, MCPAP, and IDC-HER2E); a major event found here was the decrease in frequency of 22q loss events in the histologies with a high D score (ANOVA p value < 0.0001) followed by a decrease in frequency of 20q gain (statistically not significant) (Figure 6C). Overall, a trend of decrease in overall CNA events was seen as the degree of differentiation of a histologic type increased.

Similarly, looking at mutations, we found that the *TP53* mutation percentage was high in histologies with a low D score, and *GATA3* mutation percentage was high in histologies with a higher D score (Figure 6D; Figures S4 and S5 and Table S7). *PIK3CA* mutation percentages were highest in ILC-Luminal and IDC-LumA/B and lower in all other histologic types

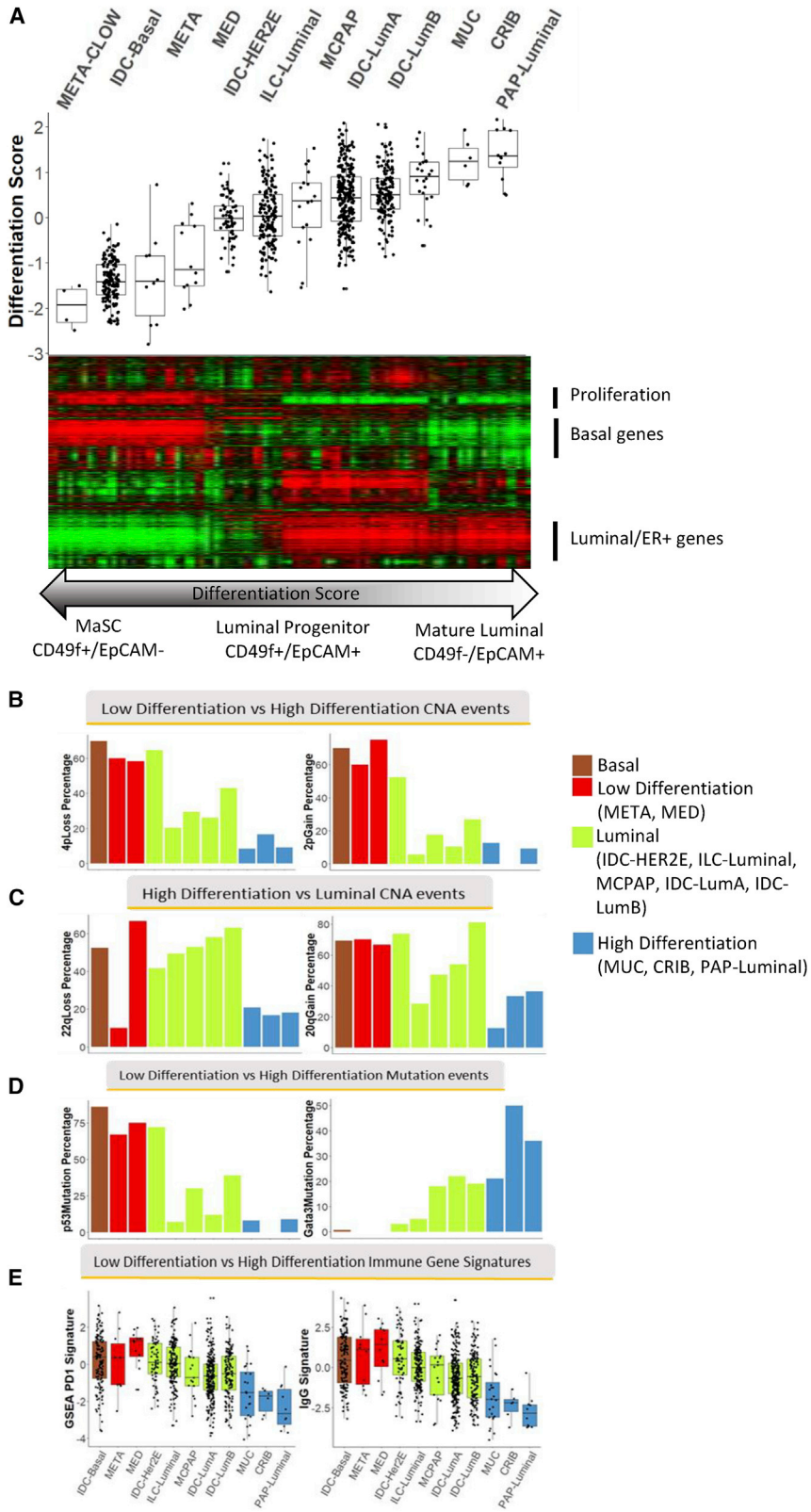


Figure 6. Histologic types of breast cancer can be grouped based on normal mammary cell type differentiation score (D score), specific CNA events, mutation events, and immunologic gene signatures

(A) Box-and-whisker plots of the 12 consensus groups defined by histology and gene expression (x axis) using the D score (y axis) indicate the median score (horizontal line), the interquartile range (IQR, box boundaries), and 1.5 times the IQR (whiskers). The heatmap indicates the clustering of the 7,000 most variable genes of 886 samples with D score.

(B) Bar plots of significant CNA events high in "high differentiation" and "low differentiation" histologic groups.

(C) Bar plots of significant CNA events unique to "high differentiation" versus "luminal" group.

(D) Bar plot of Tp53 and Gata3 mutation events in the 12 biologically relevant breast cancer groups.

