Identifying survival associated morphological features of triple negative breast cancer using multiple datasets

Chao Wang,[1,2] Thierry Pécot,[3] Debra L Zynger,[4] Raghu Machiraju,[5] Charles L Shapiro,[3,6] Kun Huang[1,7]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001538).

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
[2]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA
[3]Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA
[4]Department of Pathology, The Ohio State University, Columbus, Ohio, USA
[5]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA
[6]Division of Medical Oncology and the Breast Program James Cancer Hospital, The Ohio State University, Columbus, Ohio, USA
[7]Biomedical Informatics Shared Resource, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA

Correspondence to
Dr Kun Huang,
Department of Biomedical Informatics, The Ohio State University, Room 218,
420 w 12th, Columbus,
OH 43210, USA;
kun.huang@osumc.edu

RM, CLS, and KH are co-senior authors with equal contributions.

Received 2 December 2012
Revised 7 March 2013
Accepted 18 March 2013
Published Online First
12 April 2013

Open Access
Scan to access more
free content

To cite: Wang C, Pécot T, Zynger DL, *et al*. *J Am Med Inform Assoc* 2013;**20**:680–687.

ABSTRACT

Background and objective Biomarkers for subtyping triple negative breast cancer (TNBC) are needed given the absence of responsive therapy and relatively poor prediction of survival. Morphology of cancer tissues is widely used in clinical practice for stratifying cancer patients, while genomic data are highly effective to classify cancer patients into subgroups. Thus integration of both morphological and genomic data is a promising approach in discovering new biomarkers for cancer outcome prediction. Here we propose a workflow for analyzing histopathological images and integrate them with genomic data for discovering biomarkers for TNBC.

Materials and methods We developed an image analysis workflow for extracting a large collection of morphological features and deployed the same on histological images from The Cancer Genome Atlas (TCGA) TNBC samples during the discovery phase (n=44). Strong correlations between salient morphological features and gene expression profiles from the same patients were identified. We then evaluated the same morphological features in predicting survival using a local TNBC cohort (n=143). We further tested the predictive power on patient prognosis of correlated gene clusters using two other public gene expression datasets.

Results and conclusion Using TCGA data, we identified 48 pairs of significantly correlated morphological features and gene clusters; four morphological features were able to separate the local cohort with significantly different survival outcomes. Gene clusters correlated with these four morphological features further proved to be effective in predicting patient survival using multiple public gene expression datasets. These results suggest the efficacy of our workflow and demonstrate that integrative analysis holds promise for discovering biomarkers of complex diseases.

INTRODUCTION

Breast cancer is a highly heterogeneous disease.[1 2] During the past half century, several different subtypes of breast cancers have been discovered based on histological features, specific protein markers, and gene signatures obtained from high throughput and high-content experiments. These subtypes present diverse clinical outcomes including varying prognosis and response to treatment.

Histological images of tumor tissues play important roles in breast cancer diagnosis, staging, and prognosis.[3 4] Typically, pathologists visually review stained slides of breast cancer biopsy samples and assign scores to the detected and prevailing tumors. During this inspection, cellular composition is often assessed semi-quantitatively. This process is costly in both time and labor and the results may differ across pathologists. Recently, computer-assisted quantitative analysis of stained histology images has received wide attention in the biomedical and bioimage informatics fields.[5–9] For instance, automated quantification of the levels of salient proteins has led to the discovery of new markers of malignant cells in cancers.[10] In the study by Beck *et al*,[5] stromal morphological features were found to be strongly associated with survival time of breast cancer patients than the morphological features obtained from epithelial compartments.

Besides the clinical use of histological images, subtyping and stratification of breast cancer patients has been widely studied using high throughput gene expression data.[11] For instance, van Veer *et al* identified a 70-gene signature for predicting prognosis of breast cancer patients,[12] while Perou *et al*[13] identified basal-like and non-basal-like subtypes of breast cancers. In addition, in a recent study using The Cancer Genome Atlas (TCGA) breast cancer (BRCA) data, investigators led by Perou identified four major subtypes of breast cancers by combining five different genomic data including gene expression, exome-sequencing, copy number variance, DNA methylation, and microRNA expression.[14]

While both morphological features and genomic data are widely used for breast cancer subtyping and staging, the causal and inferential relationship between genomic data such as gene expression profiles and morphology in histological images from breast cancer patients is still not clear. In a recent study,[6] the heterogeneity of the estrogen receptor (ER) negative breast cancer cells was explained using complementary DNA copy number variance information and consequently an improved prognostic biomarker was suggested. This study suggests that the prognostic power of these two types of features (histological and genomic) should be combined and is likely to lead to better biomarkers for classification of the breast cancer and other types of cancer as well.[15 16] However, there is neither a study which associates salient gene expression biomarkers with pertinent morphology of the tissue, nor a morphology driven result whose scope is also extended to include datasets where gene expression profiles are solely available. In essence, there is a paucity of work that incorporates multiple measurements of disease from disparate sources to create biomarkers.

In this work, we present a novel workflow for integrating histopathological images analysis with gene expression analysis for Triple Negative Breast Cancers (TNBC). Specifically, we develop a workflow for identifying the morphological features that correlate well with survival outcome of patients. Signature morphological features with strong associations to survival were then analyzed by correlating with specific gene expression profiles using another large publicly available dataset and subsequently cross-validated with multiple datasets. Our pipeline provides a novel platform of translating the morphological features to any gene expression data without histopathological images. Our results clearly demonstrate that specific morphological features are correlated with specific genetic features (post-transcription) with enriched biological functions pertaining to cancer development and/or tumor microenvironment structure. These results suggest that new integrative biomarkers can be developed via such integrative approach. An additional aspect of our work is that we are able to work across datasets that were either collected in the public realm or acquired in specific laboratories.

