CANCER

Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling

Yinyin Yuan,[1,2*†] Henrik Failmezger,[3,4‡] Oscar M. Rueda,[1,2‡] H. Raza Ali,[1,2‡] Stefan Gräf,[1,2§] Suet-Feung Chin,[1,2] Roland F. Schwarz,[1,2] Christina Curtis,[5] Mark J. Dunning,[1] Helen Bardwell,[1] Nicola Johnson,[6] Sarah Doyle,[6] Gulisa Turashvili,[7,8] Elena Provenzano,[6,9] Sam Aparicio,[7,8] Carlos Caldas,[1,2,9,10] Florian Markowetz[1,2*]

(Published 24 October 2012; revised 21 November 2012)

Solid tumors are heterogeneous tissues composed of a mixture of cancer and normal cells, which complicates the interpretation of their molecular profiles. Furthermore, tissue architecture is generally not reflected in molecular assays, rendering this rich information underused. To address these challenges, we developed a computational approach based on standard hematoxylin and eosin–stained tissue sections and demonstrated its power in a discovery and validation cohort of 323 and 241 breast tumors, respectively. To deconvolute cellular heterogeneity and detect subtle genomic aberrations, we introduced an algorithm based on tumor cellularity to increase the comparability of copy number profiles between samples. We next devised a predictor for survival in estrogen receptor–negative breast cancer that integrated both image-based and gene expression analyses and significantly outperformed classifiers that use single data types, such as microarray expression signatures. Image processing also allowed us to describe and validate an independent prognostic factor based on quantitative analysis of spatial patterns between stromal cells, which are not detectable by molecular assays. Our quantitative, image-based method could benefit any large-scale cancer study by refining and complementing molecular assays of tumor samples.

INTRODUCTION

A major obstacle in refining molecular cancer signatures is the cellular heterogeneity and complex tissue architecture of most tumor samples, which consist of cancer cells as well as immune and other stromal cells. These cells form an integral part of the tumor microenvironment, but their admixture poses severe challenges for molecular assays, particularly in large-scale analyses of hundreds or thousands of patient samples, such as The Cancer Genome Atlas (*1*) or the International Cancer Genome Consortium (*2*). One major challenge is normal cell contamination, which can dilute cancer cell information and compromise detection sensitivity of somatic copy number events (*3*). Microdissection can help to selectively extract cancer cells (*4*) but is labor-intensive and cannot be easily scaled to many samples. A second challenge is that spatial features of tissue architecture are lost in molecular assays, rendering it impossible to assess co-location and other cellular interactions within the tumor microenvironment (*5*).

General cellular heterogeneity in a sample can be assessed by a pathologist visually examining stained slides of the tumor. Like microdissection, a detailed histopathological analysis of cellular composition is time-consuming and often infeasible for large-scale studies. Pathological scores reflect primarily the density of tumor cells seen in a given area (*6*) (cell-to-area ratio), and less so the cancer cell proportion (cell-to-cell ratio), which could be used to estimate the cancer DNA content. Thus, pathological scores are poorly suited to deconvolute signal to noise in molecular assays. Cellularity assessments by pathologists also generally yield qualitative results on coarse ordinal scales and rarely provide quantitative data. Conversely, computational and statistical inference from molecular data can be applied to large data collections and yield quantitative results. For example, lymphocytic infiltration (LI) can be inferred from gene expression data (*7*), as can cellularity from single-nucleotide polymorphism (SNP) data (*8*, *9*). However, compared to a pathologist's assessment, these approaches are indirect and often strongly rely on statistical assumptions (*10*).

To draw on both the richness of histopathological information and the speed and quantitative results of computational analyses, we have developed a systematic approach that exploits the most widely used method in histological diagnosis, images of hematoxylin and eosin (H&E)–stained solid tumor sections. In two large cohorts of breast cancer patients, we show how an image-based approach can increase the power of molecular assays and complement them to uncover prognostic features not visible in molecular data. Our integration of histopathology and genomics extends recent approaches that only identified morphological features predictive of patient survival by image analysis (*11*). Our approach could improve clinical assessment of breast cancer by removing the variability in subjective histological scoring between pathologists, using a combination of molecular signatures and features

[1]Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, UK. [2]Department of Oncology, University of Cambridge, Cambridge CB2 2XZ, UK. [3]Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. [4]Gene Center and Department of Biochemistry, Center for Integrated Protein Science Munich, Ludwigs-Maximilian University, Munich 81377, Germany. [5]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [6]Department of Histopathology, Cambridge University Hospital NHS Foundation Trust (Addenbrooke's Hospital), Cambridge CB2 0QQ, UK. [7]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [8]Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada. [9]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [10]Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.
*To whom correspondence should be addressed. E-mail: florian.markowetz@cancer.org.uk (F.M.); yinyin.yuan@icr.ac.uk (Y.Y.)
†Present address: The Institute of Cancer Research, London SW3 6JB, UK.
‡These authors contributed equally to this work.
§Present address: Division of Respiratory Medicine, Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.

of the tumor tissue architecture for an objective stratification of breast cancer.

RESULTS

Automated image analysis dissects cellular heterogeneity of human tissue sections

We collected matched H&E-stained images, gene expression data, and copy number variation data for a discovery set of 323 breast cancer patients and for an independent validation set of 241 breast cancer patients as part of the METABRIC collection (*12*). Each of the H&E images contained histopathological sections from the top, middle, and bottom of the tumor adjacent to the parts used for DNA and RNA profiling (Fig. 1). Our image processing approach automatically segments H&E images, detects artifacts arising from differences in section thickness and staining variation, and classifies cellular components into three categories: cancer, lymphocyte, or stromal. This is based on morphological features (table S1) using a support vector machine (SVM) classifier (*13*) that has been trained by a pathologist (Fig. 2A). The cellular categories were broadly defined by nuclear morphology in the H&E images (Fig. 2B). Malignant cells typically have large (>10 μm),

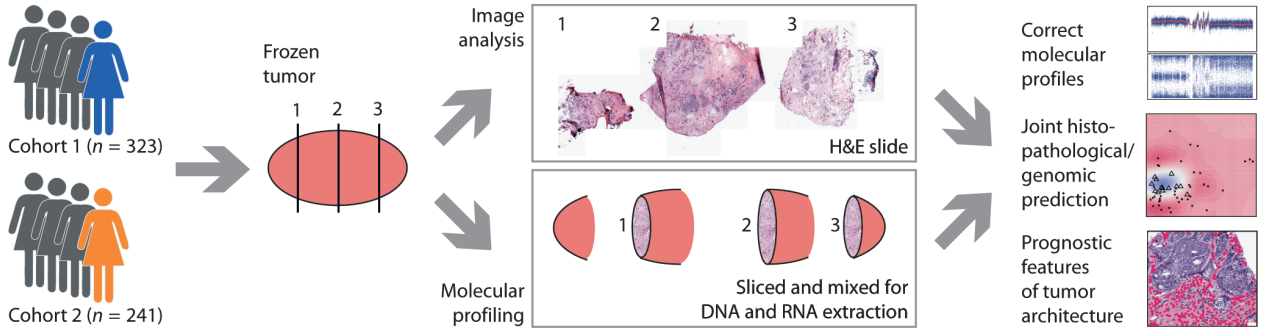

Fig. 1. Parallel image analysis and molecular profiling of breast tumors for automated assessment of tumor composition. Data were obtained from frozen tumor samples from two independent patient cohorts. To best recapitulate tumor architecture, we took 5-μm sections for imaging from the top, middle, and bottom portions of the tumor, sandwiching sections that were used for DNA and RNA profiling.