(E) Box-and-whisker plots of the GSEA PD1 and IgG transcriptomic signatures (fourth row); median score (horizontal line), the interquartile range (IQR, box boundaries), and 1.5 times the IQR (whiskers). Broad groups are separated by dotted colored lines: brown, IDC-Basal; red, low differentiation group (MED, META); green, luminal group (IDC-HER2E, ILC-Luminal, MCPAP, IDC-LumA, IDC-LumB); blue, high differentiation group (MUC, CRIB, PAP-Luminal). MaSC, mammary stem cells. The original FACS-sorted population nomenclature and cell surface markers that the D score was based upon are highlighted.[25]

See also Table S7, Data S6, and Figures S4 and S5.

(Figures S4 and S5). Finally, certain histologies exhibited unique mutation percentages, e.g., *PTEN* mutation was present in 30% of META samples and *MAP3K1* mutation was present in 24% of MCPAP samples (Table S7). There was no unequivocally clinically actionable mutation that was specific for any of the six special histologic types. Next, we examined the most common actionable (BRCA1/2 and PIK3CA) events and another set of mutations that some might consider actionable, like PTEN and MAP3K1. None of the six special types harbored more than one BRCA1/2 somatic mutation and/or germline mutation, except for the MED histologic type, which had a BRCA1 germline event detected in 3/12 patients (Table S8); however, MCPAP (5/17) and PAP-Luminal (3/11) showed common PIK3CA mutations, which might be targeted with alpelisib if treated in the metastatic setting. We also report MAP3K1 mutations in the MCPAP (5/17) histologic type, with one case having co-occurring PIK3CA and MAP3K1 mutations (Table S7).

We also identified a correlation between the D score for the 12 consensus groups with immune cell genomic signatures. Immune cell signatures including GSEA_PD1, IgG Signature (Figure 6E, significance analysis of microarrays [SAM] q value = 0) showed a low signature value for the histologies with a higher differentiation (MUC, CRIB, PAP-Luminal) and, conversely, high immune signature scores in low-differentiation histology types like META and MED. Including these two, we also report 35 statistically significant immune cell signatures that have a similar pattern (Table S9, SAM q value = 0).

DISCUSSION

Breast cancer is known to have at least 21 different histologic types, classified using a combination of architectural and cytological features. Among these, the majority are IDCs, and the rest form special histologic types[1–3]. IDC is essentially a diagnosis of exclusion: these tumors lack the defining features of any of the special histologic types. However, it is well appreciated that distinct molecular intrinsic subtypes have been identified among IDCs, which dictate much of their underlying biology.[10] These intrinsic subtypes are also found in the special histologic types, along with additional genes influencing the distinct phenotypic features found in them. For example, somatic mutation and/or loss of protein expression of *CDH1*/E-cadherin is associated with non-cohesive cells in ILCs.[13] In this work, we characterized the transcriptome profiles of six rare histologic types in the TCGA-BRCA dataset, each represented by five or more tumors.

We analyzed four conventionally ER+ special histologic types—CRIB, MUC, MCPAP, and PAP—and two TNBC histologic types, META and MED. To identify genes of histologic relevance, we performed differential gene expression analysis followed by GSEA on each of these six histologies. We found that within each histologic type, specific comparisons taking into account the predominant molecular subtype increased the number of relevant GSEA findings associated with these histologic types (Figure 2B; Data S3). For example, MCPAP was found to have an enrichment of genes associated with endothelial cell activation and angiogenic pathways, which could explain its high propensity for LVI.[32,33] This could also indicate that the MCPAP histologic type might benefit from anti-angiogenic drugs. MED signature was associated with cell adhesion genes, which could explain its "sheet-like" or "syncytial" histologic growth pattern,[34] although when we clustered this histology with all of the basal subtype histologic types, the signature was not very well highlighted (Figure S2B). This correlates with the recent removal of this entity as a distinct histology by the WHO, describing it instead as IDC with medullary-like features. We would like to note that, as a biologic group, MED was predominately composed of a basal molecular subtype. META was associated with keratinization and epithelial-mesenchymal transition-related pathways, which could explain its higher percentage of distant metastasis and skin involvement.[35] Both MED and META histologic types also had a significant enrichment of the EGFR pathway, which could indicate a possible response to anti-EGFR therapy. The presence of such biological pathways correlated well with mixed spindle and squamous components that were noted in this pathologically diverse histologic group (Data S1). Using pan-cancer analysis, we were able to further separate the META histologic type into predominant squamous component type (META-squamous) and predominant META mesenchymal component type (META-CLOW), and in the latter case we show that these rare breast tumors share more molecular features with sarcomas than they do with other breast cancers. This indicates unique genomic events occurring in these histologic types, distinct from other breast cancers. This also indicates the relevance of determining the CLOW subtype in the case of a META histologic type, as this forms a unique biologic group within this histology type.

MUC histologic type is used when > 90% of tumor clusters are found floating in pools of mucin.[1,2,36] These tumors have been described as unique transcriptomic and genomic entities when compared against IDC,[20] and through our analysis we again find this to be true, including the granin proteins that have been described before. We further validated this finding by showing that a newly derived MUC histologic signature seems to be shared in MUC carcinomas of other organ systems (Figure 5). This finding has been previously reported, whereby the mucin histochemistry shared between these tumors is associated with O-acylation of sialomucins.[37,38] Thus, we corroborate this finding and provide a unique gene signature involved with the MUC histologic type.

