The molecular basis of breast cancer pathological phenotypes

Yujing J. Heng[1,2], Susan C. Lester[3] , Gary M.K. Tse[4], Rachel E. Factor[5], Kimberly H. Allison[6], Laura C. Collins[1,2], Yunn-Yi Chen[7], Kristin C. Jensen[6,8], Nicole B. Johnson[1,2], Jong Cheol Jeong[1,2], Rahi Punjabi[1,2], Sandra J. Shin[9], Kamaljeet Singh[10], Gregor Krings[7], David A. Eberhard[11], Puay Hoon Tan[12], Konstanty Korski[13], Frederic M. Waldman[14], David A. Gutman[15], Melinda Sanders[16], Jorge S. Reis-Filho[17], Sydney R. Flanagan[1,2], Deena M.A. Gendoo[18,19], Gregory M. Chen[18], Benjamin Haibe-Kains[18,19], Giovanni Ciriello[20], Katherine A. Hoadley[21], Charles M. Perou[11, 21] and Andrew H. Beck[1,2,*]

1. Cancer Research Institute, Beth Israel Deaconess Cancer Center, Boston, MA, USA
2. Department of Pathology, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, USA
3. Department of Pathology, Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA
4. Department of Anatomical & Cellular Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, NT, Hong Kong
5. Department of Pathology, School of Medicine, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA
6. Department of Pathology, School of Medicine, Stanford Medical Center, Stanford University, Stanford, CA, USA
7. Department of Pathology, School of Medicine, University of California, San Francisco, CA, USA
8. VA Palo Alto Healthcare System, Palo Alto, CA, USA
9. Department of Pathology & Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA
10. Department of Pathology & Laboratory Medicine, Brown University, Providence, RI, USA
11. Department of Pathology & Laboratory Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC, USA
12. Department of Pathology, Singapore General Hospital, Singapore
13. Department of Pathology, Greater Poland Cancer Centre, Poznan, Poland
14. Department of Urology, School of Medicine, University of California, San Francisco, CA, USA
15. Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, USA
16. Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, TN, USA
17. Memorial Sloan Kettering Cancer Center, New York, NY, USA
18. Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
19. Departments of Medial Biophysics and Computer Science, University of Toronto, Toronto, ON, Canada
20. Department of Medical Genetics, University of Lausanne (UNIL), Lausanne, Switzerland
21. Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA

**Corresponding Author: Andrew H. Beck, Department of Pathology, Harvard Medical School, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Dana 517, Boston, MA 02115, USA; Tel: +1-617-667-4132; Email: abeck2@bidmc.harvard.edu*

Conflict of Interest: CMP is an equity stock holder, and Board of Director Member, of BioClassifier. CMP is also listed an inventor on patent applications on the Breast PAM50 assay. AHB is an equity stock holder and Board of Director Member of PathAI.

Running Title: Molecular basis of breast cancer pathological phenotypes

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.4847

Keywords: PAM50, TCGA, Bioinformatics, Genomics, mRNA, Epithelial Tubule Formation, Histological Grade

Abstract

The histopathological evaluation of morphological features in breast tumours provides prognostic information to guide therapy. Adjunct molecular analyses provide further diagnostic, prognostic and predictive information. However, there is limited knowledge of the molecular basis of morphological phenotypes in invasive breast cancer. This study integrated genomic, transcriptomic and protein data to provide a comprehensive molecular profiling of morphological features in breast cancer. Fifteen pathologists assessed 850 invasive breast cancer cases from The Cancer Genome Atlas (TCGA). Morphological features were significantly associated with genomic alteration, DNA methylation subtype, PAM50 and microRNA subtypes, proliferation scores, gene expression and/or RPPA subtype. Marked nuclear pleomorphism, necrosis, inflammation and high mitotic count were associated with Basal-like subtype and have similar molecular basis. Omics-based signatures were constructed to predict morphological features. The association of morphology transcriptome signatures with overall survival in oestrogen receptor (ER)-positive and ER-negative breast cancer was first assessed using the METABRIC dataset; signatures that remained prognostic in the METABRIC multivariate analysis were further evaluated in five additional datasets. The transcriptomic signature of epithelial tubule formation was prognostic in ER-positive breast cancer. No signature was prognostic in ER-negative. This study provided new insights into the molecular basis of breast cancer morphological phenotypes. The integration of morphological with molecular data has potential to refine breast cancer classification, predict response to therapy, enhance our understanding of breast cancer biology and improve clinical management. This work is publicly accessible at www.dx.ai/tcga_breast.

Introduction

Histopathological analysis of breast tumours plays a central role in the diagnosis of breast cancer. The assessment of histological type (e.g. invasive ductal carcinoma (IDC) or invasive lobular carcinoma (ILC)) and histological grade (a summary score of epithelial tubule formation, mitotic count and nuclear pleomorphism) are reported to guide clinical management [1–4]. The microscopic assessment of tumour infiltrating lymphocytes can predict improved response to chemotherapy and prognosis in erb-b2 receptor tyrosine kinase (HER2)-positive breast cancer [5–8]. Beyond these features, breast tumours display an array of other morphological features such as necrosis, whose clinical significance are not well characterized.

Breast cancer is a heterogeneous disease at both morphological and molecular levels. The PAM50 molecular "intrinsic" subtypes, Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like, have distinct biological properties, epidemiological risk factors, response to therapy, prognoses, and are associated with specific morphological features [9–13]. The "Normal-like" subtype is highly variable and is not reproducibly defined [14]. Morphological and molecular data complement the characterization of breast cancer phenotypes. For example, Basal-like tumours display high histological grade, necrosis, tumour infiltrating lymphocytes, fibrotic foci and are generally IDCs [15–19] while HER2-enriched displays high histological grade and may contain apocrine features and ductal carcinoma *in situ* (DCIS) [20,21].

Few studies have analysed the molecular characteristics of morphological features. These studies were limited by sample sizes (*n*=57 to 212) and investigated one to three features with one or two types of molecular data [22–25]. The Genomic Grade Index (GGI; i.e. MapQuant Dx^{TM}) is a transcriptomic signature constructed by integrating histological grade with gene expression and is associated with oestrogen receptor (ER)-positive breast cancer prognosis [22]. GGI, like most first generation prognostic signatures, is largely a measure of cellular proliferation [14,26,27]. The molecular basis of each histological grade component, nuclear pleomorphism, epithelial tubule formation and mitotic count, as well as other breast tumour morphological features remain unknown.

This study aimed to comprehensively elucidate the molecular basis of breast cancer morphological phenotypes by integrating genomic, transcriptomic and proteomic data with morphological features and determined if morphology transcriptomic signatures were prognostic in ER-positive or ER-negative breast cancer. To achieve this, a team of 15 international breast cancer pathology experts provided detailed histopathological annotation for 850 invasive breast cases in The Cancer Genome Atlas (TCGA). After we integrated the consensus assessments of 11 morphological features with TCGA's molecular profiles, we identified genomic, transcriptomic and proteomic data associated with morphological features. Next, omics-based signatures representative of morphological features were constructed and the prognostic value of each signature with overall survival was assessed using the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database [28]. Signature(s) that remained prognostic in the METABRIC multivariate analysis was further evaluated in five additional datasets.

Materials and Methods

Images and Molecular Data

TCGA data generation and processing were as previously described; samples were obtained from patients with appropriate consent from institutional review boards [29]. TCGA invasive breast cancer (n=850) images were assessed via http: //cancer.digitalslidearchive.net/ [30]. Molecular profiles were retrieved (http: //cancergenome.nih.gov/): RNAseq gene expression (Illumina HiSeq RNASeqV2 Level 3.1.9.0); DNA methylation subtype 1, 2, 3, 4 and 5 (Illumina Infinium DNA chips); microRNA subtype 1, 2, 3, 4, 5, 6 and 7 (Illumina sequencing); and reverse-phase protein assay (RPPA) subtype Basal, HER2-enriched, Luminal A, Luminal A/B, ReacI, ReacII and "X" (M.D. Anderson RPPA Core Facility). PAM50 classification and PAM50 proliferation score were computed [9,31,32].

Genomic alterations implicated in breast cancer (43 somatic mutations, 45 amplifications, 62 deletions and 6 multiple alterations, e.g. mutation and amplification) identified using Mutation Significance version 2 [33] and Genomic Identification of Significant Targets in Cancer [34] were retrieved from cBioPortal for Cancer Genomics [31]. Of a total of 156 genomic alterations, 127 genomic alterations were assessed in this study after excluding multiple alterations (*n*=6) and rare genomic alterations which only occurred in <5 cases (*n*=23). Those were excluded to allow separate analyses of mutation, amplification and deletion, and reduce spurious findings.

Histopathology Morphological Assessment

Cases were randomly assigned to the pathologists and images were graded using an electronic scoring sheet adapted from the Collage of American Pathologists' protocol for invasive breast examination [35] (Supplementary Figure S1A). For routine clinical features such as histological type, histological grade (nuclear

pleomorphism, mitotic count and epithelial tubule formation), lobular carcinoma *in situ* (LCIS) and DCIS, the pathologists used criteria applied in clinical practice. For features not commonly assessed in clinical practice, including stromal inflammation, necrosis, proportion of cancerous epithelium in invasive portion by area (i.e. epithelial area), lymphovascular invasion, stromal central fibrotic foci and apocrine features, the pathologists carried out conference calls to discuss the grading criteria and circulated images for scoring. Images with high consensus diagnoses were circulated as examples for grading. Supplementary Figure S1B shows an annotated scoring sheet with additional pathological scoring criteria and details.

To define final histological type, information from pathology reports and the pathology review committee were integrated [31]. Pathology assessments were converted to integer scores. For proportion of cancerous epithelium, in cases with discordance, discordant scores were resolved by taking the minimum value. For other morphological features, if the most frequent feature value in the dataset was the maximum of the possible feature values, discordant scores were resolved by taking the minimum value, otherwise the maximum value was used. This was done to obtain an even distribution of the scores in the final dataset. Table 1 displays the morphological features, their grading categories and frequencies.

Inter-Rater Reliability

Inter-rater reliability was assessed for each morphological feature where cases were graded (using the categories shown in Table 1) by at least two pathologists. Inter-rater reliability was calculated using Krippendorff's alpha [36] (irr, version 0.84; R, version 3.2.1) with bootstrapping (100 iterations), and average percentage agreement.

Subsequent Exclusion of Histological Type and Re-stratification of Morphological Features into Binary Groups

The molecular characterization of histological types was reported separately and we demonstrated that histological type represents a morphological continuum with a significant proportion of cases with morphological features of both ductal and lobular cancers [31]. Thus, we decided to include the full range of histological types in this manuscript to enable the robust identification of molecular profiles and signatures associated with the remaining 11 morphological phenotypes across all types of invasive breast cancer. To reduce the complexity and to increase statistical power, morphological features were re-stratified into binary categories to determine the association of each morphological feature with a type of molecular data (Table 2). For example, we investigated the association of *TP53* mutation in tumours with marked nuclear pleomorphism compared to tumours with small/moderate nuclear pleomorphism, or genes differentially expressed in tumours with DCIS compared to tumours without DCIS. All tests of statistical significance were two-sided. Statistical significance was attained when a *p*-value was <0.05 or false discovery rate (FDR) was <0.05.