MATERIALS AND METHODS

TCGA breast cancer dataset

The TCGA project[17] collects high-quality breast tumor samples and makes available the clinical information, molecular/genomic profiling data, and histopathology slide images on its data portal. Fifty-one triple negative breast tumor samples were available by selecting from subjects with reported ER (negative), progesterone receptor (PR) (negative), and Her2/neu (0 and 1+) status (using data available up to August 25, 2012). Forty-four of these 51 TNBC subjects have accompanying histopathology images of adequate quality. No other stratification of these TNBC cases was available. Owing to the short median overall follow-up (<2 years) and scarcity of survival events (6 of 51 currently marked as deceased), survival analyses for TCGA breast cancers will not be effective in the next few years.

From the TCGA portal, we downloaded ×40 magnification whole slide images in the SVS file (single-file pyramidal tiled TIFF) format. Tissue slides are available as thin slice of snap-frozen optimal cutting temperature embedded block of tissue for imaging. We used tissue slide images from frozen tissue sections instead of diagnostic slides, since their adjacent tissue samples were used to provide DNA and RNA material for generating genomic data. Typical size of these images is about 100 000×300 000 pixels. It is thus very difficult to process the entire image due to the large size and high computing cost. In addition, whole slide images contain redundant information and encompass artifacts such as folding and missing and broken tissue. Thus, four representative image patches for each whole TNBC slide image were curated by manually selecting heterogeneous and informative regions containing both tumor and stroma tissues. All patches varied between 3000 and 5000 pixels in width and height. This process eliminated artifacts and low-quality regions from consideration.

The Ohio State University TNBC tissue microarray

The Ohio State University (OSU) Pathology Core Facility collects breast cancer biopsy specimens which are stored in the OSU Tissue Archive Service. The necessary clinical information is available from the Information Warehouse at the OSU Medical Center (OSUMC). A total of 365 TNBC patients were identified between the years 1995 and 2005. After pathology review of tumors with sufficient sample for study, 175 paraffin blocks for TNBCs were selected to generate tissue microarrays (TMAs) used in this study. The TMAs were stained using H&E and digitized by an Aperio ScanScope under ×20 magnification. After filtering using measures of satisfactory image quality, TMA images for 143 patients were finally selected. The overall demographic profiles of the cohort were not altered significantly after filtering (shown in table 1).

Table 1 Demographics summary of the Ohio State University triple negative breast cancer cohort

Demographic characteristic	Complete set, 365	Pruned set, 175
Median age (range)	51 (20–84)	51 (20–84)
Race (%; White:African American)	91:8	91:9
Stage (%; I:II)	35:54	31:54
Grade 3 (%)	84	89
Basal cancers (%)	47	45
Adjuvant chemotherapy (%)	73	84
Median follow-up (months)	74 (4–272)	75 (4–272)

Public breast cancer expression datasets

There are several large public breast cancer gene expression datasets with adequate information on survival outcomes and subtyping. Identified biomarkers can be cross-validated using these datasets, of which the Perou[18] (NCBI GSE2741) and NKI[19] collections are among the most frequently used. To evaluate our discovered metagenes on the ER-negative samples, we tested them on these two breast cancer datasets to see if any of the gene signatures can predict ER-negative patient survival.

Integrated breast cancer biomarker identification workflow

In this study, we focus on the discovery of TNBC biomarkers by translating gene–morphology relationships across multiple datasets. Figure 1 shows our overall workflow. Algorithmic details of each are described in subsequent sections.

Using the TCGA dataset, in which both histopathological images and gene expression profiles are available, we computed correlations between morphological features and expression profiles of gene clusters (figure 1A). Then, we selected the morphological features which can classify patients into higher and lower risk groups using the OSU TNBC cohort (figure 1B). Assuming that the selected morphological features glean similar relationships of survival in other datasets, gene clusters with strong correlations to these morphological features can potentially serve as biomarkers for survival. We test them using public breast cancer gene expression datasets without available histology images (figure 1C).

TCGA and OSU histopathology slide image preprocessing and segmentation

During preprocessing we ensured that both tumor epithelial and stromal compartments existed on chosen slides. Magnification for digitization was enforced to be consistent so that all images within each cohort are of the same resolution. Color images were filtered to remove extreme values in the RED channel, which was used to delineate blood cells and spills. A mask was generated to separate the superpixels in each slide.

Characterizing cellular morphological features of TNBC samples

Each tissue sample is heterogeneous with existence of multiple tissues (eg, tumor and stroma) and cells (eg, tumor epithelial cells, fibroblasts, endothelial cells, macrophages). We first adopt an entropy-based image segmentation algorithm similar to that