In the pathological annotation scheme from the present study, the PAP histologic type had the presence of ER+ LumA/LumB and ER− basal-like intrinsic subtypes, which alerted us to revisit these slide images again. Upon re-examination with a panel of three expert breast cancer pathologists, it was agreed that most of the basal-like PAP carcinomas were more consistent with high-grade IDCs with pseudo-PAP growth (3/4). However, one of the cases was confirmed to be a solid PAP carcinoma (Figure S1F). This finding correlated well with a study of PAP carcinomas by Piscuoglio et al.,[27] which also reported one basal-like solid PAP carcinoma. We report the genomic and mutation profiles of these so-called basal-like PAP carcinomas here both to note the potential diagnostic pitfalls of high-grade basal-like cancers that have pseudo-papillae rather than true invasive papillary growth and to reiterate that solid PAP carcinomas rarely fall into non-luminal molecular subtypes. These

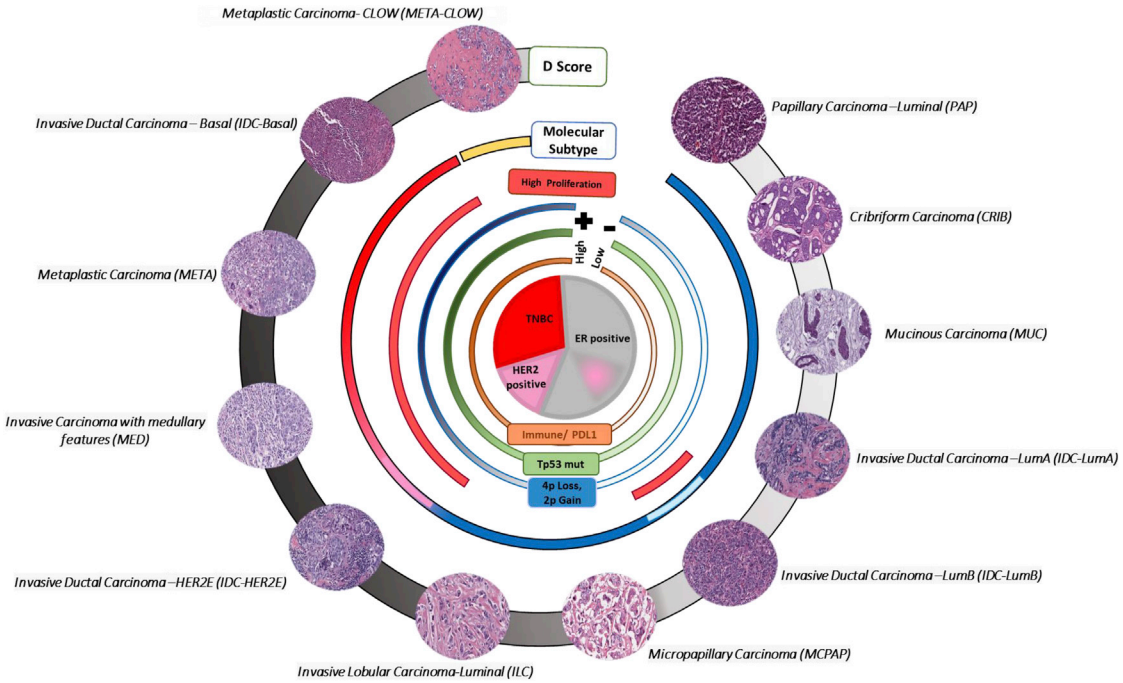


Figure 7. A TCGA breast cancer classification based on molecular and histologic features combined

Schematic representation of 12 consensus groups defined by histology and gene expression analyses of the TCGA-BRCA dataset and organized by differentiation (D) score. These groups are connected by an outer ring based on D score (lowest differentiation to highest differentiation arranged in anticlockwise direction). From the D score ring inward: The second ring exhibits PAM50 subtype association with D score (red, basal; pink, HER2E; blue, LumA and LumB; yellow, claudin low). The next ring highlights the proliferation gene signature, which is high in all biological groups with a low D score but also in one group with a high D score, namely IDC-LumB. The next two rings represent the descending abundance of CNA events (4p loss, 2p gain) and mutation events (Tp53 mutation) associated with ascending D score. The final ring exhibits decreasing immunological gene signatures in relation to ascending D score. The innermost pie chart exhibits the clinical immunohistochemistry status found in these 12 breast cancer consensus groups.

tend to harbor *TP53* mutations and should be diagnosed and managed with the caution that they may not share the favorable outcomes associated with the more typical luminal solid PAP carcinoma. Most of the PAP-Luminals were called invasive solid PAP carcinomas or an IDC with a solid PAP carcinoma component. There was transcriptomic similarity of PAP-Luminal to MUC carcinomas in having highly differentiated cell structures like cilia, suggesting a common pathobiology, and having a higher cellular differentiation, as shown by the D score (Figure 6A).

We were underpowered to identify CNA and somatic mutation events associated with these rare types with statistical significance, given our small sample sizes. However, we sought to look for broad CNA events per chromosome arm and important breast cancer somatic mutation events per individual histologic group (Figure 6B) by correlating the transcriptomic D score to genomic events. We classify 12 consensus groups derived from our histological and gene expression analyses of the TCGA-BRCA dataset and organize these according to D score. We again note that the histologic types with the lowest D scores are thought to share transcriptomic similarity to mammary stem cells, those with low to moderate scores relate to luminal progenitor cells, and those with high D scores are transcriptomically like mature luminal cells. In this way we find that the META-CLOW is the lowest, IDC subtypes are mostly in the middle, and ER+ special histologic types (MUC, CRIB, PAP-Luminal) are on the high end.

In our continuum of the 12 consensus groups organized by D score, first is META-CLOW to IDC-Basal, then next are the ER− special histologic types (META, MED), and then IDC-HER2E, ILC-Luminal, MCPAP, IDC-LumA, and IDC-LumB, in that order. Using this organization, we can place certain DNA CNA events, such as the high percentage of 4p loss, 2p gain, and *TP53* mutation, within the TNBC/IDC-Basal and special histologies with lower D scores (Figure 7), and which go lower as one leaves this section of the wheel. Similarly, we find a paucity of 22q loss, 20q gain, and increased GATA3 mutation as uniting features of higher differentiated histologic types. Besides these, we also found that the D scores also inversely correlated with 35 RNA-seq-based immune cell signatures. This finding can suggest possible use of immunotherapy in special histologies with lower D scores as is already being explored in IDC-Basal.