Determining the Association of Morphological Features with Molecular Profiles

Genomic Data

The univariate association of genomic alteration and DNA methylation subtype with morphological features was determined using a Chi-square test with Bonferroni adjustment, and Fisher's exact test with Benjamini-Hochberg multiple testing corrections, respectively.

Transcriptomic Data

The association of PAM50 and microRNA subtypes, PAM50 proliferation score, differential gene expression and gene sets/pathways with morphological features were determined using a Chi-square test with

Bonferroni's adjustment, Wilcoxon's test, limma-voom with the Benjamini-Hochberg correction (version 3.22.1) [37], and piano (version 1.6.2) [38], respectively.

Differential gene expression (*n*=15,398) was performed in all cases (i.e. overall, *n*=826) and within each PAM50 subtype, except in Normal-like (excluded due to small sample size, *n*=24). Gene set enrichment analysis utilized C2 Molecular Signatures Database which includes gene sets from Reactome, BioCarta and KEGG (version 4.0, *n*=4646, www.broadinstitute.org/gsea/msigdb/). Gene sets distinctly up or down regulated were reported.

Proteomic Data

The association of RPPA subtype with morphological features was determined using a Chi-square test with Bonferroni adjustment.

Constructing Molecular Signatures of Morphological Features

Elastic-net regularised generalised linear models (glmnet, version 2.0-2) [39] was used to construct molecular signatures of morphological features using genomic alteration, transcriptomic or both types of data (genomic and transcriptomic). Model performances were assessed using cross-validated area under the receiver operator characteristic curve (ROC AUC). To determine which type of molecular data best predicted morphological features, the ROC AUC of models built using genomic alterations, transcriptomic or both types of data were compared using paired *t*-test.

Molecular signatures for histological grade were also constructed. A histological grade "feature" was created by summing the original scores of epithelial tubule formation, nuclear pleomorphism and mitotic count. The summed scores (ranging from 3 to 9) were then stratified into low/medium (summed scores of 3 to 7) or high histological grade (summed scores of 8 or 9).

Transcriptomic signatures predicted morphological features with the highest ROC AUCs. Thus, morphology transcriptomic signatures were subjected to bootstrapping (1000 iterations) to obtain 95% confidence intervals for gene coefficient estimates. Gene coefficients with 95% confidence intervals crossing zero were dropped from the signature. Gene set enrichment of each transcriptomic signature was performed using Gene Ontology Biological Processes summarized version (DAVID 6.7 [40]). Statistical significance was achieved when the FDR was <0.05.

Survival Analyses using METABRIC Breast Cancer Dataset

To determine if the morphological features' transcriptomic signatures were prognostic for overall survival, these signatures were compared against established proliferation-based prognostic signatures (GGI [22], OncotypeDx® [41] and MammaPrint® [42]) and PAM50 subtype using the METABRIC (*n*=1992) dataset [28].

PAM50 subtype for METABRIC were retrieved from Prat *et al.* [43]. Research-based classifications of GGI, OncotypeDx® and MammaPrint® for each woman were computed using genefu (version 3.1) [44]. Each morphological feature's signature score was calculated by subtracting the average expression of genes with negative coefficients from the average expression of genes with positive coefficients.

Cox proportional hazards model was used to assess the univariate association of clinicopathological variables (age at initial morphological diagnosis, tumour size (in centimetres), node-positive (spread to regional lymph nodes; yes/no) and clinical grade (1, 2 or 3)) [45], PAM50, GGI, OncotypeDx®, MammaPrint® and morphology transcriptomic signatures with overall survival. To ensure that the association of our morphology transcriptomic signatures with overall survival was not by chance, the Significance Analysis of Prognostic Signatures (SAPS) algorithm compared the prognostic utility of each morphology transcriptomic signature with "random" transcriptomic signatures of similar size (saps, version 2.0.0) [46]. Hence, a

morphology transcriptomic signature was only considered significant when the Cox model (Wald test) *p*-value was <0.05 and an absolute adjusted SAPS score of >1.3 was obtained.

Clinicopathological variables, PAM50, GGI, OncotypeDx® and MammaPrint® were considered significantly associated with survival when *p*-values were <0.05. Significant variables and/or signatures were subsequently evaluated in a multivariate model, adjusted by treatment (chemotherapy, hormone therapy, combined chemo and hormone therapy, or untreated). Analyses were performed separately in ER-positive and ER-negative breast cancer.

Meta-Analysis of Significant Transcriptomic Signature

The transcriptomic signature of poorly differentiated epithelial tubules remained significantly prognostic in the METABRIC multivariate analysis amongst ER-positive women. This signature was further evaluated in a meta-analysis consisting of five ER-positive breast cancer gene expression datasets: CAL [47], PNC [48], NKI [42,49,50], TRANSBIG [42,49], and GSE25066 [51]. Each gene expression dataset was pre-processed using Weighted Gene Co-Expression Network (version 1.47) [52] and annotated using lumi (version 2.20.2) [53]. These datasets were chosen because they had overall survival data or distant-relapse-free survival, clinical grade information, treatment information (for CAL, NKI and GSE25066) and at least 10,000 annotated genes. The meta-analysis adjusted for clinical grade and treatment.

Website Resource

Data are available at: www.dx.ai/tcga_breast. Detailed methodologies are given in Supplementary Materials and Methods.

Results

Pathology Morphological Dataset and Assessment of Inter-Rater Reliability

From November 2011 until March 2014, 15 pathologists completed 1,524 online scoring sheets: 11 cases were reviewed >10 times, 15 cases reviewed 5-9 times, 357 cases reviewed 2-4 times and 467 cases reviewed once. The annotations and frequencies of morphological assessments are in Table 1. The prevalence of IDC, ILC and special histological types is similar previous reports (IDC: 50-80%, ILC: 5-15%, special histological types: 1-15%). IDC/ILC cases (10.9%) in this study appear to be slightly higher than the 3-7% reported by a limited number of studies [54–59]. Supplementary Table S1 (A-D) displays the frequencies stratified by PAM50 subtype for all cases, within IDCs, ILCs or special histological types. Raw annotation data are in Supplementary Table S2. Inter-rater reliability was calculated for 383 cases that were reviewed at least twice. There is a moderate agreement among pathologists with percentage agreements ranging from 78% (mitotic count) to 98% (LCIS), respectively (Table 3).

Morphological Features are Associated with Molecular Data

Table 4 summarizes the association of each morphological feature with various molecular data (details in Supplementary Table S3 and Figure S2). Differential gene expression was performed in all cases and within each PAM50 subtype, except in Normal-like (Supplementary Table S4A). Due to small sample sizes, differential gene expression associated with the presence of LCIS was performed for all cases and within Luminal A.

Inflammation, Necrosis, Nuclear Pleomorphism and Mitotic Count

Inflammation, necrosis, marked nuclear pleomorphism and medium/high mitotic counts co-occur in the tumours and share many genomic alterations (FDR<0.05; Figure 1, Table 4). In the TCGA publication, DNA methylation subtype 3 and 5 are enriched for Luminal B and Basal-like, respectively, while TCGA microRNA

subtypes 4 and 5 are associated with Basal-like [29]. Therefore, the presence of inflammation and necrosis, marked nuclear pleomorphism and medium/high mitotic counts are distinctly associated with the highly proliferative Basal-like subtype (Table 4).

These four morphological features have 15 common up-regulated genes involved in cellular proliferation (*MYBL2*, *CHEK1, CENPA, MELK, MEMO1, NASP, RCC2, LTV1* and *C1orf135*) [60,61], MYC activation (*CDCA7*) [62–65], and DNA and RNA metabolism (*PIF1, RBM17, AMD1* and *RPIA*). The function of *C17orf96* is unknown. These four morphological features also have 13 common down-regulated genes involved in membrane signalling *(CBLN4, ELFN1, LTBP3, LRP10, TENC1* and *TPCN1*) including GTPase activity (*RAPGEF3*, *TBC1D13*), transcription (*CAMTA2, CRY2* and *LOC653501),* cytoskeleton (*KIF13B)* and lysosome positioning (*C10orf32)*. The plethora of differentially expressed genes associated with necrosis, marked nuclear pleomorphism and medium/high mitotic counts were enriched for *proliferation* gene sets, while the presence of stromal inflammation was enriched for *inflammation* gene signatures (FDR <0.05; Figure 2; detailed heatmap in Supplementary Figure S3). Collectively, molecular data suggest that tumorigenesis involving these four proliferative Basal-like morphological features may be driven by MYB- and MYC-regulated pathways, and potentially in conjunction with *TP53* pathways, in invasive breast cancer [60,66,67].

Epithelial Tubule Formation

Histological grade harbours genomic alterations of *TP53*, 8q24.21 (*MYC*), 19q12 (*CCNE1*), 20p13.2 (*ZNF217*) and 9p21.3 (*MTAP*) [24,25]. Focusing on the components of histological grade, poorly differentiated epithelial tubules share only a few molecular traits with medium/high mitotic counts and marked nuclear pleomorphism: *TP53* mutation, high PAM50 proliferative score, Basal-like subtype classified using methylation and microRNA data. Molecular traits of poorly differentiated epithelial tubules were common with LCIS (i.e. *CDH1* mutation, PAM50 Luminal A subtype and *inflammation* gene sets) although there is no correlation

between the two morphological features (Supplementary Table S3B). *P2RY11*, a G-protein coupled receptor activated by extracellular adenosine and uridine [68], was the top differentially expressed gene (2.3-fold increase) in tumours with poorly differentiated epithelial tubules compared to well/moderately differentiated epithelial tubules. The role of *P2RY11* in breast tissue remains unknown and could be evaluated as a potential pharmacological target.

LCIS

LCIS are precursor lesions for ILC, defined by the hallmark *CDH1* loss-of-function mutation and are almost exclusive to Luminal A tumours [31,69]. Regardless of histological type, the presence of LCIS is also associated with DNA methylation subtype 1, downregulation of *proliferation* gene sets and enriched for cytokines/immune-signalling pathways (despite not being associated with the presence of morphological inflammation). *HMGCS2,* a breast apocrine carcinoma marker involved in the anabolic ketogenesis pathway, was increased by 8.6-fold in tumours with LCIS (Supplementary Table S4B) [70]. Other top ranking up-regulated genes were involved in inflammation (*GP2* and *C7*) [71], alcohol (*ADH1B*) and fat metabolism (*ADIPOQ*) [71–73], transcription (*TFAP2B*), are transmembrane proteins (*TMEM132C* and *SLC7A4*) as well as genes with unknown function (*TFF1* and *TUSC5*) [74,75]. At the same time, the mRNA expression levels of extracellular matrix proteins (*MMP1, CDH1, EPYC, COL11A1, HAPLN1* and *IBSP*) were significantly lower. Thus, *APIPOQ* and *HMGCS2* overexpression suggest that the manifestation of LCIS may reflect abnormal hormone and fatty acid levels in the breast tissues, impaired fatty acid oxidation and mitochondrial dysfunction [76,77]. Mitochondrial dysfunction can lead to inflammation, tumorigenesis, dysregulation of cell-cell adhesion, dis-cohesive morphology and invasion [76,78–81]. These characteristics of mitochondria dysregulation are supported by our differential gene expression analyses. It would also be interesting to investigate the association of lifestyle factors such as obesity or alcohol consumption with LCIS or histological type [72,82,83].