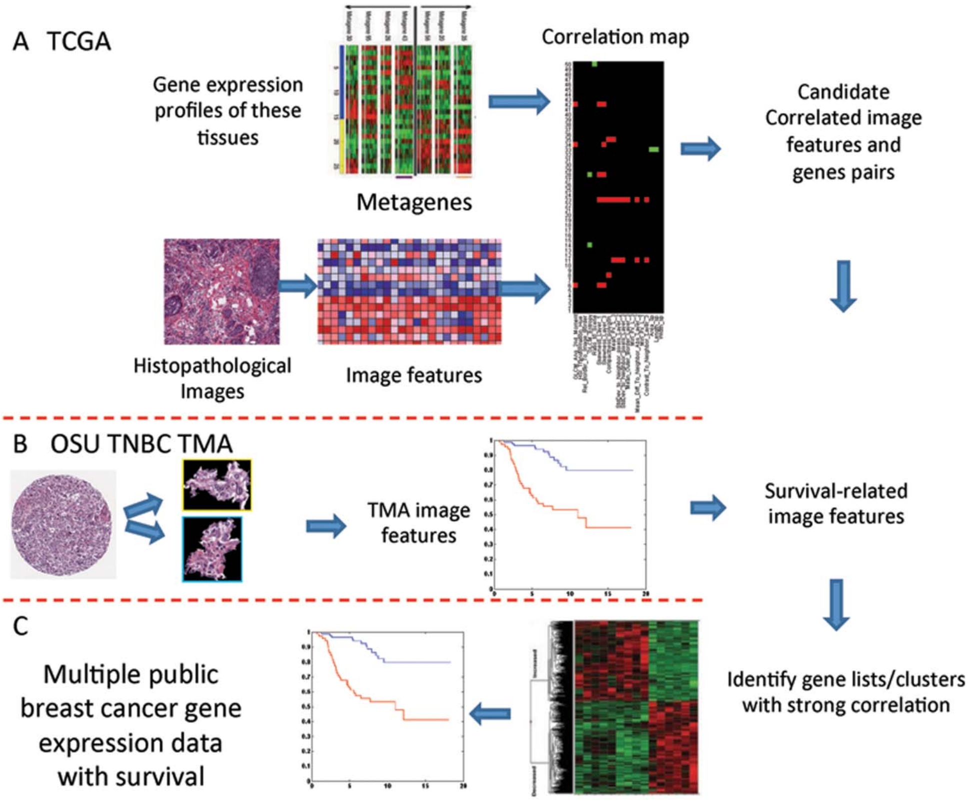
A TCGA
Gene expression profiles of these tissues
Metagenes
Correlation map
Candidate Correlated image features and genes pairs
Histopathological Images
Image features
B OSU TNBC TMA
TMA image features
Survival-related image features
C
Multiple public breast cancer gene expression data with survival
Identify gene lists/clusters with strong correlation

Figure 1 The workflow of cross-datasets feature discovery and validation. (A) Steps for discovering correlations between morphological features and expression profiles of gene clusters using The Cancer Genome Atlas (TCGA) data. (B) Survival-related morphological features are discovered using the Ohio State University (OSU) triple negative breast cancer (TNBC) cohort. (C) Gene clusters strongly correlating with the survival-related morphological features are tested for survival using public breast cancer datasets. TMA, tissue microarray.

in Beck *et al*[5] to divide the images into small regions with relatively homogeneous cellular components and morphology called 'superpixels'.

This procedure removes artifacts in slide preparation.[20] It not only retains the homogeneity of each region of interest (ROI), but also better captures the local morphological structure of the tissue and the relationship between neighboring tissues. The workflow is shown in figure 2. Usually in TMA image analysis, a preprocessing step of color normalization is often employed to mitigate bias; however, this step will also introduce artifacts in the final assessment of the morphological features. Additionally, the subsequent pixel segmentation step replies on the true texture of the images. Finally, since the selected OSU TMA images do not contain large color intensity variations, we did not include the color normalization step in our current pipeline.

Morphometric analysis for cell nuclei

For breast cancer tissue, the abundance of tumor cells is crucial to the diagnosis and prognosis of the patients. Typically, the size of the tumor is a key factor to consider, when the tumor grade is given. Additionally, it is well known that tumor with higher grades often leads to shorter survival, early recurrence, and metastasis.[21 22] The density of tumor cells along with other types of cells, for example lymphocytes and stromal cells, are essential to the quantitative analysis of the histopathological features of the cancer.[23] To evaluate these cellular characteristics, we first identify cell nuclei within each selected patch of the tissue slide images. The pipeline of cell nuclei quantification is illustrated in figure 3. After removing the background, artifacts, and white space regions, a threshold-based segmentation step using the Otsu algorithm[24] is applied to a superpixel to obtain a coarse segmentation of the cell nuclei (figure 3A–C). In histopathological images, nuclei often overlap with each other and form clumps of cells during fixation and staining. The clumps are separated with an edge-cut set selection algorithm (figure 3D).[25] Objects without enough intrinsic nucleus area are considered artifacts and not counted for the analysis.

Morphological feature extraction of tumor and its microenvironment

Besides tumor density being an important characteristic of cancer tissue, morphology of tumor tissue is heavily influenced by stromal and immune cells as well as extracellular matrix (ECM). Here, we measured three classes of morphological features describing the distribution and spatial information of the tumor and its microenvironment. Figure 3 illustrates the extraction of nucleus from other molecular compartments. After we obtained the segmentation of the nuclei, three categories of morphological features are measured on the nuclei: signal intensity, texture, and shape. Shape features include the area of the cell nuclei and eccentricity. These features are tested to be discriminatory and representative for survival of the generic breast cancer population.[5] Examples of morphological features are described in table 2. A full list and summarized descriptions of image features are shown in online supplementary table S1. After features for each individual superpixel in a slide are extracted, the mean estimate of these features is obtained to represent the feature vector for this whole slide.