Limitations of the study

There are some limitations to our study. First and foremost, we had small numbers of tumors in the six special histologic groups, which in many cases were not powered for robust statistical tests, especially for mutation-associated features. However, for gene expression, we were able to validate a number of these associations on other datasets with special histologic annotations. Given

that these special histologic groups are rare,[1–3] even these small numbers are useful in helping with their study. Additionally, to validate our gene expression results from this study, we utilized the METABRIC and NKI microarray-based gene expression datasets, which contained special histologic type.

Another limitation is that our histologic annotations are derived from one or two virtual slides and therefore were subject to less than the ideal method of definitive diagnosis recommended for these entities. Although the pathology review committee did not have access to all the original diagnostic slides, multiple expert breast cancer pathologists reviewed each virtual slide, and their agreement was taken into account when a diagnostic annotation was provided for the TCGA-BRCA dataset. In light of the 2019 revised WHO breast histopathology classification,[2] MED carcinomas are not considered a unique histologic type and are considered as a special variety of IDC known as IDC with medullary features. A re-annotation was also performed for all PAP carcinomas and META carcinomas, as described in the Results. We have also included the original TCGA histologic type diagnoses according to the pathology reports available at the time of sample collection to offer a holistic picture of the revision of these entities.

In conclusion, our aim with the TCGA breast tumor biological wheel is to present a uniting scheme for molecular, histological, and biological information on breast cancers. Based on our combined histologic and molecular subtype approach, we estimate 12 consensus groups in the TCGA-BRCA dataset, but we encourage further research in identifying additional unique and biological groupings. Here, we have provided updated reference of histological and genomic annotations for the TCGA-BRCA dataset, providing a comprehensive set of classifications for this unique and highly utilized breast cancer resource.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets used and histopathological assessment
  - mRNA-seq analysis
  - Differentiation Score Calculation
  - Elastic net modeling
  - Tumor Class specific DNA copy number identification and Mutation analysis
  - Immune Gene Signature Module Calculation and Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2021.100067.

ACKNOWLEDGMENTS

This work was supported by funds from the NCI Breast SPORE program (P50-CA58223), the Breast Cancer Research Foundation (RO1-CA148761 and RO1-CA195740), and Susan G. Komen (SAC-160074 and PDF17479425) to C.M.P. The results published here are in part based upon data from the Cancer Genome Atlas managed by the NCI and NHGRI (dbGaP accession phs000178). We would like to thank the TCGA Network, TCGA Breast Cancer Analysis Working Group, and the TCGA Breast Cancer Pathology Group for the many significant contributions to this dataset and work presented here. We also thank Dr. Tomás Pascual for his helpful comments.

AUTHOR CONTRIBUTIONS

A.T. conducted most of the analysis and wrote the paper. C.M.P. conceived the ideas of analysis and supervised all aspects of this paper. A.B. and F.B. conducted the histopathologic assessments of TCGA breast cancer virtual slides and defined the revised special histologic type annotations with support from K.A., L.C.C., G.M.T., Y.-Y.C., and S.J.S. K.A.H. performed the initial RNA-seq preprocessing of all TCGA data and aided in interpretation of the data. Y.X. performed the elastic net modeling. S.G.R. helped in providing ideas on drafting the manuscript and presenting the figures. All authors discussed and contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

C.M.P. is an equity stockholder and consultant of BioClassifier LLC; C.M.P. is also listed as an inventor on patent applications for the Breast PAM50 Subtyping assay.

Received: November 9, 2020
Revised: March 24, 2021
Accepted: July 17, 2021
Published: December 8, 2021

REFERENCES

1. Tan, P.H., Ellis, I., Allison, K., Brogi, E., Fox, S.B., Lakhani, S., Lazar, A.J., Morris, E.A., Sahin, A., Salgado, R., et al.; WHO Classification of Tumours Editorial Board (2020). The 2019 World Health Organization classification of tumours of the breast. Histopathology *77*, 181–185. https://doi.org/10.1111/his.14091.
2. WHO (2019). Breast Tumours. In WHO Classification of Tumours, Fifth Edition (WHO).
3. Dieci, M.V., Orvieto, E., Dominici, M., Conte, P., and Guarneri, V. (2014). Rare breast cancer subtypes: histological, molecular, and clinical peculiarities. Oncologist *19*, 805–813. https://doi.org/10.1634/theoncologist.2014-0108.
4. Weigelt, B., Horlings, H.M., Kreike, B., Hayes, M.M., Hauptmann, M., Wessels, L.F., de Jong, D., Van de Vijver, M.J., Van't Veer, L.J., and Peterse, J.L. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. J. Pathol. *216*, 141–150.
5. Weigelt, B., Geyer, F.C., and Reis-Filho, J.S. (2010). Histological types of breast cancer: how special are they? Mol. Oncol. *4*, 192–208.
6. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.
7. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.
8. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

9. Chia, S.K., Bramwell, V.H., Tu, D., Shepherd, L.E., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K.I., et al. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. Clin. Cancer Res. *18*, 4465–4472.

10. Network, C.G.A.; Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

11. Bertucci, F., Finetti, P., Cervera, N., Charafe-Jauffret, E., Mamessier, E., Adélaïde, J., Debono, S., Houvenaeghel, G., Maraninchi, D., Viens, P., et al. (2006). Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers. Cancer Res. *66*, 4636–4644.