DCIS

The co-existence of DCIS with prominent features (i.e. strong molecular profiles) such as marked nuclear pleomorphism and poorly differentiated epithelial tubules may have masked our ability to decipher more molecular basis of DCIS (Supplementary Table S3B). DCIS was associated with 40 differentially expressed genes and enriched for *proliferation* and cell-cell junction pathways (Supplementary Figure S3). The top ranking up-regulated genes in breast cancers associated with DCIS were epithelial (*CALML3*, *ANXA8L1*, *ANXA8*) [84,85] and extracellular matrix proteins (*KRT14, KRT6B, KRT17* and *MMP10*) [86], desmosomes (*DSG3 and DSC3*) that connect adjacent myoepithelial cells [87], involved in myoepithelial cell differentiation (*ACTA1*) [88] and *CCL21*-related chemotaxis resulting in epithelial-mesenchymal transition and metastasis [89–91]. These results support reports that the progression of DCIS to invasive breast cancer is influenced by changes in microenvironmental factors, especially in myoepithelial cells [87,92]. Proliferating cancerous ductal cells exert pressure against the myoepithelial cells and basement membrane. When the myoepithelial cells cannot sustain the pressure and rate of basement membrane turnover, they lose their cell-cell adhesion capabilities and allow the cancerous cells to invade into the surround tissues [87]. The 12 down-regulated genes in breast cancers associated with DCIS are newly-associated with breast cancer (cytoskeleton-related (*HOOK2* and *ARHGEF18*), mitochondria iron-sulphur cluster assembly pathway (*C1orf69*), gene regulation (*MAFG* and *WDR37*), GTPase activity (*TBC1D13* and *RAB43*), lipid synthesis (*CLN8*) and neuronal components (*PRX, LOC100130093* and *OPA3*; Supplementary Table S4B). Their involvement in DCIS and/or invasive breast cancer warrant further elucidation.

Apocrine Features

This is the first study to characterize the molecular basis of apocrine features (Table 4). Upregulated genes and enriched pathways associated with marked apocrine features include increased lipid and membrane transport (*ABCC2, ABCA12, ABCC11, HAPLN1, FIBCD1* and *FAM155B*), lipid and/or cell metabolism (*DHRS2*, *HGD, IYD* and *HHIPL2*), apoptotic, diabetes and cholesterol pathways. Down-regulated genes and pathways with marked apocrine features were gastro-peptides (*NPYIR* and *PI16*), serine peptidase (*KLK11*), alcohol/drug metabolism (*ADH1B* and *CYP4F22*), involved in secretion (*AQP5*), extracellular matrix (*HAPLN1*) and cytokine signalling (*C7* and *DARC*). These genes have been investigated as markers of proliferation or metastasis [93–99], breast cancer risk [100], prognosis [101,102] and response to therapy [103–106]. Our work suggest drug resistance may occur in tumours with marked apocrine features with overexpression of ATP-binding cassette transporters mRNA, and new drugs targeting aquaporin water channels may not work in these tumours [103,107].

Lymphovascular Invasion and Fibrotic Foci

Neither genomic alteration nor PAM50 subtype was associated with the presence of lymphovascular invasion [25], fibrotic foci or high proportion of cancerous epithelium. IL12 and integrin-related neutrophil pathways and extracellular matrix organization gene sets were down-regulated with the presence of lymphovascular invasion. The presence of fibrotic foci was linked to increased integrin and extracellular matrix organization gene sets, and down-regulated *inflammation* gene sets. The lack of distinct molecular profiles for lymphovascular invasion and fibrotic foci may be attributed to their low frequencies and suggests that these features remain largely morphologic.

Transcriptomic Signatures of Morphological Features

Genomic alterations, gene expression or both data types were used to construct signatures of morphological features. The ROC AUCs of multivariate models built using transcriptomic and combined data outperformed models constructed using genomic alterations ($p \leq 0.001$; Supplementary Table S5A). There was no difference in the ROC AUCs between transcriptomic and combined data indicating that the addition of genomic alteration data did not enhance the performance of transcriptomic signatures to predict morphological features ($p=0.139$; Supplementary Table S5B). Thus, only transcriptomic signatures were subjected to bootstrapping and further explored. Transcriptomic signatures of morphological features ranged from one gene (*LRRC32*) for proportion of cancerous epithelium to 110 genes for poorly differentiated epithelial tubules (Supplementary Table S5C,D).

The stromal inflammation signature is enriched for the suppression of T-cell activation, driven by its strongest (positive) coefficient, *CTLA4.* The increase in CTLA4 in breast cancer prevents anti-tumour T-cell response [108]. Its monoclonal antibody, anti-CTLA4, when used in synergy with other therapeutic agents (e.g. trastuzumab), blocks immune checkpoints, induces anti-tumour immunity resulting in tumour regression in preclinical (HER2) breast cancer models [109–112]. However, the blocking of immune checkpoints using antibodies of programmed cell death protein 1 and its ligand, is more effective than using anti-CTLA4 [109,113,114]. Future work could evaluate if this inflammation signature can identify women who may benefit from anti-CTLA4 therapy as well as investigating how CTLA4 contributes to tumour immunity [115].

Signatures for medium/high mitotic count, marked nuclear pleomorphism and high histological grade were enriched for cell proliferation, further confirming that these features are proliferation-related (Supplementary Table S5E). No enrichment was obtained for other signatures.

Epithelial Tubule Formation Transcriptomic Signature was Prognostic in ER-Positive Breast Cancer

In METABRIC ER-positive women (*n* = 1494), age at initial morphological diagnosis, tumour size, node-positive and the transcriptomic signature for poorly differentiated epithelial tubules remained prognostic in the multivariate model ($p<0.05$; Table 5). In ER-negative women (*n* = 434), no feature was prognostic in the multivariate model.

The transcriptomic signature for poorly differentiated epithelial tubules in ER-positive women was further evaluated in a meta-analysis across five publicly available gene expression datasets (Figure 3). The summary hazard ratio was 1.94 (95% confidence interval 1.51 to 2.38).

Epithelial Tubule Formation Transcriptomic Signature is Distinct and Least-Correlated with Proliferation

The transcriptomic signature for poorly differentiated epithelial tubules is distinct from signatures for medium/high mitotic count, marked nuclear pleomorphism and high histological grade (Figure 4). To determine if the transcriptomic signature for poorly differentiated epithelial tubules was the least correlated with proliferation, the PAM50 proliferation score for each woman in the METABRIC dataset was calculated and correlated with the transcriptomic signature scores of nuclear pleomorphism, mitotic count and epithelial tubule formation. PAM50 proliferation scores were more highly correlated with medium/high mitotic count (Spearman's *rho*=0.878 (ER-positive) and 0.919 (ER-negative)) and marked nuclear pleomorphism (*rho*=0.852 and 0.904) than poorly differentiated epithelial tubules (*rho*=0.351 and 0.616) in ER-positive and ER-negative invasive breast cancer ($p<0.001$).

Discussion

Little is known about the molecular characteristics of various morphological features in invasive breast cancer. We comprehensively unravelled the molecular portraits of breast cancer histopathological phenotypes by bridging histopathological annotations with the molecular profiles in the TCGA database. This manuscript represents the largest cross-section of cases and pathologists to examine breast cancer histopathological phenotypes to date. Our data support the central role of proliferation driving histological grade. Inflammation, necrosis, medium/high mitotic count and marked nuclear pleomorphism frequently co-exist in breast tumours, are associated with Basal-like subtypes and have similar molecular basis. LCIS has a distinct molecular profile that may be linked to mitochondria dysfunction while genes differentially expressed with DCIS are intimately associated with myoepithelial cells. Lymphovascular invasion and fibrotic foci are mainly morphological with few significant molecular traits.

Some morphological features harbour molecular traits that may confer drug resistance or serve as pharmacological targets. Our signatures can act as surrogate representation for morphological features enabling future studies to link the signatures to response to therapy with the long-term aim of improving clinical management. Personalised or refined breast cancer classification can be achieved by combining the observation of morphological features with molecular and immunohistochemistry data. Collectively, this study provided new insights into the molecular basis of breast cancer morphological phenotypes, and can potentially facilitate the future development of diagnostic and prognostic tools for breast cancer.

Most databases, including the METABRIC, do not provide separate epithelial tubule formation scores. We were unable to directly determine if the pathological measure of epithelial tubules is independently prognostic; or if our transcriptomic signature of epithelial tubule formation adds prognostic information or is superior to pathological assessment. However, if high histological grade can function as a surrogate for poorly differentiated epithelial tubules, our multivariate analyses show that the epithelial tubule formation signature is more prognostic than clinical histological grade, and indirectly demonstrates that our signature adds prognostic information for ER-positive breast cancer. Nevertheless, prognostic signatures for ER-positive are

well established [22,41,42], more research is needed to discover clinically useful prognostic signatures for ER-negative breast cancer.

At the molecular level, epithelial tubule formation is least-similar to mitotic count and nuclear pleomorphism, and shares traits with LCIS and inflammation. The transcriptomic signature for poorly differentiated epithelial tubules is distinct from high histological grade but not significantly enriched for any gene sets. The signature's genes are involved in proliferation, mitochondria metabolism, membrane signalling, cellular adhesion, oxidative stress, extracellular matrix organization and inflammation. These gene functions are a mix of selective molecular traits associated with medium/high mitotic count, marked nuclear pleomorphism, LCIS and inflammation. We speculate that our transcriptomic signature for poorly differentiated epithelial tubules is unique and superiorly prognostic because it contains genes that represent a wide range of tumour biology.

The failure to detect any association between DCIS, fibrotic foci or apocrine features with PAM50 subtypes may be due our studying utilizing PAM50 classification by molecular data instead of immunohistochemistry, different grading criteria and investigating these features within tumours of invasive breast cancer [116–119]. Our transcriptomic signature for fibrotic foci was not prognostic despite previous work reporting that IDC or Luminal B tumours with fibrotic foci have poorer prognosis [120,121]. The relevance of fibrotic foci as a prognostic factor requires further investigation and should take into consideration its size, breast cancer histological type and PAM50 classification.