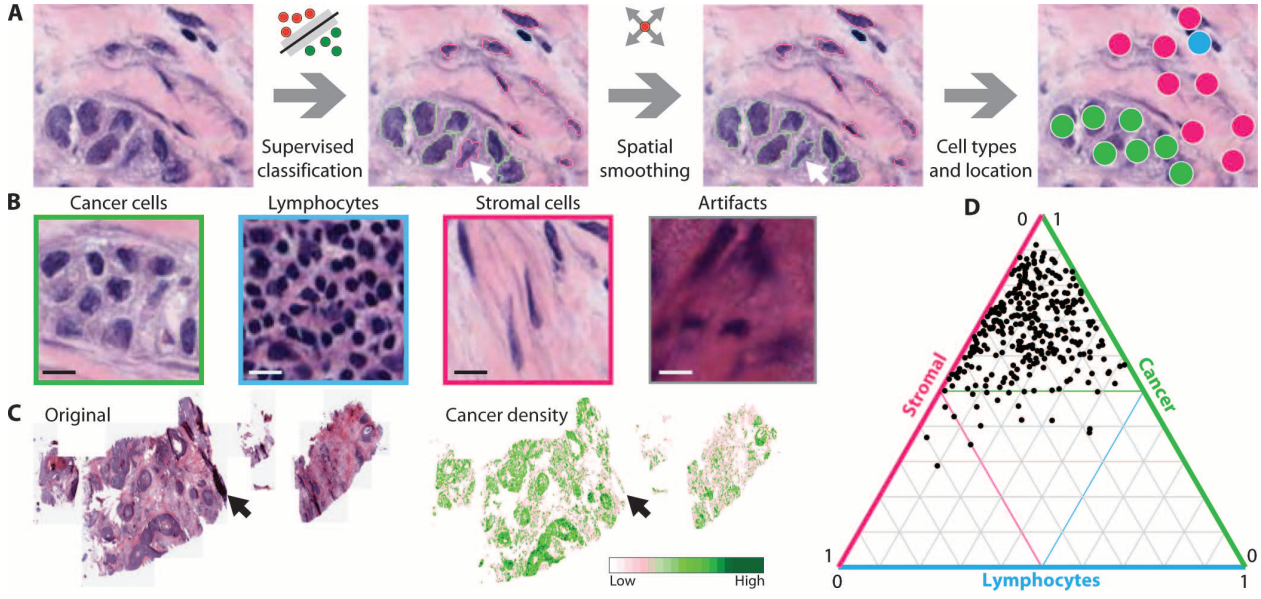

Fig. 2. The proposed image processing pipeline for uncovering cellular heterogeneity. (**A**) On the basis of their nuclear morphologies in H&E-stained images, individual cells were first classified into cancer cells (green), stromal cells (red), or lymphocytes (blue) with supervised classification techniques. In a second step, spatial smoothing refined the result by incorporating labels of neighboring cells; for example, the white arrow points to a cancer cell that was misclassified in the first step but was then corrected in the second step. Last, the system outputs the numbers and locations of cell types in each tumor. (**B**) Example images of the four different classes used in the classifier: cancer cells, stromal cells, lymphocytes, and artifacts like blurred or folded regions. Scale bars, 10 μm. (**C**) The density of cancer cells in each H&E-stained tumor slide can be visualized as a heat map. The arrow indicates an artifact of the cutting process that was automatically discarded from further analysis. (**D**) A global overview of the distribution of cell types in 323 breast tumors (discovery set) in a triangle plot, where each dot represents a tumor and the three axes represent the proportions of cancer cells, lymphocytes, and stromal cells. Thin colored lines represent the 50% mark for each cell type.

round nuclei. The stromal class was trained on spindle-shaped stromal cell nuclei (likely to be fibroblasts) and may encompass other stromal cells with similar morphology, such as endothelial cells. The lymphocyte class was trained on immune cells with the distinctive morphology of lymphocytes: small (<8 μm), dark nuclei and not much cytoplasm.

In a second step, we incorporated a spatial kernel smoothing technique (*14*) and a hierarchical multiresolution model to aggregate individual labels into a final prediction. Although spatial smoothing corrects for local, sporadic errors, the hierarchical model facilitates large-scale relabeling based on global features of the tumor by describing clusters of cells instead of individual cells using a multiresolution predictor. This hierarchical information flow enables pattern recognition in a manner close to human perception because pathologists identify cancer cells based not only on cells' appearances but also on their spatial organization. As a result, cross-validation within the training set yielded an overall classification accuracy of 90.1% (table S1). Spatial distributions of cancer cells, stromal cells, and lymphocytes in a sample can be visualized in a heat map (Fig. 2C and fig. S1), and cellular heterogeneity of a complete data set can be summarized in a triangle plot, where each axis represents one of the three cell types (Fig. 2D). Detailed description and implementation of our pipeline, together with reproducible code and data, are provided in the Supplementary Materials as Sweave files and an R package CRImage.

Image classifications are consistent with pathology and biology

To assess the validity of our image analysis platform, we first correlated it with commonly used clinical scores for cellularity and LI. Our quantitative scores of cancer cell density (cell-to-area ratio) correlated with the pathologists' scores [Fig. 3A; $P = 2 \times 10^{-11}$, Jonckheere-Terpstra (JT) test]. Our approach directly estimated tumor cellularity (cancer DNA content) by cancer cell proportions (cell-to-cell ratio), which also correlated with the pathologists' scores (Fig. 3B; $P = 6 \times 10^{-8}$, JT test). Our estimates of LI also agreed with the pathological scores (Fig. 3C; $P = 1.9 \times 10^{-24}$, JT test). Cell type proportions correlated with a pathologist's quantitative evaluation of a representative subset of 10,000 cells in 20 tumors (Fig. 3D) (Supplementary Materials).