12. Horlings, H.M., Weigelt, B., Anderson, E.M., Lambros, M.B., Mackay, A., Natrajan, R., Ng, C.K., Geyer, F.C., van de Vijver, M.J., and Reis-Filho, J.S. (2013). Genomic profiling of histological special types of breast cancer. Breast Cancer Res. Treat. *142*, 257–269.

13. Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al.; TCGA Research Network (2015). Comprehensive molecular portraits of invasive lobular breast cancer. Cell *163*, 506–519.

14. Bertucci, F., Orsetti, B., Nègre, V., Finetti, P., Rougé, C., Ahomadegbe, J.-C., Bibeau, F., Mathieu, M.-C., Treilleux, I., Jacquemier, J., et al. (2008). Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles. Oncogene *27*, 5359–5372.

15. Weigelt, B., Geyer, F.C., Natrajan, R., Lopez-Garcia, M.A., Ahmad, A.S., Savage, K., Kreike, B., and Reis-Filho, J.S. (2010). The molecular underpinning of lobular histological growth pattern: a genome-wide transcriptomic analysis of invasive lobular carcinomas and grade- and molecular subtype-matched invasive ductal carcinomas of no special type. J. Pathol. *220*, 45–57.

16. Brouckaert, O., Laenen, A., Smeets, A., Christiaens, M.R., Vergote, I., Wildiers, H., Moerman, P., Floris, G., and Neven, P.; MBC Leuven (2014). Prognostic implications of lobular breast cancer histology: new insights from a single hospital cross-sectional study and SEER data. Breast *23*, 371–377.

17. Chen, Z., Yang, J., Li, S., Lv, M., Shen, Y., Wang, B., Li, P., Yi, M., Zhao, X., Zhang, L., et al. (2017). Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma. PLoS ONE *12*, e0182397.

18. Weigelt, B., and Reis-Filho, J.S. (2009). Histological and molecular types of breast cancer: is there a unifying taxonomy? Nat. Rev. Clin. Oncol. *6*, 718–730.

19. Lopez-Garcia, M.A., Geyer, F.C., Natrajan, R., Kreike, B., Mackay, A., Grigoriadis, A., Reis-Filho, J.S., and Weigelt, B. (2010). Transcriptomic analysis of tubular carcinomas of the breast reveals similarities and differences with molecular subtype-matched ductal and lobular carcinomas. J. Pathol. *222*, 64–75.

20. Weigelt, B., Geyer, F.C., Horlings, H.M., Kreike, B., Halfwerk, H., and Reis-Filho, J.S. (2009). Mucinous and neuroendocrine breast carcinomas are transcriptionally distinct from invasive ductal carcinomas of no special type. Mod. Pathol. *22*, 1401–1414.

21. Duprez, R., Wilkerson, P.M., Lacroix-Triki, M., Lambros, M.B., MacKay, A., A'Hern, R., Gauthier, A., Pawar, V., Colombo, P.E., Daley, F., et al. (2012). Immunophenotypic and genomic characterization of papillary carcinomas of the breast. J. Pathol. *226*, 427–441.

22. Wetterskog, D., Lopez-Garcia, M.A., Lambros, M.B., A'Hern, R., Geyer, F.C., Milanezi, F., Cabral, M.C., Natrajan, R., Gauthier, A., Shiu, K.K., et al. (2012). Adenoid cystic carcinomas constitute a genomically distinct subgroup of triple-negative and basal-like breast cancers. J. Pathol. *226*, 84–96.

23. Pareja, F., Geyer, F.C., Marchiò, C., Burke, K.A., Weigelt, B., and Reis-Filho, J.S. (2016). Triple-negative breast cancer: the importance of molecular and histologic subtyping, and recognition of low-grade variants. NPJ Breast Cancer *2*, 16036.

24. Heng, Y.J., Lester, S.C., Tse, G.M., Factor, R.E., Allison, K.H., Collins, L.C., Chen, Y.Y., Jensen, K.C., Johnson, N.B., Jeong, J.C., et al. (2017). The molecular basis of breast cancer pathological phenotypes. J. Pathol. *241*, 375–391.

25. Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. *12*, R68.

26. Weigelt, B., Ng, C.K., Shen, R., Popova, T., Schizas, M., Natrajan, R., Mariani, O., Stern, M.H., Norton, L., Vincent-Salomon, A., and Reis-Filho, J.S. (2015). Metaplastic breast carcinomas display genomic and transcriptomic heterogeneity [corrected]. Mod. Pathol. *28*, 340–351. https://doi.org/10.1038/modpathol.2014.142.

27. Piscuoglio, S., Ng, C.K., Martelotto, L.G., Eberle, C.A., Cowell, C.F., Natrajan, R., Bidard, F.C., De Mattos-Arruda, L., Wilkerson, P.M., Mariani, O., et al. (2014). Integrative genomic and transcriptomic characterization of papillary carcinomas of the breast. Mol. Oncol. *8*, 1588–1602. https://doi.org/10.1016/j.molonc.2014.06.011.

28. Chen, L., Fan, Y., Lang, R.G., Guo, X.J., Sun, Y.L., Cui, L.F., Liu, F.F., Wei, J., Zhang, X.M., and Fu, L. (2008). Breast carcinoma with micropapillary features: clinicopathologic study and long-term follow-up of 100 cases. Int. J. Surg. Pathol. *16*, 155–163.

29. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al.; Cancer Genome Atlas Network (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell *173*, 291–304.e6.

30. Pommier, R.M., Sanlaville, A., Tonon, L., Kielbassa, J., Thomas, E., Ferrari, A., Sertier, A.S., Hollande, F., Martinez, P., Tissier, A., et al. (2020). Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution. Nat. Commun. *11*, 3431. https://doi.org/10.1038/s41467-020-17249-7.