Fourteen TCGA cases were inadequate for scoring due to poor image quality or insufficient invasive cancer present. Despite adhering to clinical definitions or agreed consensus scoring criteria, our histopathological analyses may be influenced by the variation in histology quality and using images instead of slides for scoring. For example, the high power field to count mitotic bodies at the highest magnification (40x) on a web browser is influenced by computer monitor size (hence high power field for each pathologist varies) and the difficulty to distinguish between mitotic figures from pynknotic nuclei due to lack of a Z axis. The pathologists used their best judgment in counting cells in mitosis. However, the mitotic count (as it was

scored) was highly concordant with PAM50 proliferation score and enriched for proliferation gene sets in this study, signifying that both mitotic count and gene expression were adequately tracking proliferation. Another limitation of this study is that we focused exclusively on a set of known morphological features that could be scored manually by experienced breast pathologists. It is likely that there are additional morphological patterns (e.g., various types of stromal reaction patterns) beyond those included in our study that are biologically important and will provide additional insight into the molecular underpinnings of breast cancer pathology.

In conclusion, breast tumour pathological phenotypes are driven by distinct underlying sets of molecular alterations. The integration of morphological with molecular data has great potential to refine breast cancer classification, predict response to therapy, enhance our understanding of breast cancer biology and improve clinical management.

Acknowledgements

The data used in this study were in whole or part based upon the data generated by the TCGA Research Network: http: //cancergenome.nih.gov/. Funding of this project was provided by the Klarman Family Foundation (AHB), the National Cancer Institute of the National Institutes of Health (SPORE grant P50CA168504; AHB) and the National Library of Medicine of the National Institutes of Health Career Development Award (Number K22LM011931; AHB).

Author Contributions

AHB and CMP were involved in the conception and design of the study. AHB, SCL, GMKT, REF, KHA, LCC, YC, KCJ, NBJ, SJS, KS, GK, DAE, PHT, KK, FMW and JSRF provided pathology annotations and clinical input. YJH, AHB, BHK, DMAG, GMC, GC, KAH and CMP retrieved and acquired other relevant data. YJH and AHB analysed and interpreted the data. JCJ, SRF, RP, DMAG, BHK, GMC and DAG provided website and/or database support. YJH, AHB, CMP, SCL, GMKT, DMAG, BHK, KCJ MS, JSRF, PHT, SJS, DAE, REF, KAH and KK were all involved in the writing, reviewing and revision of the manuscript.

References

1 Elston CW, Elston CW, Ellis IO, *et al.* Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991; **19**: 403–10. DOI:10.1111/j.1365-2559.1991.tb00229.x

2 Galea MH, Blamey RW, Elston CE, *et al.* The Nottingham prognostic index in primary breast cancer. *Breast Cancer Res Treat* 1992; **22**: 207–19. DOI:10.1007/BF01840834

3 Loi S, Sirtaine N, Piette F, *et al.* Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98. *J Clin Oncol* 2013; **31**: 860–7. DOI:10.1200/JCO.2011.41.0902

4 Rakha EA, Reis-Filho JS, Baehner F, *et al.* Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* 2010; **12**: 207. DOI:10.1186/bcr2607

5 Salgado R, Denkert C, Campbell C, *et al.* Tumor-Infiltrating Lymphocytes and Associations With Pathological Complete Response and Event-Free Survival in HER2-Positive Early-Stage Breast Cancer Treated With Lapatinib and Trastuzumab: A Secondary Analysis of the NeoALTTO Trial. *JAMA Oncol* 2015; **1**: 448–54. DOI:10.1001/jamaoncol.2015.0830

6 Adams S, Gray RJ, Demaria S, *et al.* Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. *J Clin Oncol* 2014; **32**: 2959–66. DOI:10.1200/JCO.2013.55.0491

7 Gentles AJ, Newman AM, Liu CL, *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* 2015; **21**: 938–45. DOI:10.1038/nm.3909

8 Denkert C, Loibl S, Noske A, *et al.* Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol* 2010; **28**: 105–13. DOI:10.1200/JCO.2009.23.7370

9 Parker JS, Mullins M, Cheung MCU, *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; **27**: 1160–7. DOI:10.1200/JCO.2008.18.1370

10 Sorlie T, Perou CM, Tibshirani R, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001; **98**: 10869–74. DOI:10.1073/pnas.191367098

11 Millikan RC, Newman B, Tse CK, *et al.* Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat* 2008; **109**: 123–39. DOI:10.1007/s10549-007-9632-6

12 Chia SK, Bramwell VH, Tu D, *et al.* A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clin Cancer Res* 2012; **18**: 4465–72. DOI:10.1158/1078-0432.CCR-12-0286

13 Perou CM, Sørlie T, Eisen MB, *et al.* Molecular portraits of human breast tumours. *Nature* 2000; **406**: 747–52. DOI:10.1038/35021093

14 Wirapati P, Sotiriou C, Kunkel S, *et al.* Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008; **10**: R65. DOI:10.1186/bcr2124

15 Livasy CA, Karaca G, Nanda R, *et al.* Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Mod Pathol* 2006; **19**: 264–71. DOI:10.1038/modpathol.3800528

16 Fulford LG, Easton DF, Reis-Filho JS, *et al.* Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast. *Histopathology* 2006; **49**: 22–34. DOI:10.1111/j.1365-2559.2006.02453.x

17 Carey LA, Perou CM, Livasy CA, *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 2006; **295**: 2492–502. DOI:10.1001/jama.295.21.2492

18 Dabbs DJ, Chivukula M, Carter G, *et al.* Basal phenotype of ductal carcinoma in situ: recognition and immunohistologic profile. *Mod Pathol* 2006; **19**: 1506–11. DOI:10.1038/modpathol.3800678

19 Reis-Filho JS, Milanezi F, Steele D, *et al.* Metaplastic breast carcinomas are basal-like tumours. *Histopathology* 2006; **49**: 10–21. DOI:10.1111/j.1365-2559.2006.02467.x

20 Tamimi RM, Baer HJ, Marotti J, *et al.* Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Res* 2008; **10**: R67. DOI:10.1186/bcr2128

21 Allison KH. Molecular pathology of breast cancer: What a pathologist needs to know. *Am J Clin Pathol* 2012; **138**: 770–80. DOI:10.1309/AJCPIV9IQ1MRQMOO

22 Sotiriou C, Wirapati P, Loi S, *et al.* Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006; **98**: 262–72. DOI:10.1093/jnci/djj052

23 Farshid G, Balleine RL, Cummings M, *et al.* Morphology of breast cancer as a means of triage of patients for BRCA1 genetic testing. *Am J Surg Pathol* 2006; **30**: 1357–66. DOI:10.1097/01.pas.0000213273.22844.1a

24 Fidalgo F, Rodrigues TC, Pinilla M, *et al.* Lymphovascular invasion and histologic grade are associated with specific genomic profiles in invasive carcinomas of the breast. *Tumor Biol* 2015; **36**: 1835–48. DOI:10.1007/s13277-014-2786-z

25 Langerod A AL, Zhao HH, Borgan OO, *et al.* TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res* 2007; **9**: R30. DOI:10.1186/bcr1675

26 Desmedt C, Haibe-Kains B, Wirapati P, *et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin Cancer Res* 2008; **14**: 5158–65. DOI:10.1158/1078-0432.CCR-07-4756

27 Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011; **7**: e1002240. DOI:10.1371/journal.pcbi.1002240

28 Curtis C, Shah SP, Chin S-F, *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; **486**: 346–52. DOI:10.1038/nature10983

29 The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70. DOI: 10.1038/nature11412

30 Gutman DA, Cobb J, Somanna D, *et al.* Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inf Assoc* 2013; **20**: 1091–8. DOI:10.1136/amiajnl-2012-001469

31 Ciriello G, Gatza ML, Beck AH, *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 2015; **163**: 506–19. DOI:10.1016/j.cell.2015.09.033

32 Nielsen TO, Parker JS, Leung S, *et al.* A comparison of PAM50 intrinsic subtyping with

immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res* 2010; **16**: 5222–32. DOI:10.1158/1078-0432.CCR-10-1282

33 Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**: 214–8. DOI:10.1038/nature12213

34 Mermel CH, Schumacher SE, Hill B, *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011; **12**: R41. DOI:10.1186/gb-2011-12-4-r41

35 Lester SC, Bose S, Chen Y-Y, *et al.* Protocol for the Examination of Specimens From Patients With Invasive Carcinoma of the Breast. *Arch Pathol Lab Med* 2009; **133**: 1515–38.

36 Shelley M, Krippendorff K. Content Analysis: An Introduction to its Methodology. *J Am Stat Assoc* 1984; **79**: 240. DOI:10.2307/2288384

37 Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**: e47. DOI:10.1093/nar/gkv007

38 Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* 2013; **41**: 4378–91. DOI:10.1093/nar/gkt111

39 Kodama I, Niida S, Sanada M, *et al.* Estrogen regulates the production of VEGF for osteoclast formation and activity in op/op mice. *J Bone Miner Res* 2004; **19**: 200–6. DOI:10.1359/JBMR.0301229

40 Dennis Jr G, Sherman BT, Hosack DA, *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; **4**: P3. DOI:10.1186/gb-2003-4-9-r60

41 Cristofanilli M. A multigene assay for predicting the recurrence of tamoxifen-treated, node-negative breast cancer. *Breast Dis* 2005; **16**: 219–20. DOI:10.1016/S1043-321X(05)80168-7

42 van 't Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Lett To Nat* 2002; **415**: 530–6. DOI:10.1038/415530a

43 Prat A, Carey LA, Adamo B, *et al.* Molecular features and survival outcomes of the intrinsic subtypes within HER2-positive breast cancer. *J Natl Cancer Inst* 2014; **106**: dju152. DOI:10.1093/jnci/dju152

44 Gendoo DMA, Ratanasirigulchai N, Schröder MS, *et al.* Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* 2016; **32**: 1097–9. DOI:10.1093/bioinformatics/btv693

45 Haibe-Kains B, Desmedt C, Loi S, *et al.* A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst* 2012; **104**: 311–25. DOI:10.1093/jnci/djr545

46 Beck AH, Knoblauch NW, Hefti MM, *et al.* Significance Analysis of Prognostic Signatures. *PLoS Comput Biol* 2013; **9**: e1002875. DOI:10.1371/journal.pcbi.1002875

47 Chin K, DeVries S, Fridlyand J, *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 2006; **10**: 529–41. DOI:10.1016/j.ccr.2006.10.009

48 Dedeurwaerder S, Desmedt C, Calonne E, *et al.* DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med* 2011; **3**: 726–41. DOI:10.1002/emmm.201100801

49 van de Vijver MJ, He YD, van 't Veer LJ, *et al.* A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med* 2002; **347**: 1999–2009. DOI:10.1056/NEJMoa021967

50 Desmedt C, Piette F, Loi S, *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007; **13**: 3207–14. DOI:10.1158/1078-0432.CCR-06-2765

51 Hatzis C, Pusztai L, Valero V, *et al.* A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer. *JAMA* 2011; **305**: 1873–81. DOI:10.1001/jama.2011.593

52 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**: 559. DOI:10.1186/1471-2105-9-559

53 Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. *Bioinformatics* 2008; **24**: 1547–8. DOI:10.1093/bioinformatics/btn224

54 Weigelt B, Geyer FC, Reis-Filho JS. Histological types of breast cancer: How special are they? *Mol Oncol* 2010; **4**: 192–208. DOI:10.1016/j.molonc.2010.04.004

55 Albrektsen G, Heuch I, Thoresen SØ, *et al.* Histological type and grade of breast cancer tumors by parity, age at birth, and time since birth: a register-based study in Norway. *BMC Cancer* 2010; **10**: 226. DOI:10.1186/1471-2407-10-226

56 Li CI, Uribe DJ, Daling JR. Clinical characteristics of different histologic types of breast cancer. *Br J Cancer* 2005; **93**: 1046–52. DOI:10.1038/sj.bjc.6602787

57 Phipps AI, Li CI. Breast Cancer Biology and Clinical Characteristics. In: *Breast Cancer Epidemiology*, Li CI (ed). Springer: New York, 2010; 21–46. DOI:10.1007/978-1-4419-0685-4_2

58 Dillon D, Guidi A, Schnitt S. Pathology of Invasive Breast Cancer. In: *Diseases of the Breast*, (5th edn), Harris J, Lippman M, Morrow M, *et al.* (eds). Lippincott Williams & Wilkins Health: Philadelphia, 2014.