In addition to the distribution and morphology of the nuclei, stromal components such as fibroblasts, ECM, cellular

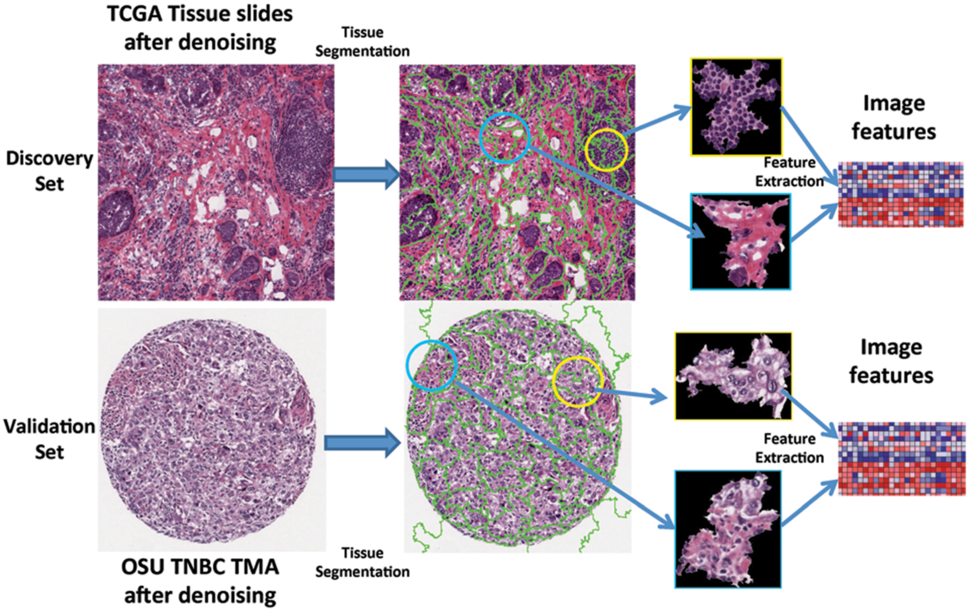

Figure 2 The workflow of the histopathological image analysis. First, after the removal of background and noise, each tissue slide (or tissue microarray, TMA) image was segmented into 'superpixels' delineating the tumor and the stromal compartments of the tissue (green lines mark the boundary of the superpixels in the slide). Then, each superpixel is represented by a series of quantitative morphological features. OSU, Ohio State University; TCGA, The Cancer Genome Atlas; TNBC, triple negative breast cancer.

constituents of the vasculature, inflammatory/immune cells, and adipose tissue arrangement and interaction with the cancer cells are also essential to the tumor development and growth. Especially at the site of the primary tumor in the breast, the interaction between tumor cells and their surrounding milieu is reciprocal; tumor cells influence the stroma and vice versa, ultimately fueling tumor progression.[21] In our study, we also measure the structure of the stromal compartment of the microenvironment in a systematic way. Within each homogeneous region (figure 1), we measured the spatial features, texture, intensity, and morphology of the tissue and build the descriptive features of each of these patches, so that other components in the tissue were measured by these features. The complete list of features is given in online supplementary table S1.

Correlations between gene expression and tissue morphology using TCGA data

mRNA expression profiles for the 41 selected TNBC tumors in TCGA were transformed from RPKM (reads per kilobase per million) normalized Illumina HiSeq 2000 RNA-seq readcounts. The mRNA data were preprocessed as follows: first, we selected genes with top 75% variance. Next, we clustered the mRNA expressions into K gene clusters (metagenes) using an iterative K-means clustering algorithm after 100 iterations. After examining the cluster homogeneity of these metagenes, we determined 50 (K=50) clusters that represented the measured gene expressions the best. Finally, each cluster was represented by its eigengene, which was defined as the first principal component from principal component analysis on the expression profiles of genes in this cluster.[26]

Pearson correlation coefficients (PCCs) between metagene expression and morphological features were calculated as $\mathrm{PCC}(\mathrm{f_i}, \mathrm{m_j}) = \mathrm{cov}(\mathrm{f_i}, \mathrm{m_j})/\sigma_{\mathrm{f_i}}\sigma_{\mathrm{m_j}}$, where f_i is a feature vector for all samples and m_j is the eigengene expression of metagene j for the same samples. All pairs of correlations between features and eigengenes were obtained and the correlation matrix was then formed. Enrichment analysis on the selected metagenes was carried out using TOPPGene (http://toppgene.cchmc.org/).

Survival analysis of OSU and public datasets

Survival analysis was performed with OSU TNBC datasets, with median survival time of 75 months. Survival was calculated

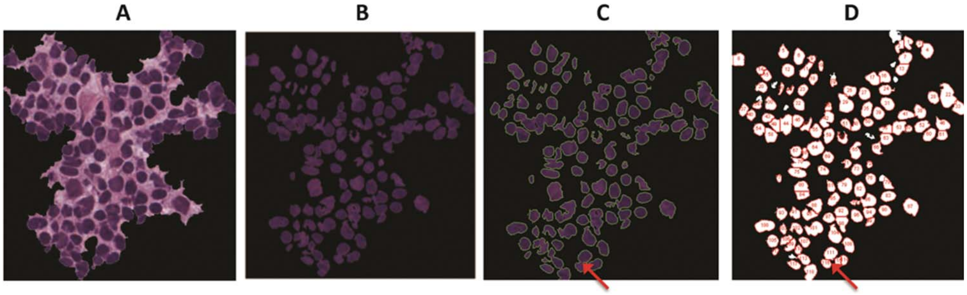

Figure 3 Nuclei segmentation within each region of interest. (A) An example of the original superpixel. (B) The result after Ostu cellular segmentation. Some of the nuclei overlap and form large clumps. (C) Boundary of the cell nuclei. (D) Final segmentation of the cell nuclei after using edge cut (example indicated by red arrow).