31. Fougner, C., Bergholtz, H., Norum, J.H., and Sørlie, T. (2020). Re-definition of claudin-low as a breast cancer phenotype. Nat. Commun. *11*, 1787. https://doi.org/10.1038/s41467-020-15574-5.

32. Kim, M.-J., Gong, G., Joo, H.J., Ahn, S.-H., and Ro, J.Y. (2005). Immunohistochemical and clinicopathologic characteristics of invasive ductal carcinoma of breast with micropapillary carcinoma component. Arch. Pathol. Lab. Med. *129*, 1277–1282.

33. Marchiò, C., Iravani, M., Natrajan, R., Lambros, M.B., Savage, K., Tamber, N., Fenwick, K., Mackay, A., Senetta, R., Di Palma, S., et al. (2008). Genomic and immunophenotypical characterization of pure micropapillary carcinomas of the breast. J. Pathol. *215*, 398–410.

34. Wang, X.-X., Jiang, Y.-Z., Liu, X.-Y., Li, J.-J., Song, C.-G., and Shao, Z.-M. (2016). Difference in characteristics and outcomes between medullary breast carcinoma and invasive ductal carcinoma: a population based study from SEER 18 database. Oncotarget *7*, 22665–22673.

35. Abouharb, S., and Moulder, S. (2015). Metaplastic breast cancer: clinical overview and molecular aberrations for potential targeted therapy. Curr. Oncol. Rep. *17*, 431.

36. Pareja, F., Lee, J.Y., Brown, D.N., Piscuoglio, S., Gularte-Mérida, R., Selenica, P., Da Cruz Paula, A., Arunachalam, S., Kumar, R., Geyer, F.C., et al. (2019). The Genomic Landscape of Mucinous Breast Cancer. J. Natl. Cancer Inst. *111*, 737–741. https://doi.org/10.1093/jnci/djy216.

37. Sáez, C., Japón, M.A., Poveda, M.A., and Segura, D.I. (2001). Mucinous (colloid) adenocarcinomas secrete distinct O-acylated forms of sialomucins: a histochemical study of gastric, colorectal and breast adenocarcinomas. Histopathology *39*, 554–560.

38. García-Labastida, L., Garza-Guajardo, R., Barboza-Quintana, O., Rodríguez-Sanchez, I.P., Ancer-Rodríguez, J., Flores-Gutierrez, J.P., and Gómez-Macías, G.S. (2014). CDX-2, MUC-2 and B-catenin as intestinal markers in pure mucinous carcinoma of the breast. Biol. Res. *47*, 43.

39. Gutman, D.A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., Saltz, J.H., Brat, D.J., and Cooper, L.A.D. (2013). Cancer Digital Slide Archive: an

informatics resource to support integrated in silico analysis of TCGA pathology data. J. Am. Med. Inform. Assoc. *20*, 1091–1098.

40. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al.; METABRIC Group (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.
41. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.
42. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
43. Reich, M., Liefeld, T., Gould, J., Lerner, J., and Tamayo, P.G. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.
44. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.
45. Şenbabaoğlu, Y., Michailidis, G., and Li, J.Z. (2014). Critical limitations of consensus clustering in class discovery. Sci. Rep. *4*, 6207. https://doi.org/10.1038/srep06207.
46. Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. *49*, 1476–1486. https://doi.org/10.1038/ng.3934.
47. Fan, C., Prat, A., Parker, J.S., Liu, Y., Carey, L.A., Troester, M.A., and Perou, C.M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC Med. Genomics *4*, 3. https://doi.org/10.1186/1755-8794-4-3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA-BRCA mRNA-seq data	NCI GDC	https://portal.gdc.cancer.gov dbGaP accession phs000178
Pan-Cancer mRNA-seq data	PanCanAtlas	RNA (final) https://gdc.cancer.gov/about-data/publications/pancanatlas dbGaP accession phs000178
TCGA-BRCA GISTIC2 gene-level copy number data	TCGA GDAC Firehose	http://firebrowse.org
TCGA-BRCA somatic mutation data	PanCanAtlas	https://gdc.cancer.gov/about-data/publications/pancanatlas
TCGA Cancer digital slide archive (CDSA)	Emory University	https://cancer.digitalslidearchive.org/
Pathology histologic type annotation and scoring sheet	Heng et al. (2017)[24]	PMID: 27861902
TCGA-BRCA 2016 histologic type annotations	This study	Data S1
TCGA-BRCA 2016 histologic morphologic scores	This study	Data S2
METABRIC mRNA-seq data	European Genome-Phenome Archive	accession number: EGAS00000000083
NKI mRNA-seq data	Weigelt et al. (2008)[4]	PMID: 18720457 Array Express (http://www.ebi.ac.uk/arrayexpress), experiment number E-NCMF-3
Pan-Cancer histologic type annotations	PanCanAtlas	TCGA-CDR-SupplementalTableS1.xlsx https://gdc.cancer.gov/about-data/publications/pancanatlas
TCGA Papillary carcinoma 2019 re-annotation	This study	Table S1
TCGA Metaplastic carcinoma 2019 re-annotation	This study	Table S4
Software and algorithms		
R (version 3.5)	R Development Core Team	https://www.R-project.org
Cluster 3.0 (version 1.59)	Stanford University	http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
Java Treeview (version 1.2.0)	Stanford University	http://jtreeview.sourceforge.net/
GenePattern (version 3.9.10)	Broad Institute	https://www.genepattern.org/
Differentiation Score DWD calculator (R package version)	UNC microarray database static publication website	https://genome-publications.bioinf.unc.edu/clow/
PAM50 molecular subtyping method (No version)	UNC microarray database static publication website	https://genome-publications.bioinf.unc.edu/PAM50/
Claudin low centroid predictor (No version)	UNC microarray database static publication website	https://genome-publications.bioinf.unc.edu/clow/

RESOURCE AVAILABILITY

Lead Contact

All future information regarding the datasets and results in this work can be directed to Charles M. Perou (cperou@med.unc.edu).