Because cellular heterogeneity can be reflected on the molecular level, we correlated cellular proportions with gene expression data. Functional enrichment analysis (*15*, *16*) showed expected enriched pathways and biological processes for the genes highly correlated with the cellular proportions of each cell type, using both KEGG (*15*) (fig. S2 and table S2) and gene ontology (*16*) (table S3) databases. For example, in KEGG, "cell division" and "cell cycle" pathways were correlated with the proportion of cancer cells; "angiogenesis" and "ECM-receptor interactions" with the proportion of stromal cells; and "B cell receptor signaling pathway" and "T cell activation" with the proportion of lymphocytes. In addition, known cell type–specific markers positively correlated with cell proportions (fig. S2). The proportion of cancer cells correlated with overexpression of cell cycle genes, including *E2F5* and *MCM7*, and the proportion of stromal cells correlated with overexpression of genes encoding extracellular matrix proteins, including *FBLN1*, *FBLN2*, *COL6A2*, and *COL6A3*. Last, the proportion of lymphocytes correlated with overexpression of genes encoding apoptosis-related proteins, such as *DFFB*. Together, these observations based on pathological review, curated biological databases, and known gene markers support the biological validity of our image analysis results.

Quantitative tumor cellularity estimates correct cancer copy number profiles

Existing statistical methods estimate tumor cellularity indirectly from molecular data by assuming independence of allele-specific signals, identification of discrete copy number events (*17*), or equality of copy number in different clones, which can often be unrealistic. We avoided these assumptions by calculating tumor cellularity as the cancer cell

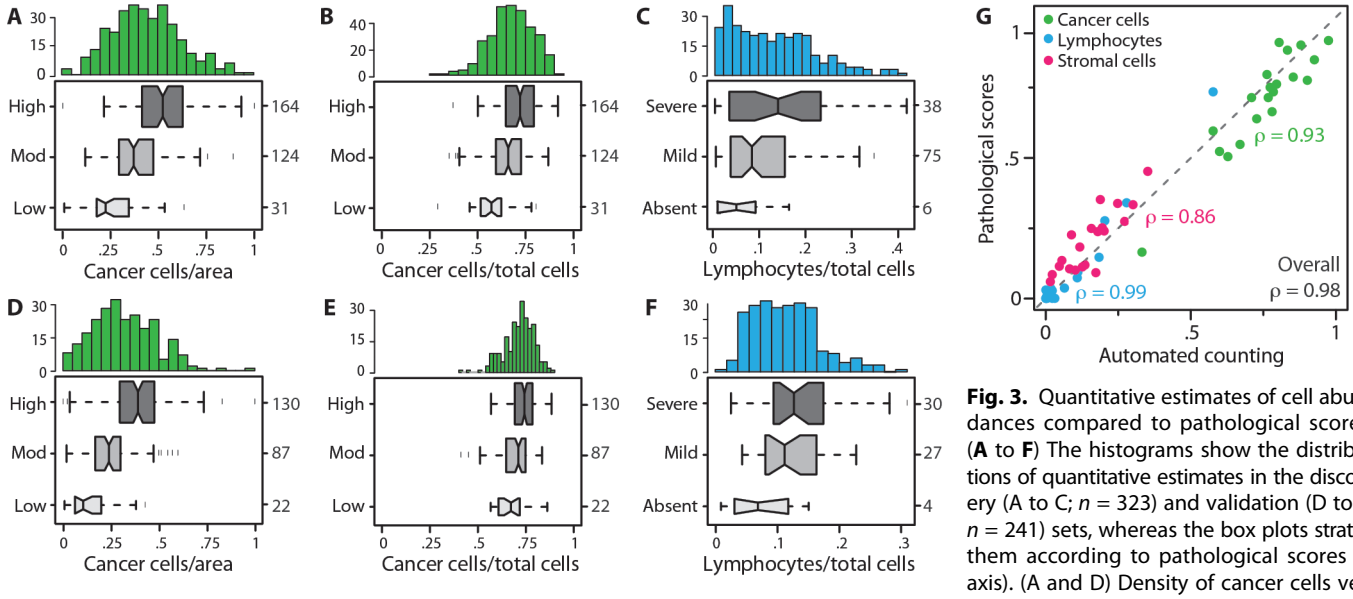

Fig. 3. Quantitative estimates of cell abundances compared to pathological scores. (**A** to **F**) The histograms show the distributions of quantitative estimates in the discovery (A to C; $n = 323$) and validation (D to F; $n = 241$) sets, whereas the box plots stratify them according to pathological scores (y axis). (A and D) Density of cancer cells versus pathological cellularity scores of high, moderate, or low. (B and E) Proportion of cancer cells in all cells (an estimate of cancer DNA content) versus pathological cellularity scores. (C and F) Proportion of lymphocytes versus pathological scores of "absent," "mild," or "severe" LI. (**G**) Cellular proportions obtained by automated image analysis compared to counts by a pathologist for a total of 10,000 cells in a representative set of 20 tissue samples.

proportion directly from H&E images and then correcting copy number data using a statistical model for the log ratio (LR) between observed and expected signal intensities and the B-allele frequency (BAF), a normalized measure of the allelic intensity ratio. We first segmented copy number events on the basis of LR values and then rescaled LR and BAF in each segment according to quantitative cancer cell proportions. In this way, we avoided an inflation of spurious events by only scaling copy number segments called on the original, uncorrected data.

Cellularity-corrected copy number profiles are able to remove effects of normal cell contaminations in the original profiles. For instance, in Fig. 4A, the uncorrected data exhibited a contradictory pattern where the *q* arm had four bands in the BAF plot (indicating an amplified region), whereas the corresponding LR values were negative (indicating a loss). This effect was due to normal cell contamination, which made the typical two-banded pattern of a loss appear as four bands, and was greatly reduced in the corrected BAF plot. Thus, cellularity correction produces cleaner profiles that make it easier to discriminate between copy number regions and to delimit breakpoints. To computationally evaluate our algorithm on a global scale, we contrasted the ratio of between-segment variance to within-segment variance (signal-to-noise ratio) in the original and corrected profiles for all samples in both the discovery and the validation sets (Fig. 4B). All points lie above the main diagonal, indicating that our algorithm globally increased the signal-to-noise level of somatic aberrations.

To experimentally evaluate the accuracy in recovering important aberration lost due to contamination, we concentrated on HER2 (human epidermal growth factor receptor 2) amplifications—one of the most important aberrations for breast cancer diagnostics. Estimates of HER2 amplification from corrected and uncorrected microarray data were compared with events validated by the current clinical gold standard, fluorescence in situ hybridization (FISH), on paraffin-embedded samples from the same tumors (Supplementary Materials). In a set of 78 samples, FISH found 4 samples with low-level and 12 samples with high-level HER2 amplifications [a total of 21% of the cases, consistent with previous studies (*18*, *19*)]. Corrected microarray copy number data showed higher concordance to FISH scores than the uncorrected data (Fig. 4C).