Table 2 Examples of morphological features of cancer tissue images.

Feature name	Description
GLCM_Ang_2nd_Moment	Haralick Texture. GLCM angular second moment (akin to variance)
Rel_Border_To_Image_Border	Rel. Border To determines the relative border length an object shares with the image border
GLCM_Entropy	The value for entropy is high, if elements of GLCM are distributed equally
Rel_Area_Cell_Nuclei	Mean value of areas within cell nuclei
Density_Cell_Nuclei_Stddev	SD of the densities of cell nuclei

GLCM, gray-level co-occurrence matrix.

from the time of initial diagnosis of breast cancer to the time of death. Patients were divided into two groups as determined by feature values being greater or lower than the median value. Univariate Cox proportional hazards regression models were fitted to estimate the hazards of death among patients using each morphological feature. p Values were calculated based on univariate regression models to determine the significance of each covariate of interest, where $p<0.05$ was considered significant. Kaplan–Meier estimators were computed to plot the survival curves for covariates which were deemed to be significant. For survival analysis of metagenes on public datasets, a prognostic index of each patient was calculated by the sum of gene expression weighted by the hazard coefficients that were estimated by Cox proportional hazards model. After 10 times of 10-fold cross-validation, the patients were divided by the 50 percentile of the tested prognostic index with statistical significance evaluated by log-rank test.

RESULTS AND DISCUSSION

Translational discovery of survival-related morphological features by cancer-related genes

In this study, we applied our proposed image analysis workflow on the more than 400 slide H&E images and extracted 37 previously-tested morphological features. We investigated the correlations between these features and the transcriptional expression profiles and found 23 significant positive statistically associations and 25 negative ones. Analysis on the 143 OSU TNBC TMA images as the validation set unveils four morphological features that have strong correlations with survival. The corresponding gene clusters of these features were validated using two other independent datasets.

Metagenes with strong correlations to morphological features

PCCs between expression of the 50 metagenes and morphological features were calculated and are shown in figure 4A. Forty-eight strong correlations (|PCC|>0.5) are highlighted in figure 4B, of which 23 are positive correlations and 25 are negative (see online supplementary table S3). Examples of morphology-correlated metagenes are listed in table 3, along with their major molecular functions and regulated human phenotypes obtained from enrichment analysis. Some of the metagenes strongly correlate with multiple morphological features. For instance, MetaGene_2 includes genes (eg, MYOT, ACTA1) regulating molecular structural constituent of muscle motor activity and it controls the abnormality of protein fibers, which is the major component of the tumor microenvironment and is associated with fibroblast cells. It is noteworthy that MetaGene_2 negatively correlates with most morphological features.

Survival of TNBC based on morphological biomarkers

We conducted univariate survival analysis of the morphological features measuring variability of the TNBC tissue slides. Cox proportional hazard models were fitted based on patient's survival time and morphological features. Survival tests of the top four predictive image features are shown in figure 5. Feature 'Area_Cell_Nuclei_stddev' measures the SD of the size of the nuclei. Higher values of this feature imply larger variations in nuclei sizes in the poorer prognosis group. This statistic was obtained by analyzing more than 30 000 cell nuclei of TNBCs in the validation cohort. Another marked characteristic of the poor survival group is the pixel density gradient among neighborhood of pixels. A high value indicates more dramatic deviations from the normal uniformly distributed tissue texture. The larger discontinuity, as noted in the survival curves among patients with poor prognosis, may result from a larger proportion of tumor cells. Since we kept the tumor size bias as small as possible when the ROIs were selected and size measurements were normalized by the size of the ROIs, the measurement bias is minimized.

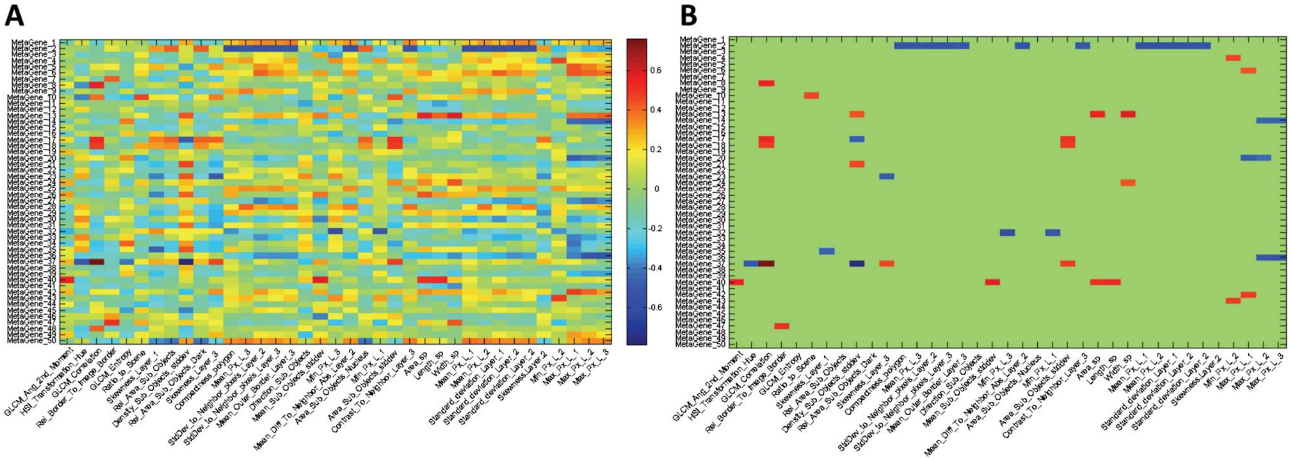

Figure 4 Pairwise correlation heat map between metagene expression and morphology of tissue in The Cancer Genome Atlas discover set. (A) Continuous correlation without threshold. The blue color demonstrates negative correlation; the red color demonstrates the positive correlation. (B) Thresholded correlation (|PCC|>0.5).