59 Arps DP, Healy P, Zhao L, *et al.* Invasive ductal carcinoma with lobular features: a comparison study to invasive ductal and invasive lobular carcinomas of the breast. *Breast Cancer Res Treat* 2013; **138**: 719–26. DOI:10.1007/s10549-013-2493-2

60 Thorner A, Hoadley K, Parker JS, *et al.* In vitro and in vivo analysis of B-Myb in basal-like breast cancer. *Oncogene* 2009; **28**: 742–51. DOI:10.1038/onc.2008.430

61 Ma CX, Cai S, Li S, *et al.* Targeting Chk1 in p53-deficient triple-negative breast cancer is therapeutically beneficial in human-in-mouse tumor models. *J Clin Invest* 2012; **122**: 1541–52. DOI:10.1172/JCI58765

62 Osthus RC, Karim B, Prescott JE, *et al.* The Myc target gene JPO1/CDCA7 is frequently overexpressed in human tumors and has limited transforming activity in vivo. *Cancer Res* 2005; **65**: 5620–7. DOI:10.1158/0008-5472.CAN-05-0536

63 Goto Y, Hayashi R, Muramatsu T, *et al.* JPO1/CDCA7, a novel transcription factor E2F1-induced protein, possesses intrinsic transcriptional regulator activity. *Biochim Biophys Acta - Gene Struct Expr* 2006; **1759**: 60–8. DOI:10.1016/j.bbaexp.2006.02.004

64 Prescott JE, Osthus RC, Lee LA, *et al.* A Novel c-Myc-responsive Gene, JP01, Participates in Neoplastic Transformation. *J Biol Chem* 2001; **276**: 48276–84. DOI:10.1074/jbc.M107357200\rM107357200 [pii]

65 Gill RM, Gabor T V, Couzens AL, *et al.* The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation. *Mol Cell Biol* 2013; **33**: 498–513. DOI:10.1128/MCB.00276-12

66 Xu J, Chen Y, Olopade OI. MYC and Breast Cancer. *Genes Cancer* 2010; **1**: 629–40. DOI:10.1177/1947601910378691.

67 Ulz P, Heitzer E, Speicher MR. Co-occurrence of MYC amplification and TP53 mutations in human cancer. *Nat Genet* 2016; **48**: 104–6. DOI:10.1038/ng.3468

68 Moore DJ, Chambers JK, Wahlin JP, *et al.* Expression pattern of human P2Y receptor subtypes: a quantitative reverse transcription-polymerase chain reaction study. *Biochim Biophys Acta* 2001; **1521**: 107–19. DOI:S0167478101002913 [pii]

69 Vos CB, Cleton-Jansen AM, Berx G, *et al.* E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. *Br J Cancer* 1997; **76**: 1131–3.

70 Gromov P, Espinoza JA, Talman ML, *et al.* FABP7 and HMGCS2 are novel protein markers for apocrine differentiation categorizing apocrine carcinoma of the breast. *PLoS One* 2014; **9**: e112024. DOI:10.1371/journal.pone.0112024

71 Clive KS, Tyler JA, Clifton GT, *et al.* The GP2 peptide: A HER2/neu-based breast cancer vaccine. *J Surg Oncol* 2012; **105**: 452–8. DOI:10.1002/jso.21723

72 Karaduman M, Bilici A, Ozet A, *et al.* Tissue levels of adiponectin in breast cancer patients. *Med Oncol* 2007; **24**: 361–6. DOI:MO: 24: 4: 361 [pii]

73 Libby EF, Liu J, Li Y, *et al.* Globular adiponectin enhances invasion in human breast cancer cells. *Oncol Lett* 2016; **11**: 633–41. DOI:10.3892/ol.2015.3965

74 Gillesby B, Zacharewski T. pS2 (TFF1) levels in human breast cancer tumor samples: correlation with clinical and histological prognostic markers. *Breast Cancer Res Treat* 1999; **2**: 253–65. DOI:10.1023/a:1006215310169

75 Bubnov V, Moskalev E, Petrovskiy Y, *et al.* Hypermethylation of TUSC5 genes in breast cancer tissue. *Exp Oncol* 2012; **34**: 370–2.

76 Camarero N, Mascaró C, Mayordomo C, *et al.* Ketogenic HMGCS2 Is a c-Myc Target Gene Expressed in Differentiated Cells of Human Colonic Epithelium and Down-Regulated in Colon Cancer Ketogenic HMGCS2 Is a c-Myc Target Gene Expressed in Differentiated Cells of Human Colonic Epithelium and Down-Regulat. *Mol Cancer Res* 2006; **4**: 645–53. DOI:10.1158/1541-7786.MCR-05-0267

77 Le May C, Pineau T, Bigot K, *et al.* Reduced hepatic fatty acid oxidation in fasting PPARα null mice is due to impaired mitochondrial hydroxymethylglutaryl-CoA synthase gene expression. *FEBS Lett* 2000; **475**: 163–6. DOI:10.1016/S0014-5793(00)01648-3

78 Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. *Nat Rev Cancer* 2013; **13**: 227–32. DOI:10.1038/nrc3483

79 Yadava N, Schneider SS, Jerry DJ, *et al.* Impaired mitochondrial metabolism and mammary carcinogenesis. *J Mammary Gland Biol Neoplasia* 2013; **18**: 75–87. DOI:10.1007/s10911-012-9271-3

80 Kamp DW, Shacter E, Weitzman SA. Chronic inflammation and cancer: the role of the mitochondria. *Oncology* 2011; **25**: 400–410,413.

81 Jeong YJ, Bong JG, Park SH, *et al.* Expression of leptin, leptin receptor, adiponectin, and adiponectin receptor in ductal carcinoma in situ and invasive breast cancer. *J Breast Cancer* 2011; **14**: 96–103. DOI:10.4048/jbc.2011.14.2.96

82 Shield KD, Soerjomataram I, Rehm J. Alcohol Use and Breast Cancer: A Critical Review. *Alcohol Clin Exp Res* 2016; **40**: 1166–81. DOI:10.1111/acer.13071

83 Kaklamani VG, Hoffmann TJ, Thornton TA, *et al.* Adiponectin pathway polymorphisms and risk of breast cancer in African Americans and Hispanics in the Women's Health Initiative. *Breast Cancer Res Treat*

2013; **139**: 461–8. DOI:10.1007/s10549-013-2546-6

84 Stein T. Annexin A8 Is Up-Regulated During Mouse Mammary Gland Involution and Predicts Poor Survival in Breast Cancer. *Clin Cancer Res* 2005; **11**: 6872–9. DOI:10.1158/1078-0432.CCR-05-0547

85 Rogers MS, Foleyt MA, Crotty TB, *et al.* Loss of Immunoreactivity for Human Calmodulin-Like Protein is an Early Event in Breast Cancer Development. *Neoplasia* 1999; **1**: 220–5. DOI:http: //dx.doi.org/10.1038/sj.neo.7900029

86 Köhrmann A, Kammerer U, Kapp M, *et al.* Expression of matrix metalloproteinases (MMPs) in primary human breast cancer and breast cancer cell lines: New findings and review of the literature. *BMC Cancer* 2009; **9**: 188. DOI:10.1186/1471-2407-9-188

87 Adriance MC, Inman JL, Petersen OW, *et al.* Myoepithelial cells: good fences make good neighbors. *Breast Cancer Res* 2005; **7**: 190–7. DOI:10.1186/bcr1286

88 Tamiolakis D, Papadopoulos N, Cheva A, *et al.* Immunohistochemical expression of alpha-smooth muscle actin in infiltrating ductal carcinoma of the breast with productive fibrosis. *Eur J Gynaecol Oncol* 2002; **23**: 469–71.

89 Li F, Zou Z, Suo N, *et al.* CCL21/CCR7 axis activating chemotaxis accompanied with epithelial-mesenchymal transition in human breast carcinoma. *Med Oncol* 2014; **31**: 180. DOI:10.1007/s12032-014-0180-8

90 Weitzenfeld P, Kossover O, Körner C, *et al.* Chemokine axes in breast cancer: factors of the tumor microenvironment reshape the CCR7-driven metastatic spread of luminal-A breast tumors. *J Leukoc Biol* 2016; **99**: 1009–25. DOI:10.1189/jlb.3MA0815-373R

91 Pang M., Georgoudaki A., Lambut L, *et al.* TGF-Beta1-induced EMT promotes targeted migration of breast cancer cells through the lymphatic system by the activation of CCR7/CCL21-mediated chemotaxis. Oncogene. 2015; **35**: 1–13. DOI:10.1038/onc.2015.133

92 Schnitt SJ. The transition from ductal carcinoma in situ to invasive breast cancer: the other side of the coin. *Breast Cancer Res* 2009; **11**: 101. DOI:10.1186/bcr2228

93 Crawford NPS, Walker RC, Lukes L, *et al.* The Diasporin Pathway: A tumor progression-related transcriptional network that predicts breast cancer survival. *Clin Exp Metastasis* 2008; **25**: 357–69. DOI:10.1007/s10585-008-9146-6

94 Shi Z, Zhang T, Luo L, *et al.* Aquaporins in human breast cancer: Identification and involvement in carcinogenesis of breast cancer. *J Surg Oncol* 2012; **106**: 267–72. DOI:10.1002/jso.22155

95 Jung HJ, Park J-Y, Jeon H-S, *et al.* Aquaporin-5: a marker protein for proliferation and migration of human breast cancer cells. *PLoS One* 2011; **6**: e28492. DOI:10.1371/journal.pone.0028492

96 Su ML, Chang TM, Chiang CH, *et al.* Inhibition of chemokine (C-C Motif) receptor 7 sialylation suppresses CCL19-stimulated proliferation, invasion and anti-anoikis. *PLoS One* 2014; **9**: e98823. DOI:10.1371/journal.pone.0098823

97 Liu L, Xu Q, Cheng L, *et al.* NPY1R is a novel peripheral blood marker predictive of metastasis and prognosis in breast cancer patients. *Oncol Lett* 2015; **9**: 891–6. DOI:10.3892/ol.2014.2721

98 Ehrenfeld P, Manso L, Pavicic MF, *et al.* Bioregulation of kallikrein-related peptidases 6, 10 and 11 by the Kinin b1 receptor in breast cancer cells. *Anticancer Res* 2014; **34**: 6925–38.