To validate our cellularity correction algorithm in the controlled setting of a dilution series, we obtained copy number profiles from human breast cancer cell line HCC2218 and its matched normal breast cell line HCC2218BL. We applied our copy number correction algorithm and compared diluted cancer DNA profiles to the pure cancer profile. Cellularity correction increased sensitivity to detect DNA aberrations by an average of 1.6-fold (±0.6/SD) while keeping specificity stable over the whole range of tumor cellularity found in the patient samples (30 to 90%, Fig. 4D). In summary, our algorithm enabled discovery of focal aberrations that could have been lost due to heterogeneous cellular composition and made copy number events comparable across samples of varying tumor cellularity.

Histopathologic-genomic integration predicts estrogen receptor–negative breast cancer survival

In a second application, we show how our quantitative analysis of LI can predict survival of estrogen receptor (ER)–negative breast cancer

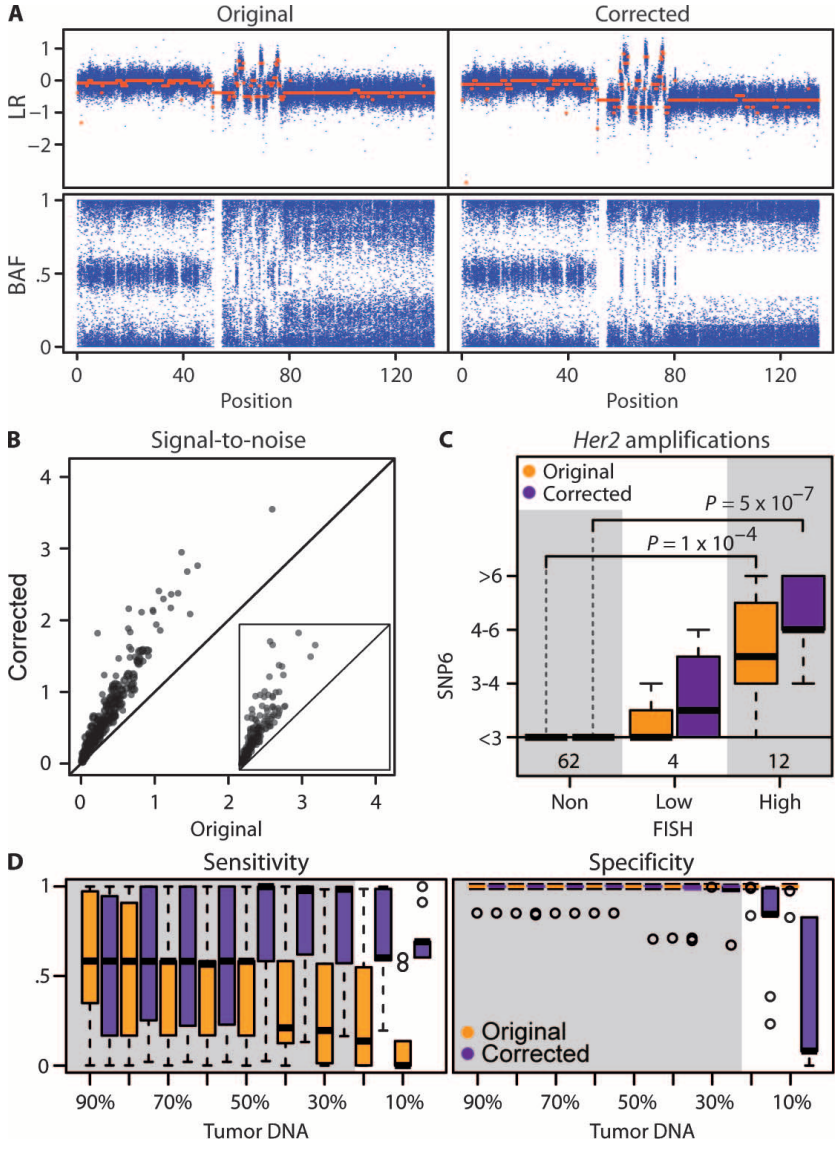

Fig. 4. Quantitative cellularity scores correct copy number profiles. (**A**) A genomic deletion on chromosome 11 is shown by LR and BAF before and after correction. (**B**) Cellularity correction increases the signal-to-noise ratio in copy number profiles. The large plot is the discovery cohort (n = 323), and the inset is the validation cohort (n = 241). Each dot in the scatter plot represents the ratio of between-segment variance to within-segment variance for one chromosome in a sample before (x axis) and after (y axis) correction. (**C**) Comparison of HER2 status established by FISH to HER2 copy number levels from microarray data (n = 78). P values were determined by Student's t test. (**D**) Sensitivity and specificity for original and corrected copy number profiles in a tumor DNA dilution series compared to the profile obtained from the pure (100%) cancer cell line. Box plots show variation over a range of possible cutoffs to identify aberrations. The gray area spans the region of tumor cellularity (≥30%) observed in the patient samples.

patients. Within ER-negative breast cancers, there are different clinical outcomes, which are thought to be influenced by the tumor microenvironment (*20*), and gene expression signatures of LI have identified a subset of ER-negative breast cancer patients with better prognostic outcomes (*7*, *21*, *22*). High LI may reflect an immune recognition of the tumor, leading to an active immune response. ER-negative samples with few lymphocytes may thus be associated with poor immune function and unimpeded tumor growth. Therefore, quantifying the lymphocyte proportion in ER-negative cancer may distinguish cases with active versus less-active antitumor immune responses.

Our quantitative approach directly interrogated lymphocytic proportions in ER-negative patients (discovery, $n = 54$; validation, $n = 61$) using H&E-stained images. Using the discovery set to calibrate a threshold for lymphocytic proportions (fig. S3), we stratified both discovery and validation cohorts into LI-low (<8%) and LI-high (≥8%) subgroups. Kaplan-Meier survival curves revealed distinctly different outcomes for these two groups in both cohorts, where higher lymphocyte proportion is associated with good outcome (Fig. 5A). Multivariate Cox regression showed that lymphocytic proportion is a prognostic factor independent of node status, size, and grade in both sample sets (Table 1).

Compared to pathological scores, our image-based LI estimates have consistent prognostic values indicated by both concordance and hazard ratio (Table 1). The categorical pathological scores in the discovery set showed no difference in outcome among different LI groups, whereas the validation set showed a poor prognosis in the LI-low group (Fig. 5A).