Table 3 Examples of enriched gene ontology and human phenotype terms of the metagenes strongly associated with morphology in figure 4

MetaGenes_ID	Top molecular functions	Top human phenotype
MetaGene_2	Structural constituent of muscle Motor activity Cytoskeletal protein binding	Myopathy Abnormality of muscle fibers Muscle fiber cytoplasmic inclusion bodies
MetaGene_ 13	Structural molecule activity Structural constituent of cytoskeleton	Abnormal epidermal layer morphology
MetaGene_37	3′,5′-Cyclic-AMP phosphodiesterase activity	Smooth muscle contraction Regulation of smooth muscle contraction
MetaGene_40	Zinc-finger transcription factor	

Survival analysis of the identified biomarkers on multiple public datasets

Most of the public breast cancer gene expression datasets do not possess histological images. In order to test the above morphological markers, we tested on these datasets using metagenes which are highly correlated with the above four predictive morphological features. In the TCGA data, three features (‘Area_Cell_Nuclei_stddev’, ‘StdDev_to_Neighbor_pixels, and ‘GLCM_Correlation’) out of the above four features have top correlations with one metagene (listed in online supplementary table S4). In the Perou and NKI datasets, this metagene can separate the ER-negative patients into two cohorts with different outcomes. The Kaplan–Meier curves for this metagene are shown in figure 6.

An interesting finding on these two validation datasets is that for the high risk cohort in the Perou dataset, the survival drops dramatically after about 2 years (figure 6B). These gene biomarkers are prognostic in both studies, with p values 0.027 and 0.008, respectively.

DISCUSSION

In this work, we present a workflow for correlating histological imaging features with gene expression profiles. Since TNBC is a subtype of breast cancer with poor prognosis and without clear predictive biomarkers, this study is part of our larger effort trying to establish more effective prognostics and predictive biomarkers for subtyping TNBC by integrating morphological features with molecular/genomic profiles. By establishing the relationship between morphological features in histological images with gene expression profiles, not only can we derive novel insights on the molecular basis for different cell and tissue morphologies, but we also practically suggest that specific survival related molecular signatures can be manifest as morphological artifacts and features and thus avoid the cumbersome process of collecting gene expression profiles from patients. Moreover, among all the 37 previously-tested morphological