Materials availability

All revised histopathologic type annotations including 11 various pathologic feature scores are included in Data S2.

Data and code availability

- This study utilized previously published RNA-seq data available for TCGA-BRCA and TCGA Pan-Cancer Atlas pan-cancer datasets. There was no new sequencing data generated. The TCGA-BRCA RNA-seq dataset was downloaded from the GDC Data Portal (based on GDC data release 12) using genome reference build GRCh38.p0 (https://portal.gdc.cancer.gov). The TCGA Pan-Cancer RNA-seq data - RNA (Final) - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv was downloaded from the PanCanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas). The TCGA-BRCA GISTIC2 gene-level copy number data was downloaded from The Broad Institute TCGA GDAC Firehose with no further processing (all_data_by_genes.txt). The TCGA-BRCA somatic mutation data (Mutations – mc3.v0.2.8.PUBLIC.maf.gz) was downloaded from the PanCanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas).
- There were no new codes generated for any analysis in this paper. We used well known default functions available for our DNA copy number, RNA-seq DE gene analysis (DESeq2), and classifier (elastic net model). For RNA-seq differential gene expression analysis, we used DESeq2 (version 3.7, R software package version 3.5; https://bioconductor.org/packages/release/bioc/html/DESeq2.html). All cluster based analysis was performed using Cluster 3.0 using the C clustering library version 1.59 and heatmaps were visualized using Treeview version 1.2.0. Gene set enrichment analysis was performed using the GSEA module hosted by Genepattern browser version 3.9.10. ConsensusClusterPlus version 3.8, R package version 3.5 was used for consensus clustering analysis. GISTIC version 2.0 was used to generate gene level copy number data (https://software.broadinstitute.org/software/cprg/?q=node/31). We provide the elastic net features and weights in Figure 4 and Data S4.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Datasets used and histopathological assessment

There was no new sequencing data generated. The TCGA-BRCA RNA-seq dataset was downloaded from the GDC Data Portal (based on GDC data release 12) using genome reference build GRCh38.p0 (https://portal.gdc.cancer.gov). The TCGA Pan-Cancer RNA-seq data - RNA (Final) - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv was downloaded from the PanCanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas). The TCGA-BRCA GISTIC2 gene-level copy number data was downloaded from The Broad Institute TCGA GDAC Firehose with no further processing (all_data_by_genes.txt). The TCGA-BRCA somatic mutation data (Mutations – mc3.v0.2.8.PUBLIC.maf.gz) was downloaded from the PanCanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas). TCGA PanCanAtlas data was retrieved and processing were performed as previously described; samples were obtained from patients with appropriate consent from institutional review boards.[13] The raw data, processed data and clinical data can be downloaded from the PanCanAtlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas). TCGA histologic interpretation was made through images that were assessed via TCGA digital slide archive (CDSA) (http://cancer.digitalslidearchive.net/;[39]). Two independent pathologists reviewed each sample photomicrograph. For the diagnosis of special histologic types, 90% of tumor area exhibiting the specific morphological appearance as outlined by the WHO was set as a diagnostic criterion. Briefly, pathologists used a scoring sheet previously described by Heng et al.[24] and all these scores and diagnosis were integrated to reach a final consensus by the pathology review committee. These scores and histologic type annotations are what we term as the 2016 annotation scheme and provide in Data S2. For the 2019 re-annotation, a conference call was done, and all the papillary and metaplastic carcinoma virtual slides were discussed by K.A., L.C.C, S.J.S. In the case of papillary carcinomas, they were re-classified into encapsulated, solid and invasive as defined by the WHO 2019 classification.[2] In cases where a true fibrotic core was not observed in papillary projections, the pathologists agreed to give an annotation of IDC with pseudo-papillary features. Each case was re-annotated upon the final agreement arrived after the discussion between the 3 pathologists.

BRCA microarray Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset was obtained from the European Genome-Phenome Archive (accession number: EGAS00000000083).[40] The NKI113 microarray dataset was publicly available, retrieved[4] and preprocessed as previously described.[25] In all datasets, genes were median-centered within each dataset and samples were standardized to zero mean and unit variance before other analyses were performed.

mRNA-seq analysis

Molecular profiles were retrieved from sources as mentioned above. Intrinsic subtyping was done using the PAM50 R function as previously described[8] and additional CLOW subtyping was done as previously described.[25] Briefly, Claudin-low (CLOW) molecular subtype can be identified using breast cancer cell line gene expression centroid-based predictors where some cell-lines show low expression of claudin genes, which is one characteristic of claudin-low samples. Using the cell-line gene expression as training data, the claudin-low centroid predictor predicts bulk tumor samples that are claudin-low. To add robustness to this centroid predictor, we simultaneously also examine the hierarchical clustering of these samples using the 1800 intrinsic gene list of Parker et al.,[8] and similarly look a group/cluster of tumors that show low expression of claudin genes. Lastly we call as "claudin-low"

those samples that are identified as claudin-low by both methods. The cluster of cluster assignments (CoCa) group information was retrieved.[41]

DESeq2 (version 3.7, R software package version 3.5; https://bioconductor.org/packages/release/bioc/html/DESeq2.html) R package was used for differentially expressed gene analysis.[42] Raw mRNA data was used for TCGA-BRCA in DESeq2 differential gene analysis as the package performs its own internal normalization. Differential expressed genes were considered significant using a threshold of FDR < 0.05. Gene set enrichment analysis (GSEA) was conducted using Genepattern browser version 3.9.10.[43] For GSEA, the raw data was log transformed, upper quartile normalized and median centered. For GSEA reporting, we utilized the hallmark signatures (h.all.v6.2.symbols.gmt), immunological signatures (c7.all.v6.2.symbols.gmt), oncogenic signatures (c6.all.v6.2.-symbols.gmt), KEGG pathway signatures (c2.cp.kegg. v6.2.symbols.gmt) and the gene ontology signatures (c5.all.v6.2.symbols.gmt). In terms of the algorithm options, we set the "min.gene.size" to 10 genes.