We compared the performance of our image-based method and previous gene expression–based estimates (*7*) to determine LI and patient outcome. An SVM predictor constructed using a published expression signature (*7*) (Supplementary Materials) achieved 67 ± 4.7% cross-validation accuracy in predicting ER-negative breast cancer survival (mean and SD of 20 repeats), whereas an SVM using our image-based signature improved the accuracy to 75 ± 1.5%. Both approaches accurately identified the nonsurviving patients as LI-low

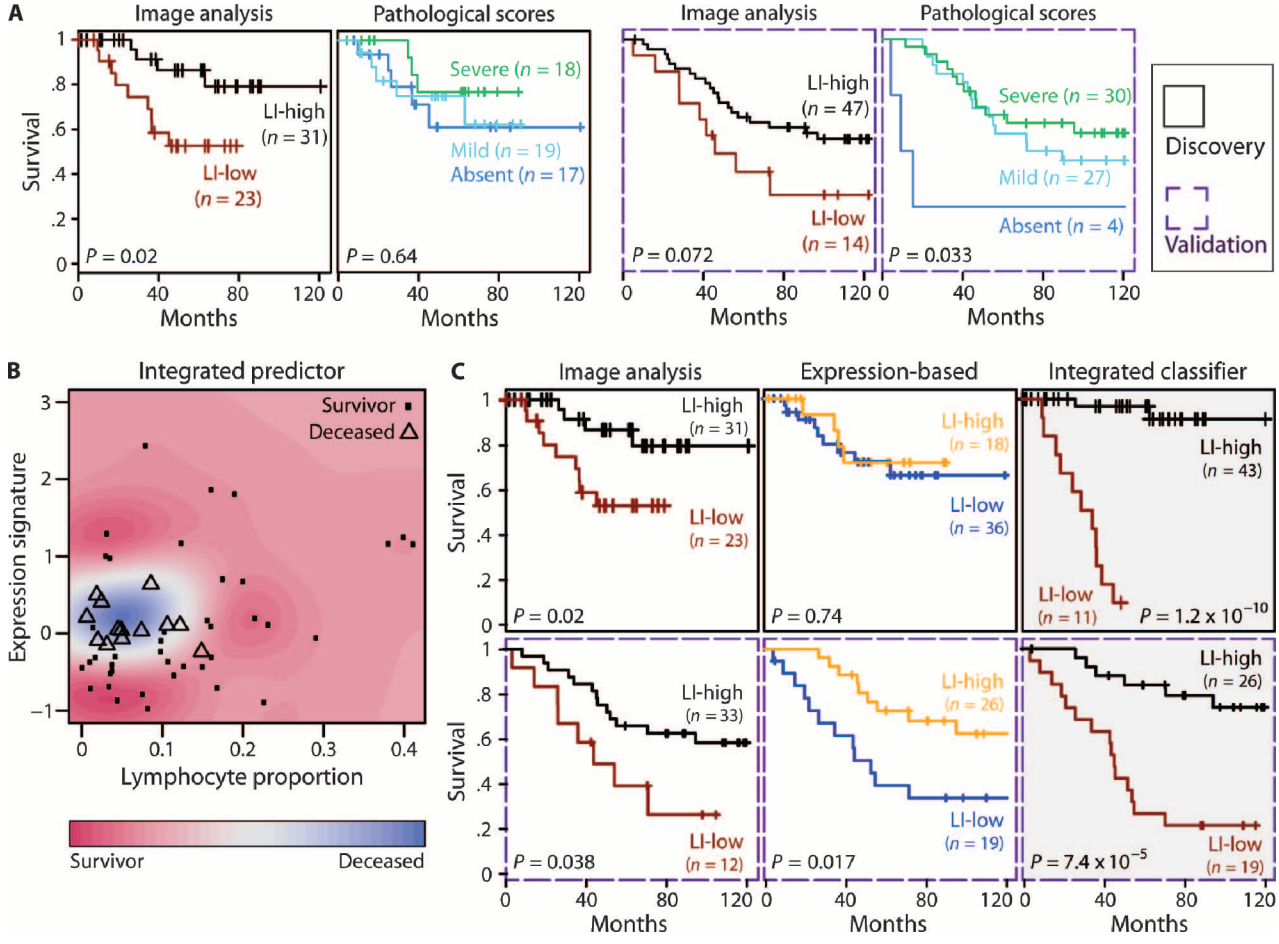

Fig. 5. LI scored by image analysis, pathologists, a gene expression signature, and a proposed integrated predictor in ER-negative breast cancer. (**A**) Kaplan-Meier curves of ER-negative tumors comparing LI scored by image analysis and by pathologists. For both discovery ($n = 54$) and validation ($n = 61$) cohorts, tumors were split into two groups on the basis of image-based lymphocyte proportions, where weak LI (≤8%) was called LI-low and severe LI (≥8%) was called LI-high. The same samples were also categorized by pathologists as absent, mild, or severe LI. (**B**) Scatter plot of image-based LI scores and a published gene expression signature for LI (*7*). Colors correspond to an integrated classifier using both types of information. (**C**) Kaplan-Meier curves for samples with expression data (discovery set, $n = 54$; validation set, $n = 45$) stratified by image-based lymphocyte proportions, expression signature, and integrated predictors. *P* values in (A) and (C) were calculated with the log-rank test.

Table 1. Prognostic values of various factors in ER-negative tissue samples with expression data. Both univariate (uni-) and multivariate (multi-) Cox regression models were considered for the discovery (n = 54) and validation (n = 61) cohorts. Proposed new prognostic factors in this paper are shown in italics. Nodes were characterized as negative or positive. Tumor size was described as small, medium, or large. Tumor grade was classified as low, medium, or high. HR (CI), hazard ratio (lower-upper 95% confidence interval).

Prognosticator	Variable	Discovery set			Validation set		
		HR (CI)	*P*	Concordance	HR (CI)	*P*	Concordance
Image-based lymphocyte proportion	Uni-	0.27 (0.08–0.88)	0.03	0.685	0.4 (0.16–0.98)	0.044	0.6
	Multi-	0.21 (0.06–0.71)	0.013	0.751	0.3 (0.11–0.79)	0.015	0.697
	Multi-node	1.79 (0.54–5.97)	0.345		0.85 (0.29–2.51)	0.77	
	Multi-size	2.9 (0.89–9.47)	0.078		2.58 (1.09–6.13)	0.031	
	Multi-grade	3.56 (0.3– 41.97)	0.313		0.77 (0.29–2.08)	0.612	
Pathological LI score	Uni-	0.69 (0.22–2.12)	0.516	0.545	0.24 (0.07–0.81)	0.022	0.577
	Multi-	0.47 (0.14–1.58)	0.223	0.702	0.09 (0.02–0.47)	0.004	0.615
	Multi-node	2.08 (0.6–7.22)	0.246		1.34 (0.41–4.38)	0.626	
	Multi-size	2.46 (0.76–7.9)	0.132		2.69 (1.2–6.05)	0.017	
	Multi-grade	1.82 (0.16–21)	0.631		1.03 (0.35–3.07)	0.955	
Expression LI signature	Uni-	0.82 (0.25–2.66)	0.74	0.531	0.36 (0.15–0.87)	0.022	0.646
	Multi-	0.59 (0.17–2.04)	0.409	0.694	0.36 (0.13–1.01)	0.053	0.699
	Multi-node	2.32 (0.62–8.68)	0.213		0.74 (0.26–2.12)	0.579	
	Multi-size	1.91 (0.64–5.73)	0.245		1.47 (0.66–3.25)	0.344	
	Multi-grade	1.1 (0.1–11.61)	0.935		0.85 (0.34–2.15)	0.736	
Integrated LI	Uni-	0.03 (0.01–0.16)	0.00012	0.836	0.17 (0.07–0.46)	0.00039	0.711
	Multi-	0.02 (0–0.12)	6×10^{-5}	0.882	0.08 (0.02–0.3)	0.00013	0.772
	Multi-node	0.4 (0.08–2)	0.266		0.25 (0.07–0.9)	0.034	
	Multi-size	1.44 (0.56–3.69)	0.448		2.57 (0.99–6.72)	0.053	
	Multi-grade	10.4 (0.83–130.71)	0.07		1.17 (0.46–2.96)	0.739	
Stromal spatial pattern	Uni-	3.55 (1.09–11.59)	0.036	0.666	2.84 (1.09–7.41)	0.033	0.618
	Multi-	4.82 (1.25–18.49)	0.022	0.769	3.58 (1.27–10.06)	0.016	0.693
	Multi-node	1.59 (0.47–5.41)	0.46		0.84 (0.3–2.33)	0.739	
	Multi-size	2.49 (0.88–7.06)	0.087		2.65 (1.15–6.12)	0.022	
	Multi-grade	0.64 (0.06–7.4)	0.723		1.1 (0.41–2.96)	0.853	