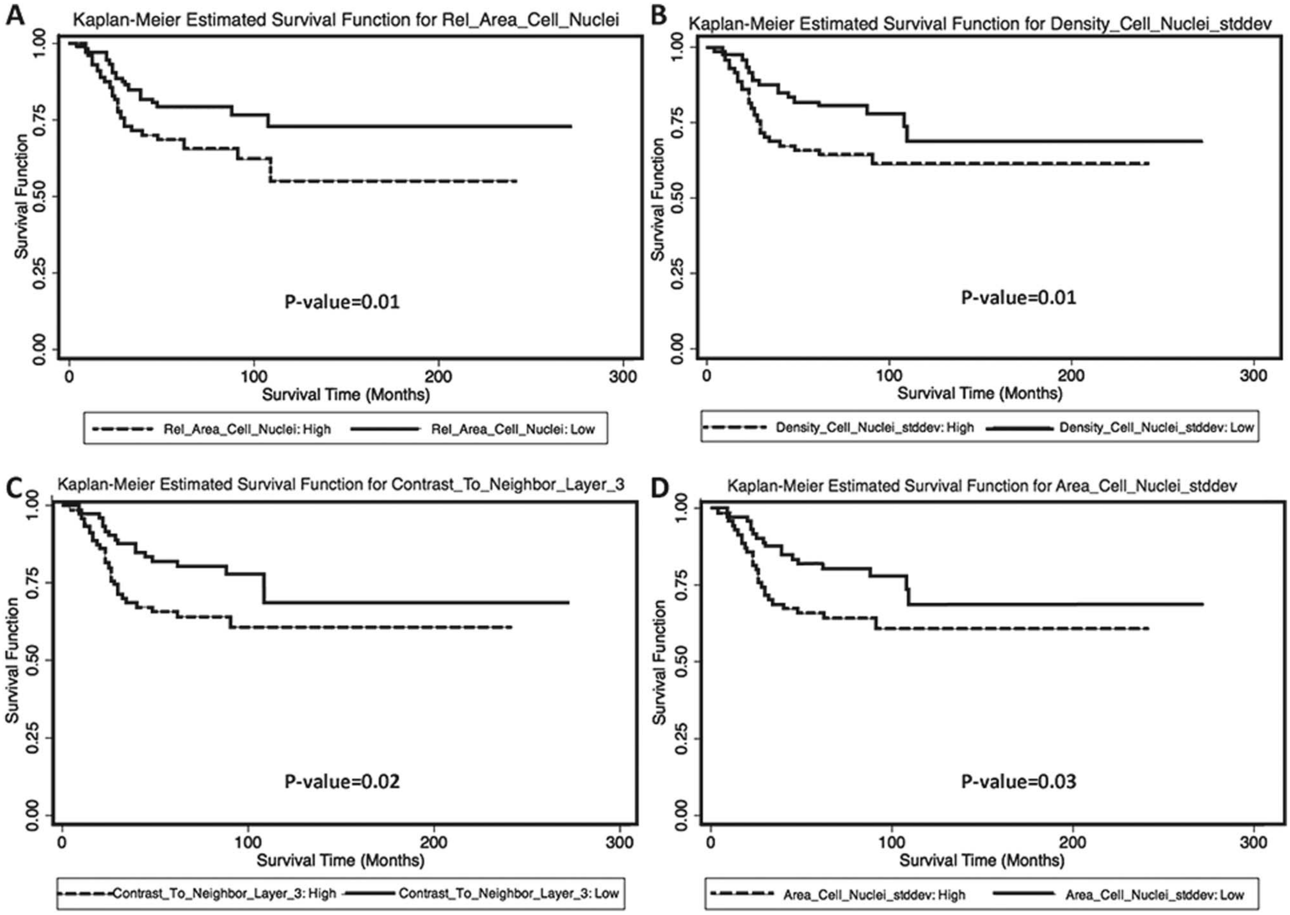

Figure 5 Kaplan–Meier survival curves of prognostic model in Ohio State University triple negative breast cancer tissue microarray. Higher values of the image features are plotted as blue lines, and lower values are plotted as red lines. (A) Survival on the two groups with distinct values of ‘Rel_Area_Cell_Nuclei’. (B) Survival of feature ‘Density_Cell_Nuclei_stddev’. (C) Survival of feature ‘Contrast_To_Neighbor_Layer_3’. (D) Survival of feature ‘Area_Cell_Nuclei_stddev’.

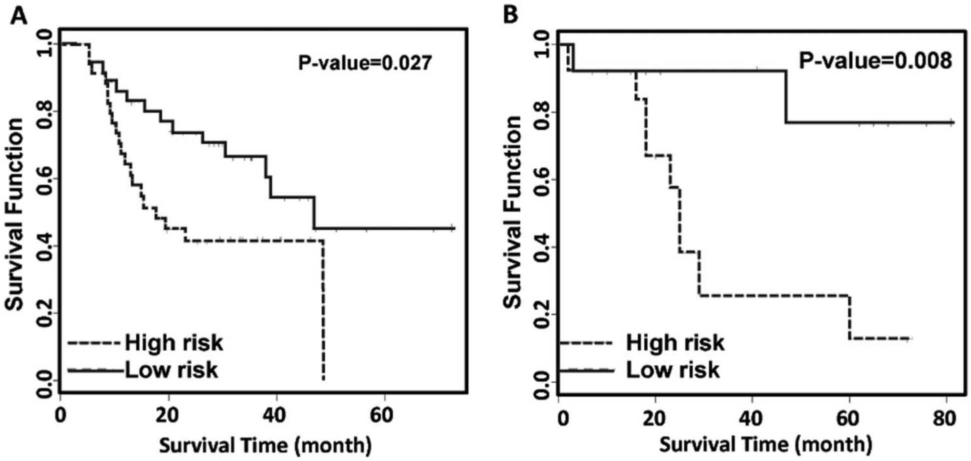

Figure 6 Kaplan–Meier curves for the metagene that shows the highest correlation with the image features. All time is represented in months. (A) Survival on the NKI ER-negative patient subset. A higher risk group was revealed by this metagene. (B) Survival on the Perou dataset, ER-negative patient subset. This list of genes can separate the patients into groups with very significantly different outcomes.

features, the top prognostic ones are related to cell nuclei. Therefore, our study validated in a novel manner that nuclear features, which have been used in tumor grading in clinical practice, are prognostic in TNBC.

To achieve this goal, we took further advantage of TCGA data, which proved to be an invaluable resource for such integrative genomic research. Even though the TCGA breast cancer data is relatively new and the follow-up time for patients is not long enough for effective survival analysis on TNBC patients, the matched histological images with gene expression profiles nevertheless provides the bridge for these two data modalities allowing us to also use multiple modalities of data from different sources (eg, OSU cohort as well as NCBI GEO).

Specifically in this study we identified metagenes that are highly correlated with the four morphological features that can predict TNBC patient survival using the OSU cohort. We further determined that the features that best separated the better and worse survival groups were the cellularity of epithelial cells and the shape of the cancer cells. Essentially, the area of cancer cell nuclei and the diversity of the area of nuclei show great prognostic power for survival. To demonstrate a distinction between low-value features and high-value features, the box plot of the area-based features and the patches with extreme values are shown in online supplementary figure S1.

While we currently do not have another independent dataset to validate these morphological features, we were able to test if the expression profiles of the associated metagenes have similar predictive power for multiple large datasets. We found that two of five gene clusters show predictive power in at least one of the test data. In particular, MetaGene_2 has strong predictive power in both public datasets. This metagene is enriched with cytoskeleton and fiber genes, which is not only consistent with its association with cell morphology, but also implies its potential roles in the development of tumor microenvironment including the stroma. The other metagene (MetaGene_13) is enriched with epidermal layer development indicating its association with tumor epithelial cells, which may explain its role in cancer development and relationship with tumor cell morphology. These observations strongly suggest that our approach effectively identified gene clusters that can partially explain the morphological characteristics and can be used as predictive markers. Additionally, our methods to process whole tissue slides improve the current state-of-the-art pathological TMA image processing.

This study has several limitations: first, we only utilized the tissue slide images of TCGA to measure the patients' phenotypes. The tissue slides were adjacent to tissues from which genomic data were derived. Thus, correlations between the morphology of these areas and genomic data are better reflected. However, these slide images are obtained from frozen sections and may possess larger artifacts than the diagnostic images. Another limitation of this study is that we did not differentiate tumor epithelial, adipose, and stromal tissue in our measurement of the morphology. Classification of these cell types is an ongoing study and will appear in a future publication. Recently Yuan *et al*[8] showed that immune cells can be used as an effective biomarker for prognosis, suggesting tissue specific morphological features should be explored. We plan to carry out a systematic analysis on different compartments of the tumor region.