99 Yau C, Esserman L, Moore DH, *et al.* A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res* 2010; **12**: R85.

DOI:10.1186/bcr2753

100 Wang J, Scholtens D, Holko M, *et al.* Lipid metabolism genes in contralateral unaffected breast and estrogen receptor status of breast cancer. *Cancer Prev Res (Phila)* 2013; **6**: 321–30. DOI:10.1158/1940-6207.CAPR-12-0304

101 Lee SJ, Chae YS, Kim JG, *et al.* AQP5 Expression Predicts Survival in Patients with Early Breast Cancer. *Ann Surg Oncol* 2013; **21**: 375–83. DOI:10.1245/s10434-013-3317-7

102 Yamada A, Ishikawa T, Ota I, *et al.* High expression of ATP-binding cassette transporter ABCC11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast Cancer Res Treat* 2013; **137**: 773–82. DOI:10.1007/s10549-012-2398-5

103 Park S, Shimizu C, Shimoyama T, *et al.* Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the morphological response to neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Res Treat* 2006; **99**: 9–17. DOI:10.1007/s10549-006-9175-2

104 Hlaváč V, Brynychová V, Václavíková R, *et al.* The expression profile of ATP-binding cassette transporter genes in breast carcinoma. *Pharmacogenomics* 2013; **14**: 515–29. DOI:10.2217/pgs.13.26

105 Sensorn I, Sirachainan E, Chamnanphon M, *et al.* Association of CYP3A4/5, ABCB1 and ABCC2 polymorphisms and clinical outcomes of Thai breast cancer patients treated with tamoxifen. *Pharmgenomics Pers Med* 2013; **6**: 93–8. DOI:10.2147/PGPM.S44006

106 Litviakov N V., Cherdyntseva N V., Tsyganov MM, *et al.* Deletions of multidrug resistance gene loci in breast cancer leads to the down-regulation of its expression and predict tumor response to neoadjuvant chemotherapy. *Oncotarget* 2016; **5**: 7829–41. DOI:10.18632/oncotarget.6953

107 Mobasheri A, Barrett-Jolley R. Aquaporin water channels in the mammary gland: From physiology to pathophysiology and neoplasia. *J Mammary Gland Biol Neoplasia* 2014; **19**: 91–102. DOI:10.1007/s10911-013-9312-6

108 Mao H, Zhang L, Yang Y, *et al.* New insights of CTLA-4 into its biological function in breast cancer. *Curr Cancer Drug Targets* 2010; **10**: 728–36. DOI:10.2174/156800910793605811

109 Loi S. Tumor infiltrating lymphocytes (TILs) indicate trastuzumab benefit in early-stage HER2-positive breast cancer (HER2+ BC). *Cancer Res* 2013; **73:** S1-5. DOI:10.1158/0008-5472.SABCS13-S1-05

110 Jure-Kunkel M, Masters G, Girit E, *et al.* Synergy between chemotherapeutic agents and CTLA-4 blockade in preclinical tumor models. *Cancer Immunol Immunother* 2013; **62**: 1533–45. DOI:10.1007/s00262-013-1451-5

111 Wang Q, Li SH, Wang H, *et al.* Concomitant targeting of tumor cells and induction of T-cell response synergizes to effectively inhibit trastuzumab-resistant breast cancer. *Cancer Res* 2012; **72**: 4417–28. DOI:10.1158/0008-5472.CAN-12-1339-T

112 Demaria S, Kawashima N, Yang AM, *et al.* Immune-mediated inhibition of metastases after treatment with local radiation and CTLA-4 blockade in a mouse model of breast cancer. *Clin Cancer Res* 2005; **11**: 728–34. DOI:11/2/728 [pii]

113 Mittendorf EA, Philips A V, Meric-Bernstam F, *et al.* PD-L1 Expression in Triple-Negative Breast Cancer. *Cancer Immunol Res* 2014; **2**: 361–70. DOI:10.1158/2326-6066.CIR-13-0127

114 Stagg J, Loi S, Divisekera U, *et al.* Anti-ErbB-2 mAb therapy requires type I and II interferons and synergizes with anti-PD-1 or anti-CD137 mAb therapy. *Proc Natl Acad Sci U S A* 2011; **108**: 7142–7. DOI:10.1073/pnas.1016569108

115 Janakiram M, Abadi YM, Sparano JA, *et al.* T cell coinhibition and immunotherapy in human breast cancer. *Discov Med* 2012; **14**: 229–36. DOI:10.1016/j.micinf.2011.07.011

116 Hwang ES, McLennan JL, Moore DH, *et al.* Ductal carcinoma in situ in BRCA mutation carriers. *J Clin Oncol* 2007; **25**: 642–7. DOI:10.1200/JCO.2005.04.0345

117 Hannemann J, Velds A, Halfwerk JBG, *et al.* Classification of ductal carcinoma in situ by gene expression profiling. *Breast Cancer Res* 2006; **8**: R61. DOI:10.1186/bcr1613

118 Done SJ, Eskandarian S, Bull S, *et al.* p53 Missense Mutations in Microdissected High-Grade Ductal Carcinoma In Situ of the Breast. *JNCI J Natl Cancer Inst* 2001; **93**: 700–4. DOI:10.1093/jnci/93.9.700

119 Carraro DM, Elias E V, Andrade VP. Ductal carcinoma in situ of the breast: morphological and molecular features implicated in progression. *Biosci Rep* 2014; **34**: 19–28. DOI:10.1042/BSR20130077

120 Mujtaba SS, Ni Y-B, Tsang JYS, *et al.* Fibrotic focus in breast carcinomas: relationship with prognostic parameters and biomarkers. *Ann Surg Oncol* 2013; **20**: 2842–9. DOI:10.1245/s10434-013-2955-0

121 Hasebe T, Sasaki S, Imoto S, *et al.* Prognostic significance of fibrotic focus in invasive ductal carcinoma of the breast: a prospective observational study. *Mod Pathol* 2002; **15**: 502–16. DOI:10.1038/modpathol.3880555

Table 1. Twelve morphological features graded by the international breast cancer pathology expert committee using clinical categories for 850 TCGA invasive breast cancer cases. Detailed annotation for each TCGA case, consensus annotation, example images of each morphological feature and other details can be found at: www.dx.ai/tcga_breast.

Morphological Feature	Clinical Grading Categories	All cases (n=850), *n* (%)	IDC (n=523), *n* (%)	ILC (n=117), *n* (%)	Special Histological Types (n=117), *n* (%)
Histological Type	Invasive Ductal Carcinoma (IDC)	523 (61.5)			
	Invasive Lobular Carcinoma (ILC)	117 (13.8)			
	Mixed (IDC/ILC)	93 (10.9)			
	Special types (others)	117 (13.8)			
Histological Grade					
Epithelial Tubule Formation	>75% (well differentiated)	91 (11.0)	61 (11.7)	2 (1.8)	23 (22.5)
	10-75% (moderately differentiated)	161 (19.4)	134 (25.6)	5 (4.5)	16 (15.7)
	<10% (poorly differentiated)	576 (69.6)	328 (62.7)	104 (93.7)	63 (61.8)
Nuclear Pleomorphism	Small regular nuclei	67 (8.1)	17 (3.3)	33 (29.7)	11 (10.8)
	Moderate increase in size	372 (44.9)	200 (38.2)	65 (58.6)	57 (55.9)
	Moderate to marked variation in size	389 (47.0)	306 (58.5)	13 (11.7)	34 (33.3)
Mitotic Count (HPF, high powered fields)	0-5 per 10 HPF (low)	383 (46.7)	172 (33.3)	100 (90.1)	54 (53.5)
	6-10 per 10 HPF (medium)	194 (23.6)	146 (28.2)	9 (8.1)	18 (17.8)
	>10 per 10 HPF (high)	244 (29.7)	199 (38.5)	2 (1.8)	29 (28.7)
In Situ Cancer					
DCIS	Present	376 (45.5)	298 (57.0)	5 (4.5)	48 (47.1)
	Absent	450 (54.5)	225 (43.0)	105 (95.5)	54 (52.9)
LCIS	Present	60 (7.3)	8 (1.6)	43 (39.1)	2 (1.9)
	Absent	767 (92.7)	506 (98.4)	67 (60.9)	101 (98.1)
Other Features					
Stromal Inflammation	Present	262 (31.8)	207 (39.9)	11 (9.9)	22 (21.4)
	Absent	562 (68.2)	312 (60.1)	100 (90.1)	81 (78.6)
Necrosis	Present	264 (32.0)	217 (41.6)	3 (2.7)	30 (29.7)
	Absent	562 (68.0)	305 (58.4)	108 (97.3)	71 (70.3)
Proportion of Cancerous Epithelium in Invasive Portion by Area (excluding Areas of Necrosis)	<25% (Low)	78 (9.5)	29 (5.6)	20 (18.2)	13 (12.7)
	25-75% (Moderate)	506 (61.5)	325 (62.5)	76 (69.1)	54 (52.9)
	>75% (High)	239 (29.0)	166 (31.9)	14 (12.7)	35 (34.3)

Apocrine Features	Absent	669 (81.0)	413 (79.0)	98 (88.3)	83 (83)
	1-5% (Minimum)	24 (2.9)	16 (3.1)	2 (1.8)	3 (3.0)
	6-50% (Moderate)	57 (6.9)	36 (6.9)	6 (5.4)	6 (6.0)
	>50% (Marked)	76 (9.2)	58 (11.1)	5 (4.5)	8 (8.0)
Lymphovascular Invasion	Present	204 (24.8)	142 (27.5)	15 (13.6)	24 (23.3)
	Absent	617 (75.2)	375 (72.5)	95 (86.4)	79 (76.7)
Stromal Central Fibrotic Focus	Multiple fibrotic foci	263 (32.1)	184 (35.6)	22 (19.8)	21 (21.2)
	Absent	556 (67.9)	333 (64.4)	89 (80.2)	78 (78.8)

Table 2. The re-stratification of the eleven morphological features into binary grading levels to integrate with molecular data.

Morphological Features	**Binary Categories**
Histological Grade	
Epithelial Tubule Formation	>10% (Well / Moderately differentiated) <10% (Poorly differentiated)
Nuclear Pleomorphism	Small regular nuclei / Moderate increase in size Moderate to marked variation in size
Mitotic Count (HPF, high powered fields)	0-5 per 10 HPF (Low) >6 per 10 HPF (Medium / High)
In Situ Cancer	
DCIS	Present or Absent
LCIS	Present or Absent
Other Features	
Stromal Inflammation	Present or Absent
Necrosis	Present or Absent
Proportion of Cancerous Epithelium in Invasive Portion by Area (excluding Areas of Necrosis)	<75% (Low / Moderate) >75% (High)
Apocrine Features	Absent / 1-5% (Minimum) / 6-50% (Moderate) >50% (Marked)
Lymphovascular Invasion	Present or Absent
Stromal Central Fibrotic Focus	Present or Absent

Table 3. Inter-rater reliability on cases graded by at least two pathologists and Krippendorff's alpha with bootstrap resampling and percentage agreement. Each morphological feature's grading categories are shown in Table 1.