(Fig. 5B) but agreed less well on the set of survivors (correlation ρ = 0.24), suggesting that respective biases of survival are orthogonal. Therefore, we reasoned that a joint classifier would yield improved diagnostic performance. Our SVM predictor integrating gene expression and image data achieved 86 ± 3.0% cross-validation accuracy and improved stratification of the patient cohorts (Fig. 5C). Death rates within 5 years of diagnosis of the LI-low groups as determined by both data sources were 92 and 78% in discovery and validation cohorts, respectively. Integrating images with molecular data improved the separation of outcome in patient groups and located patients with poor prognosis in both cohorts.

Spatial distribution of stromal cells is an independent prognostic factor for survival

Our image-based approach allowed us to quantitatively analyze tumor architecture and base prognosis on cellular patterns in the tumor microenvironment that are invisible to molecular assays. Spatial variations between cell types have previously been reported in small collections of breast cancers (*23*). To describe whether specific cell types are confined in clusters or uniformly scattered, we computed a quantitative score based on Ripley's K function (*24*) (Supplementary Materials). The K score summarizes pairwise distances between a single type of cells and assigns a high score to highly clustered patterns and a low score to randomly scattered patterns (fig. S4). Only in stromal cells, this score identified a high-risk subgroup of ER-negative patients within the interquartile range (25 to 75% quantiles) (Fig. 6A) of the population in both discovery and validation sample sets. Patients with high or low K score (extreme K), indicating highly clustered or randomly scattered stromal cell patterns, have a significantly better outcome than other patients (medium K) (Fig. 6B).

The K scores did not correlate with the cell-to-area ratio, indicating that this association is not related to cell density (Fig. 6C). ER-positive patients show no stratification difference (fig. S5), suggesting that the prognostic value of stromal cell patterns is specific to ER-negative

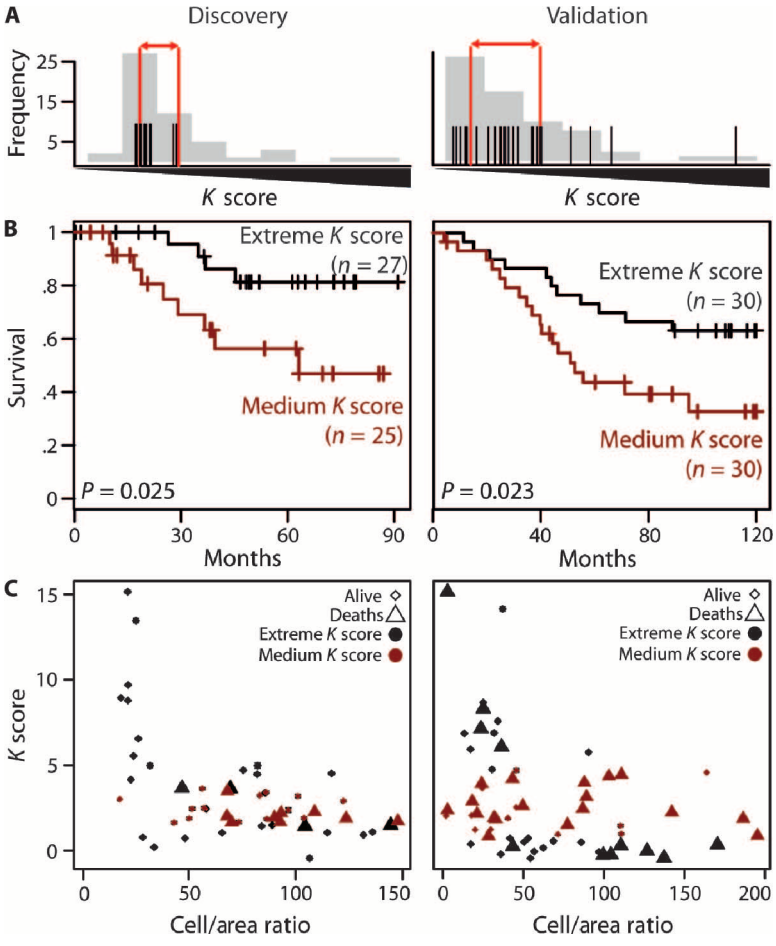

Fig. 6. Spatial patterns of cells define prognostic subtypes for ER-negative tumor samples. The *K* score summarizes the degree of clustering for a single type of cells and assigns a high score to highly clustered cell patterns and a low score to randomly scattered patterns. (**A**) Breast cancer–specific deaths (black lines) are enriched in the interquartile range (red bars) of the *K* score distribution (gray histogram). (**B**) Kaplan-Meier curves of patients in both cohorts that have sufficient stromal cells for computing the spatial scores (discovery set, $n = 52$; validation set, $n = 60$), with medium *K* scores (within interquartile range of the distribution) and extreme *K* scores (outside interquartile range). *P* values were calculated with the log-rank test. (**C**) Scatter plots comparing stromal cell *K* score and stromal cell-to-tissue area.