For supervised hierarchical clustering analysis, we took the top 500 upregulated and top 500 downregulated genes for each of the six individual histologic types as identified by DESeq2 DE analysis and combined them and removed the duplicates. Supervised hierarchical clustering was done using Cluster 3.0 using the C clustering library version 1.59 and exported using Java Tree View version 1.2.0. Genelists utilized for clusterings can be found in Data S1. ConsensusClusterPlus R package version 3.8 was utilized for consensus clustering analysis.[44] CDF plots within the package were utilized to determine the optimal number of clusters.Proportion of ambiguous clustering (PAC) was calculated using the methodology described by enbabaoğlu et al.[45]

Differentiation Score Calculation

Differentiation score is an ancillary score to determine the level of differentiation of a breast tumor sample in regards to how closely the gene expression of the tumor is similar to normal mammary cell types.[25] For differentiation score, we utilized the algorithm developed previously to define a differentiation axis from DNA microarray datasets of three epithelial cell-enriched subpopulations: mammary stem cells (MaSC), luminal progenitors (pL) and mature luminal cells (mL).[25] Briefly this method utilizes distance weighted discrimination (DWD) to determine the distance of greatest variation from MaSC to pL and pL to mL. Each tumor sample was projected onto the MaSC ® pL axis and the pL ® mL axis by calculating the inner product of the sample and the MaSC or mL vectors identified by DWD. The difference of the two projected positions of each sample along the MaSC ® pL ® mL axis is referred to as the differentiation score.[25]

Elastic net modeling

We analyzed the TCGA Pan-Cancer Atlas dataset, including the histologic annotations found within the clinical details file in the PanCanAtlas publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas) – (TCGA-Clinical Data Resource (CDR) Outcome* - TCGA-CDR-SupplementalTableS1.xlsx) were used to identify mucinous histologic samples. The images of these non-BRCA cases with a mucinous histologic type annotation were then re-validated by examining their respective virtual slides in the digital archive database (http://cancer.digitalslidearchive.net/) and designated as mucinous carcinoma or MUC samples. The rest of the cancers in those tissue systems were then termed as non-MUC samples. The final cohort was 6017 cases which were divided into 30% testing and 70% training datasets balanced for MUC cases through R package – sampling version 2.8. R package – caret was used to build elastic net generalized linear models using the training dataset. Tuning grid were determined with alphas over a range from 0.1 to 1 by 0.1 and a sequence of 100 lambdas. The minimum and maximum of lambda values were determined by fitting generalized linear models with each alpha value on training set (R package glmnet version 2.0.16). 200 rounds of Monte-Carlo cross validation with default training percentage of 0.75 (R package caret version 6.0.8) were used to select the tuning parameters. The optimal parameter combination was determined to have the best classification accuracy. This model was then applied to both the training and the testing data and ROC curves were generated to identify the performance of the model using R package ROCR.

Tumor Class specific DNA copy number identification and Mutation analysis

To identify tumor class specific copy number alteration (CNA), the TCGA-BRCA GISTIC2 gene-level copy number data was downloaded from The Broad Institute TCGA GDAC Firehose with no further processing (all_data_by_genes.txt) which is available for 1070 breast cancer patients in TCGA. To identify CNA patterns across the 4 groups based on the transcriptomic differentiation score (Basal, Low Differentiation, Luminal, and High Differentiation), ANOVA followed by Tukey's post-test for pairwise comparisons was used. The ANOVA analysis used GISTIC2 gene-level gain/loss calls (−2 for high-level deletion, −1 for loss, 0 for neutral events, 1 for gain and 2 for high-level amplification) to compare copy number values among the 4 broader groups. ANOVA F statistics and p values were reported. P values were further adjusted by Benjamini-Hochberg multiple tests correction.

For individual tumor class mutation analysis, publicly available TCGA-BRCA somatic mutation data was downloaded (Mutations – mc3.v0.2.8.PUBLIC.maf.gz) (https://gdc.cancer.gov/about-data/publications/pancanatlas) and the waterfall plots were constructed using R package GenVisR version 3.9. The somatic mutation data was available for 1066 samples.

The BRCA1/2 germline and somatic events data was available for 990 TCGA samples and were annotated as bi-allelic-inactivation, mono-allelic-inactivation and epigenetic-silencing based upon previous published work.[46]

Immune Gene Signature Module Calculation and Statistical Analysis

Immune Gene Signature Module Calculation and

For each group (Basal, Low Differentiation, Luminal, and High Differentiation), we calculated 115 immune gene expression modules, representing multiple published immune related biological pathways and cell types.[47] All gene expression module scores were calculated as the median of all individual gene expression values present in the module for each sample in the 4 groups. Significant immune gene modules were then analyzed between the Low Differentiation and High Differentiation groups using the two-class Significance Analysis of Microarrays (SAM) implemented by the ''samr'' package version 3.0 in R.