Morphological Feature	Cases (*n*)	Krippendorff's Alpha	Krippendorff's Alpha Bootstrap Resampling (95% confidence interval)	Agreement (%)
Histological Type	358	0.471	0.472 (0.402-0.532)	85.6
Histological Grade				
Epithelial Tubule Formation	316	0.544	0.547 (0.463-0.621)	87.4
Nuclear Pleomorphism	318	0.522	0.520 (0.457-0.590)	80.8
Mitotic Count	311	0.488	0.493 (0.421-0.576)	77.7
In Situ Cancer				
Ductal Carcinoma *In Situ*	317	0.526	0.521 (0.451-0.592)	89.0
Lobular Carcinoma *In Situ*	317	0.298	0.303 (0.088-0.507)	97.5
Other Features				
Stromal Inflammation	315	0.544	0.534 (0.442-0.593)	89.8
Necrosis	317	0.591	0.581 (0.474-0.669)	90.6
Proportion of Cancerous Epithelium in Invasive Portion by Area (excluding Areas of Necrosis)	312	0.472	0.467 (0.387-0.538)	79.2
Apocrine Features	314	0.164	0.189 (0.076-0.318)	90.3
Lymphovascular Invasion	312	0.423	0.413 (0.327-0.515)	90.1
Stromal Central Fibrotic Focus	311	0.256	0.262 (0.155-0.367)	82.7

Table 4. An overview of molecular data significantly associated with morphological features.

Stromal inflammation, Necrosis, Nuclear Pleomorphism and Mitotic Count

The presence of necrosis and inflammation, medium/high mitotic count and marked nuclear pleomorphism were associated with:

- *TP53* loss-of-function mutation, and chr12p13.3, ch8q24.21 (*MYC*) and chr3q26.3 amplifications
- PAM50 Basal-like subtype, higher PAM50 proliferation score
- DNA methylation subtypes 4 and 5, microRNA subtype 4 (these subtypes are linked to Basal-like subtypes [29])
- RPPA (Basal subtype)
- Presence of necrosis, medium/high mitotic count and marked nuclear pleomorphism were enriched for *proliferation* gene sets
- Presence of inflammation was enriched for *inflammation* gene sets

In general, these four features are linked to Basal-like subtypes and have similar molecular basis.

Epithelial Tubule Formation

Poorly differentiated epithelial tubules were associated with:

- *TP53* and *CDH1* loss-of-function mutation
- chr12p13.3, ch8q24.21 (*MYC*) and chr3q26.3 amplifications
- PAM50 Luminal A subtype, higher PAM50 proliferation score
- DNA methylation subtypes 4 and 5, microRNA subtype 4
- Enriched for *inflammation* gene sets

Poorly differentiated epithelial tubules shares selective molecular traits with medium/high mitotic count, marked nuclear pleomorphism and LCIS.

LCIS

The presence of LCIS was associated with:

- *CDH1* loss-of-function mutation
- PAM50 Luminal A subtype, lower PAM50 proliferation score
- DNA methylation subtype 1
- Downregulation of *proliferation* gene sets

The molecular profile of LCIS may be linked to mitochondria dysfunction.

DCIS

Tumours with DCIS were enriched for *proliferation* gene sets.

Up-regulated genes in DCIS are linked to the breast microenvironment, especially myoepithelial cells.

Apocrine Features

The presence of marked apocrine features was associated with:

- chr20q13.2 and chr17q11.2.q12.17q21.1 amplifications
- Enriched gene sets linked to lipid and membrane transport, lipid and/or cell metabolism
- Down-regulated alcohol/drug metabolism gene sets and cytokine signalling

Tumours with marked apocrine features overexpress ATP-binding cassette transporters.

Lymphovascular Invasion

Lymphovascular invasion feature remains mainly morphologic but was also associated with:

- RPPA Basal subtype
- Down-regulation of IL12 and integrin-related neutrophil pathways, and extracellular matrix organization gene sets

Stromal Central Fibrotic Foci

The presence of fibrotic foci remains mainly morphologic.

Fibrotic focus was not associated with any gene but displays down-regulated *inflammation* gene sets

Proportion of Cancerous Epithelium in Invasive Portion by Area (excluding Areas of Necrosis)

High proportion of cancerous epithelium was associated with:

- PAM50 Luminal A, higher PAM50 proliferation score
- RPPA Luminal A/B subtype

Table 5. The variables significantly associated with overall survival in METABRIC ER-positive and ER-negative women in the univariate and subsequent multivariate analyses.

		Univariate Analysis				Multivariate analysis			
	Type of Variable	Hazard Ratio (e^b)	95% CI for Hazard Ratio Lower	Upper	*p*-value	Hazard Ratio (e^b)	95% CI for Hazard Ratio Lower	Upper	*p*-value
ER-Positive Breast Cancer									
Age at Initial Pathological Diagnosis	Clinicopathological	1.048	1.040	1.056	<1.00E-16	1.046	1.037	1.055	<1.00E-16
Tumour Size	Clinicopathological	1.217	1.172	1.264	<1.00E-16	1.148	1.098	1.199	1.03E-09
Node-Positive	Clinicopathological	1.916	1.642	2.235	1.11E-16	1.534	1.257	1.872	2.50E-05
Clinical Grade (METABRIC)	Clinicopathological	1.340	1.182	1.519	4.85E-06	-	-	-	-
Genomic Grade Index	Established Signature	1.639	1.406	1.911	2.64E-10	-	-	-	-
OncotypeDx®	Established Signature	1.619	1.309	2.002	8.86E-06	-	-	-	-
MammaPrint®	Established Signature	1.421	1.191	1.695	9.67E-05	-	-	-	-
Marked Nuclear Pleomorphism	Transcriptomic Signature	1.334	1.231	1.446	2.41E-12	-	-	-	-
High Histological Grade	Transcriptomic Signature	1.311	1.211	1.420	2.08E-11	-	-	-	-
Medium/High Mitotic Count	Transcriptomic Signature	1.339	1.229	1.460	2.93E-11	-	-	-	-
Poorly Differentiated Epithelial Tubules	Transcriptomic Signature	1.907	1.554	2.338	5.83E-10	1.308	1.005	1.703	4.57E-02
Necrosis	Transcriptomic Signature	1.348	1.173	1.548	2.54E-05	-	-	-	-
HER2-Enriched	PAM50	1.423	1.141	1.776	1.75E-03	-	-	-	-
Luminal A	PAM50	0.787	0.670	0.925	3.60E-03	-	-	-	-
Basal-Like	PAM50	1.511	1.138	2.006	4.36E-03	-	-	-	-
Luminal B	PAM50	1.252	1.064	1.473	6.72E-03	-	-	-	-
Normal-Like	PAM50	0.638	0.503	0.811	2.32E-04	-	-	-	-
ER-Negative Breast Cancer									
Node-Positive	Clinicopathological	1.628	1.229	2.156	6.69E-04	-	-	-	-
Tumour Size	Clinicopathological	1.084	1.023	1.149	6.03E-03	-	-	-	-

Confidence interval, CI; Clinicopathological variables were obtained from METABRIC. Transcriptomic signatures for morphological features were developed in this study. Research-based classifications of Genomic Grade Index, OncotypeDx® and MammaPrint® (i.e. Established Signatures) were computed using genefu.

Figure Legends

Figure 1. Heatmap and unsupervised hierarchical clustering of the 38 significant genomic alterations and 11 morphological features based on the degree and direction of the associations. The presence of inflammation and necrosis, marked nuclear pleomorphism and medium/high mitotic counts are clustered together as they share many common genomic alterations.

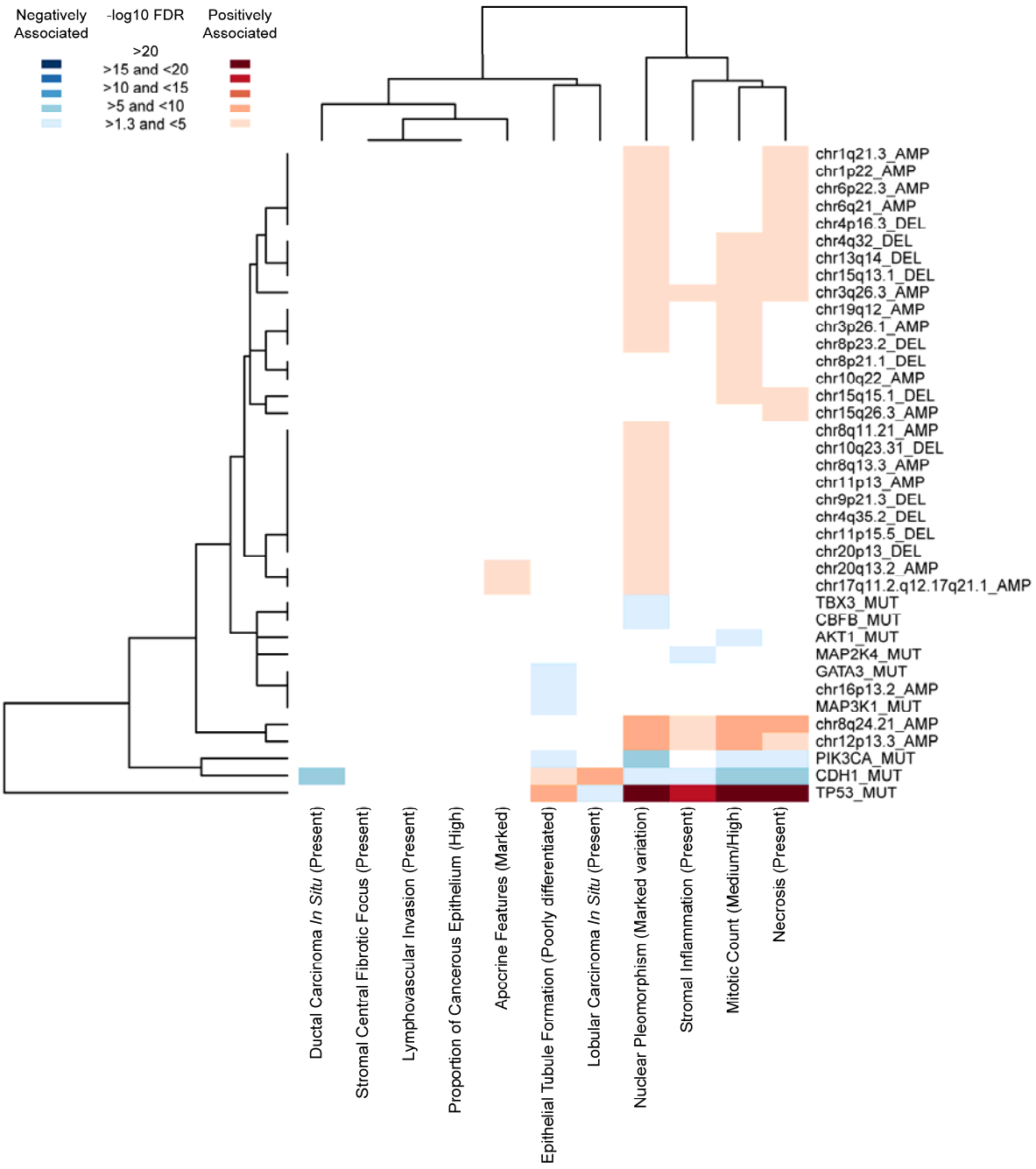


Figure 2. Heatmap summarizing the false discovery rates of 485 significant pathways and unsupervised hierarchical clustering of morphological features. Features are clustered into two groups characterized mainly by *proliferation* and *inflammation*. Detailed pathways are presented in Supplementary Figure S3. The *proliferation* cluster had increased cell proliferation and metabolism, and decreased inflammation and membrane receptor signalling. The *inflammation* cluster comprised largely of immune-related signatures.