DISCUSSION

Pathological analysis of tumor architecture and cell morphology has a long history, reaching back to the 19th century. Applying modern, automated image analysis procedures to large sample collections promises new quantitative insights in this well-established field. For example, by analyzing tissue microarrays, a recent study by Beck *et al.* (*11*) found a surprising role for stromal tissue in predicting overall survival of breast cancer patients. Here, we have extended this line of research by integrating pathology with genomics in two large independent cohorts to advance prognostic markers as they relate to cancer subtypes. Gene expression and copy number alteration data have refined tumor classification strategies; we show here how automated pathology that exploits a widely available, yet underused, resource (H&E-stained tissue sections) can improve clinical diagnosis. Our approach—freely available as an R package—quantifies cellular heterogeneity of tumors from H&E images and uses this information to complement molecular data. In particular, we show that the proportion of lymphocytes in a tumor is complementary to gene expression signatures for lymphocyte infiltration and that integrating both types of evidence boosts survival prediction: The gene expression classifier had 67% cross-validation accuracy in predicting disease-specific deaths, the image-based classifiers had 75%, and the integrated classifier reached 86%.

It is crucial that the cellular heterogeneity quantifications are accurate and objective for the purpose of integration with molecular data. Our image processing system was accurate in quantifying cellular contents, as evaluated by both categorical and cell-counting pathological assessment. Meanwhile, it avoids biases or discrepancies arising from manual scoring. Pathological scores of LI were prognostic in one cohort scored by a single pathologist but lost power in the other cohort scored by multiple pathologists. In contrast, image-based LI scores differentiated good and poor outcome groups in both cohorts, thereby providing a solution for removing variances between different pathologists.

Our approach has the potential to provide new clinical predictors and better understanding of the tumor microenvironment by using already available histopathological information. Here, we found that stromal cell patterns (spatial arrangement) in tumors can be independent prognosticators for breast cancer. These findings add to an increasing body of evidence demonstrating the role of the stroma in mediating breast tumor development and in promoting tumor growth (*11*, *25*–*28*). Although the mechanistic basis underpinning the prognostic value of stromal patterns is not known, our findings highlight the importance of accounting for architectural differences in attempting to understand tumor-microenvironment interactions.

The biggest limitation of this approach is that it requires matched molecular and image data. Fortunately, major international consortia (*1*, *2*, *12*) that collect large compendia of molecular data also routinely collect histopathological images. H&E stainings as we use them only provide coarse information on specific cell types, and more specific stainings are needed to identify, for example, different types of immune cells. This might partly explain the lack of correlation we observe between gene expression signatures for LI and lymphocyte counts from images.

tumors. In multivariate Cox regression, we verified that the stromal cell spatial pattern is a prognostic factor independent of node, size, and grade in both cohorts (Table 1). After multiple testing corrections for correlations between gene expression and this spatial stromal pattern, we found no significant correlation, indicating that we identified features in the higher-order organization of the tumor that are augmenting—not supplanting—molecular characterizations. Little overlap between the two image-based stratifications (LI and stromal spatial pattern) (fig. S6) suggests potential of aggregating these two factors. Our observations corroborate a recent analysis of morphological features predictive for survival (*11*), which showed the importance of spatial relationships between cells in the tumor tissue and found a strong impact of stromal features on survival.

Another limitation is the quality of images. In particular, frozen tumor slides contain many artifacts of the cutting process. Additionally, large collections can suffer from variability in stainings and batch effects. We screen for artifacts during the automated classification, but more work has to be done to further refine automated quality assessment of images, normalization of varying degrees of staining, and removal of batch effects.

Our image analysis methods can be applied to paraffin-embedded sections, which in general have fewer artifacts than frozen sections. The approach we describe may also be used for other epithelial cancer types and to deconvolute different molecular assays; for example, proportions of different cell types could be incorporated into established methods for gene expression deconvolution (*29*, *30*), and next-generation sequencing data can be corrected much like how we treated microarray copy number data (*31*). Because our algorithms rely on the quality of histopathological stainings and images, with more and better quantitative systems for computational analysis of pathological images, we expect such analytical methods to become more accepted and widely implemented in pathology laboratories.

What has limited translation to clinical practice to date is the gap between visual pathological and quantitative molecular analyses, which our integrated approach will help to bridge by using unbiased quantitative methods in both areas. The availability of tools for digital image analysis, such as the ones reported here, together with widespread genomic characterization of tumors, will provide us with an unprecedented systems-level view of cancer, which will have profound practical consequences in the way cancer hospitals and clinics will be set up in the future. Pathology, clinical genomics, and bioinformatics/computational biology will have to be fully integrated into cancer care.

MATERIALS AND METHODS

Sample collection

Primary frozen breast tumors from the discovery and validation sets were collected and stained independently in different laboratories contributing to the METABRIC consortium (*12*). All samples included H&E images, and the majority had molecular data passing quality control (discovery, 318 of 323; validation, 225 of 241). Control DNA from match normal breast tissues was available for 473 samples and RNA from 144 of these individuals. Details about the sample makeup and selection can be found in (*12*). All histopathological images were scanned with a ScanScope TX scanner (Aperio Technologies Inc.) providing images at 200-fold magnification. DNA and RNA were extracted from each primary tumor specimen and subjected to copy number and genotype analysis on the Affymetrix SNP 6.0 platform and transcriptional profiling on the IlluminaHT-12 v3 platform (Illumina_Human_WG-v3).

Image processing pipeline

Our image pipeline used an SVM classifier to provide initial classification of cells based on morphological features, a kernel smoother to account for the neighbors of a cell, and a hierarchical model for incorporating a global view of the image by a multiresolution information flow (details provided in the Supplementary Sweave file). To train the SVM classifier, a pathologist (H.R.A.) labeled 871 cells. The kernel smoother then performed spatial smoothing for cells in spatial proximity to cancer cells. Subsequently, a hierarchical model provided a top-down view of the image by outlining cancer cell clusters to further improve classification accuracy. This entire image processing pipeline and the subsequent cellularity correction method are supplied as an R package CRImage, together with code and data for reproducing all our results available as Supplementary Sweave files.

Algorithm for cellularity correction

The algorithm we used to obtain corrected microarray copy number data is described in detail in the Supplementary Materials.

Quantitative lymphocyte proportions contribute to accurate prediction for ER-negative patient survival

We used the lymphocyte proportions as our image-based estimation of LI in ER-negative samples. To train this LI estimation in the discovery set and separate patients into LI-low and LI-high groups, we first calculated *P* values using log-rank test with various cutoffs in the range of 5 to 20% lymphocytes. The range was chosen to obtain reasonable patient group sizes. The optimal *P* value was obtained at 8% cutoff (fig. S5). Hence, to dichotomize the continuous image-based signature into two groups, we used 8% lymphocytes as the threshold to differentiate low- and high-LI groups.

In ER-negative samples, we compared the image-based lymphocyte proportions with pathological scores in both discovery and validation sets. Because there are different numbers of groups in the signatures, we used the concordance index (*32*) to quantify the predictive ability of a survival model. For each type of scores, we also calculate the *P* values from univariate Cox proportional hazards regression models. For comparison among different variables in Table 1, we merged mild and severe pathological scores as "high" so that all signatures had LI-low and LI-high groups.