CONCLUSION

We present a novel workflow for discovering the associations between histological features with gene expression profiles. Our analysis reveals 48 pairs of strongly correlated morphological features and gene clusters. Four of the morphological features were identified as potential biomarkers separating TNBC patients into groups with different survival in a large validation cohort. Gene ontology analysis suggests that the high correlations are consistent with development and tumor related functions. Additionally, these morphological features on the tumor tissues can be extended as prognostic biomarkers for ER-negative breast cancers as the top gene cluster correlated with these morphological features was shown to be effective for predicting patient survival for ER-negative breast cancers in two independent public datasets.

Correction notice This paper has been corrected since it was published Online First. In table 1, the number for 'Race' and 'Pruned set' has been changed from 81:9 to 91:9. On page 5, two sentences have been added to the end of the first paragraph.

Contributors KH, RM, and CLS conceived the data integration idea and led the study. KH, RM, and CW designed the computational workflow. CW implemented the computational workflow and carried out the data analysis with KH and RM. CLS provided the OSU TNBC images. DZ reviewed the histology images. CLS and DZ provided the clinical interpretations. TP conducted the survival analysis for public gene expression datasets. CW, KH, and RM wrote the manuscript. CLS and DZ revised the manuscript. All authors reviewed and approved the manuscript.

Funding This work is partially supported by OSU-Howard Hughes Medical Institution MED into Grad Scholar Program (56006765), Stefanie Spielman Fund for

Breast Cancer Research, NCI R01CA141090, and UKY-OSU CCTS Joint Pilot Award (UL1RR033173-01, 8UL1TR000090-05).

Competing interests None.

Ethics approval The Ohio State University IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Irshad S, Ellis P, Tutt A. Molecular heterogeneity of triple-negative breast cancer and its clinical implications. *Curr Opin Oncol* 2011;23:566–77.

2 Perou CM. Molecular stratification of triple-negative breast cancers. *Oncologist* 2011;16:61–70.

3 Nafe R, Franz K, Schlote W, *et al*. Morphology of tumor cell nuclei is significantly related with survival time of patients with glioblastomas. *Clin Cancer Res* 2005;11:2141–8.

4 Khan OA, Fitzgerald JJ, Field ML, *et al*. Histological determinants of survival in completely resected T1-2N1M0 nonsmall cell cancer of the lung. *Ann Thorac Surg* 2004;77:1173–8.

5 Beck AH, Sangoi AR, Leung S, *et al*. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* 2011;3:108–13.

6 Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph* 2011;35:515–30.

7 Pham NA, Morrison A, Schwock J, *et al*. Quantitative image analysis of immunohistochemical stains using a CMYK color model. *Diagn Pathol* 2007;2:8.

8 Yuan Y, Failmezger H, Rueda OM, *et al*. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* 2012;4:157ra143.

9 Basavanhally A, Feldman M, Shih N, *et al*. Multi-field-of-view strategy for image-based outcome prediction of multi-parametric estrogen receptor-positive breast cancer histopathology: comparison to Oncotype DX. *J Pathol Inform* 2011;2:S1.

10 Gil J, Wu HS. Applications of image analysis to anatomic pathology: realities and promises. *Cancer Invest* 2003;21:950–9.

11 Sotiriou C, Neo SY, McShane LM, *et al*. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003;100:10393–8.

12 Mook S, Schmidt MK, Viale G, *et al*. The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1–3 positive lymph nodes in an independent validation study. *Breast Cancer Res Treat* 2009;116:295–302.

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

14 Koboldt DC, Fulton RS, McLellan MD, *et al*. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.

15 Cooper LA, Kong J, Gutman DA, *et al*. Integrated morphologic analysis for the identification and characterization of disease subtypes. *J Am Med Inform Assoc* 2012;19:317–23.

16 Chang H, Fontenay GV, Han J, *et al*. Morphometic analysis of TCGA glioblastoma multiforme. *BMC Bioinformatics* 2011;12:484.

17 The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2008;490:61–70.

18 Weigelt B, Hu ZY, He XP, *et al*. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* 2005;65:9155–8.

19 van't Veer LJ, Dai HY, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.

20 Liu MY, Tuzel O, Ramalingam S, *et al*. Entropy rate superpixel segmentation. Proc CVPR IEEE Colorado Springs, USA; 2011.

21 Frost AR, Hurst DR, Shevde LA, *et al*. The influence of the cancer microenvironment on the process of metastasis. *Int J Breast Cancer* 2012;2012: Article ID 756257.

22 Schnitt SJ. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod Pathol* 2010;(23 Suppl 2):S60–4.

23 Baxevanis CN, Dedoussis GV, Papadopoulos NG, *et al*. Tumor specific cytolysis by tumor infiltrating lymphocytes in breast cancer. *Cancer* 1994;74:1275–82.

24 Cheriet M, Said JN, Suen CY. A recursive thresholding technique for image segmentation. *IEEE Trans Image Process* 1998;7:918–21.

25 Wen Q, Chang H, Parvin B. A delaunay triangulation approach for segmenting clumps of nuclei. 2009 IEEE International Symposium on Biomedical Imaging Boston, Massachusetts, USA: From Nano to Macro, Vols 1 and 2*;* 2009: 9–12.

26 Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 2000;97:10101–6.