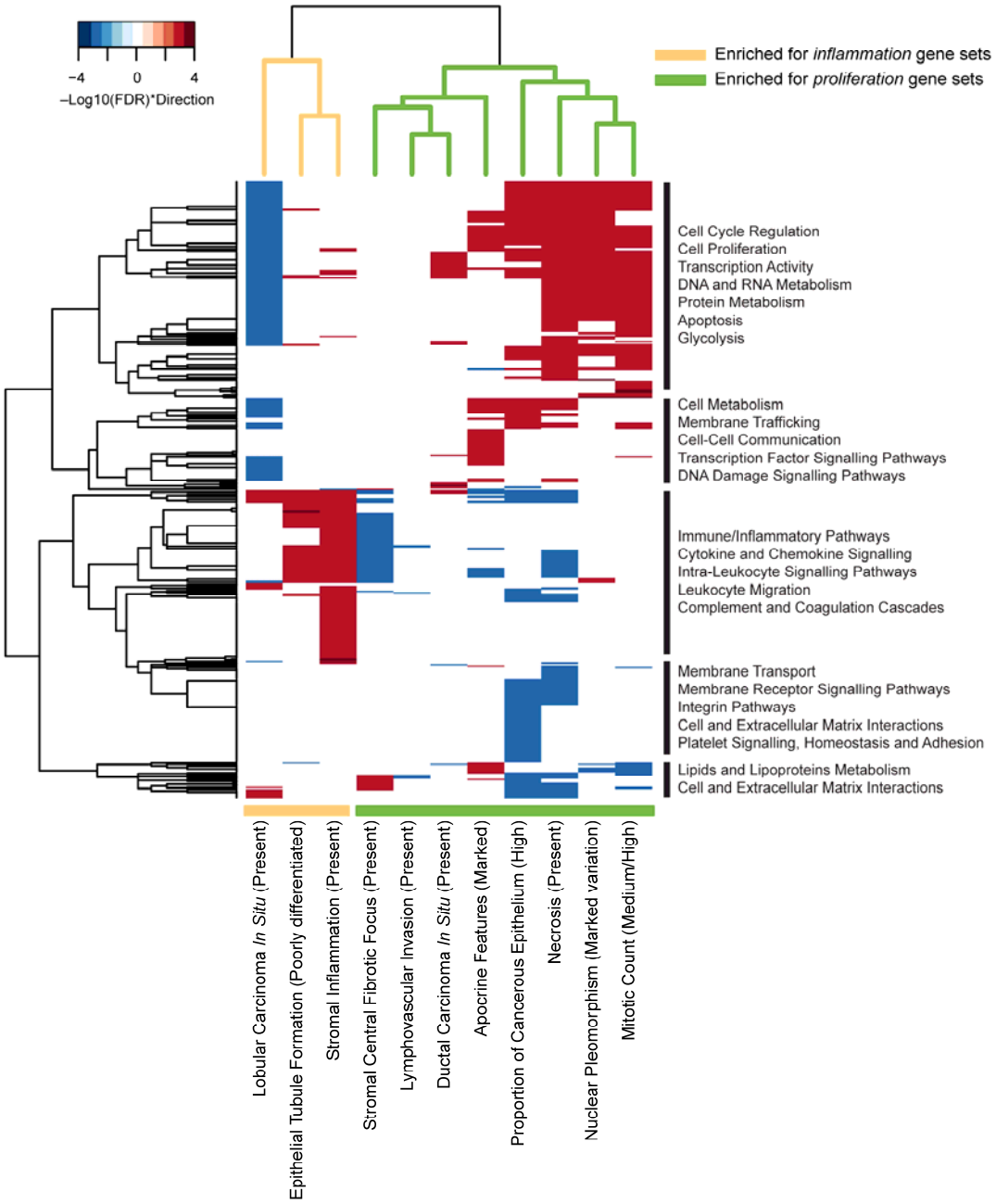


Figure 3. The prognostic significance of the transcriptomic signature for poorly differentiated epithelial tubules in oestrogen-receptor positive women was further validated in a meta-analysis across five cohorts. * The endpoint for GSE25066 is distant-relapse-free survival; the endpoints for all other datasets are overall survival. The summary hazard ratio estimate is a weighted average. Weights are the reciprocal of the estimated variance (square of standard error for the analysis).

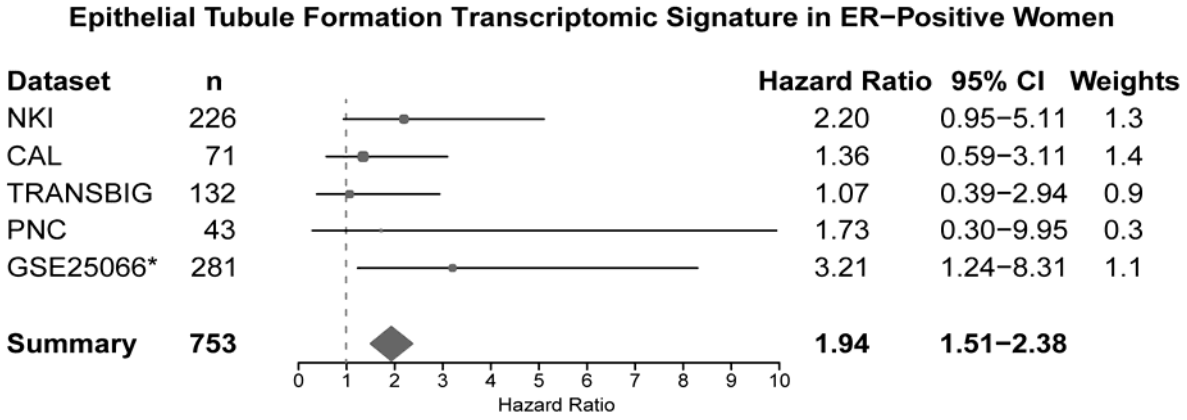
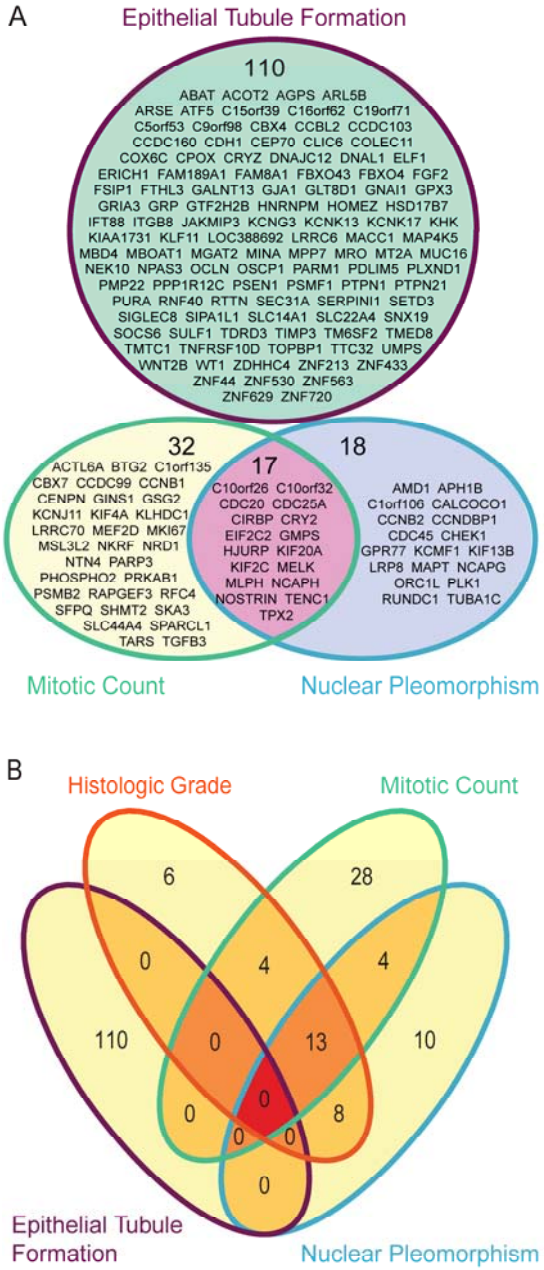


Figure 4. A. There were 17 overlapping genes between the transcriptomic signatures for medium/high mitotic count and marked nuclear pleomorphism while genes predictive of poorly differentiated epithelial tubules were distinct. **B.** Most of the genes in the transcriptomic signature for high histological grade were common to the signatures for medium/high mitotic count and marked nuclear pleomorphism but were distinct from of poorly differentiated epithelial tubules.



SUPPLEMENTARY MATERIAL ONLINE

Supplementary Materials and Methods

Figure S1. Images of the electronic scoring sheets: (A) used by the pathologists to annotate images; (B) annotated scoring sheet with pathological scoring criteria and details.

Figure S2. Boxplots displaying PAM50 proliferation scores for each morphological feature.

Figure S3. Expanded heatmap of Figure 2, with detailed gene sets.

Table S1 A-D. Tables summarising the distribution of morphological features within each PAM50 subtype for (A) all cases; (B) invasive ductal cases; (C) invasive lobular cases; and (D) special histological types.

Table S2. Detailed data for morphological features, PAM50, DNA methylation, miRNA and RPPA subtypes, PAM50 proliferation scores and genomic alterations for 850 TCGA cases.

Table S3. (A) The association of molecular subtypes PAM50, DNA methylation, RPPA, and microRNA with morphological features (Chi-square test with Bonferroni correction). (B) Association of morphological features with each other (Chi-square test with Bonferroni correction).

Table S4. (A) The frequency of each morphological feature in binary groups and the number of differentially expressed genes, stratified by PAM50 subtype, excluding Normal-like. (B) Top 10 up- or down-regulated differentially-expressed genes (limma FDR<0.05) for each morphological feature, ranked by Log2 fold change (Log2FC).

Table S5. (A) Multivariate ROCAUC performances of genomic alterations, gene expression and combined data. (B) Paired *t*-test of ROCAUC of genomic alterations, gene expression and combined data models. (C) Summary of transcriptomic signatures and their enrichments. (D) Transcriptomic signatures of morphological features. (E) Gene set enrichments for transcriptomic signatures of morphological feature.