The expression signature. The expression signature of Calabrò *et al.* (*7*) took the mean of standardized expression of 18 lymphocyte marker genes, of which 13 were profiled and passed quality check in our data: *LCK*, *CD8A*, *CD14*, *LTB*, *MS4A1*, *CD3E*, *CD3D*, *CD3G*, *CD19*, *CCL5*, *CD79B*, *CD79A*, and *CD37*. Patients were partitioned into two groups with a size ratio of 1:2 (an LI-low group comprising one-third of the patients, and an LI-high group with the remaining), following (*7*).

The SVM predictors. SVM predictors were built with a Gaussian radial basis kernel (*33*) using the ksvm function from the kernlab R package (*34*). Accuracy for the SVM predictions was calculated by running 10-fold cross-validation 20 times.

Spatial patterns with the *K* function

Ripley's *K* function provides descriptions of spatial point patterns at different distance scales (*24*). With distance variable *r* and a total number of *N* points in the sampling region, *K* function is defined as follows:

$$K(r) = \sum_{i,j} \frac{I(d_{i,j} < r) \times e_{i,j}}{\lambda \times N} \quad (1)$$

where d_{ij} is the distance between points *i* and *j*, λ is the cell density in the window, and e_{ij} is an edge correction function that is important to avoid a biased estimation of $K(r)$ owing to edge effects. *K* captures spatial interdependencies between points over a range of distances *r*, generating quantitative scores summarizing $K(r)$ to characterize patterns of points. However, our task is of high complexity and would be computationally prohibitive: There are 61,090 ± 59,285 (median ± SD) points/cells in a whole slide image, and the window space is highly complex because of irregular sections and fragmented tissues. Therefore, we adapted *K* function to a resampling scheme as our problem-specific solution for our objective (Supplementary Materials).

Quantitative scoring of stromal cell spatial patterns in ER-negative tumors

Although spatial statistics have been applied to explore cell patterns before (*23*), our task was challenging because frozen tissue sections are of irregular shapes and often fragmented; as such, the window for computing the *K* function (*24*) is complex. Moreover, whole-slide images have a large number of points. The number of cells on a slide ranged from 1000 to 250,000, which would be computationally prohibitive for computing *K* with complex windows and edge correction. Hence, we focused on adapting this statistical method to generate scores comparable across samples (Supplementary Materials). Note that some samples failed the computation if there were few stromal cells or the tissues were too fragmented (discovery, 6 of 323; validation, 12 of 241).

Statistical methods

Survival analysis was performed with breast cancer–specific 10-year survival data. Kaplan-Meier estimator was used for patient stratification, and *P* values were calculated with the log-rank test. Cox proportional hazards regression model (*35*) was fitted and 95% confidence intervals were computed to determine the prognostic values of various factors, where $P < 0.05$ was considered significant. Correlation was computed with Spearman's ρ. Tests for trends of continuous variables among groups were performed with JT trend test (*36*) for an alternative hypothesis that these variables have a monotone trend.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/4/157/157ra143/DC1
Materials and Methods
Fig. S1. Visualization of cancer cell distribution in a tumor section.
Fig. S2. Top genes expressed in different cell types and their enrichment.
Fig. S3. Selecting cutoffs for determining the image-based low lymphocyte infiltration (LI) group in the discovery set of ER-negative samples.
Fig. S4. A toy example of quantifying stromal cell spatial patterns with *K* statistics.
Fig. S5. Stromal cell spatial pattern is not prognostic in ER-positive breast cancer in either the discovery or the validation cohort.
Fig. S6. Patient stratification using combined image-based LI and stromal spatial pattern in ER-negative samples.
Table S1. Classification accuracy by 10-fold cross-validation for artifacts, cancer cells, lymphocytes, and stromal cells based on the training set.
Table S2. KEGG pathways enriched in the top 500 genes expressed in cancer cells, stromal cells, and lymphocytes.
Table S3. Gene Ontology Biological Process enriched in the top 500 genes expressed in cancer cells, stromal cells, and lymphocytes.

REFERENCES AND NOTES

1. Cancer Genome Atlas Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
2. International Cancer Genome Consortium, International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
3. C. Garnis, B. P. Coe, S. L. Lam, C. MacAulay, W. L. Lam, High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples. *Genomics* **85**, 790–793 (2005).
4. N. Tørring, M. Borre, K. D. Sørensen, C. L. Andersen, C. Wiuf, T. F. Ørntoft, Genome-wide analysis of allelic imbalance in prostate cancer using the Affymetrix 50K SNP mapping array. *Br. J. Cancer* **96**, 499–506 (2007).
5. R. Kalluri, M. Zeisberg, Fibroblasts in cancer. *Nat. Rev. Cancer* **6**, 392–401 (2006).
6. R. Rajan, A. Poniecka, T. L. Smith, Y. Yang, D. Frye, L. Pusztai, D. J. Fiterman, E. Gal-Gombos, G. Whitman, R. Rouzier, M. Green, H. Kuerer, A. U. Buzdar, G. N. Hortobagyi, W. F. Symmans, Change in tumor cellularity of breast carcinoma after neoadjuvant chemotherapy as a variable in the pathologic assessment of response. *Cancer* **100**, 1365–1373 (2004).
7. A. Calabrò, T. Beissbarth, R. Kuner, M. Stojanov, A. Benner, M. Asslaber, F. Ploner, K. Zatloukal, H. Samonigg, A. Poustka, H. Sültmann, Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Res. Treat.* **116**, 69–77 (2009).
8. P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A. L. Børresen-Dale, V. N. Kristensen, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16910–16915 (2010).
9. G. Assié, T. LaFramboise, P. Platzer, J. Bertherat, C. A. Stratakis, C. Eng, SNP arrays in heterogeneous tissue: Highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am. J. Hum. Genet.* **82**, 903–915 (2008).
10. P. Neuvial, H. Bengtsson, T. P. Speed, Statistical analysis of single nucleotide polymorphism microarrays in cancer studies, in *Handbook of Statistical Bioinformatics, Springer Handbooks of Computational Statistics*, H. H.-S. Lu, B. Schölkopf, Z. Hongyu, Eds. (Springer, Berlin, 2011).
11. A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, D. Koller, Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
12. C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney; METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, S. Aparicio, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
13. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
14. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).
15. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
16. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
17. G. Yu, B. Zhang, G. S. Bova, J. Xu, I. M. Shih, Y. Wang, BACOM: In silico detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics* **27**, 1473–1480 (2011).
18. M. A. Owens, B. C. Horten, M. M. Da Silva, HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clin. Breast Cancer* **5**, 63–69 (2004).
19. D. J. Slamon, W. Godolphin, L. A. Jones, J. A. Holt, S. G. Wong, D. E. Keith, W. J. Levin, S. G. Stuart, J. Udove, A. Ullrich, M. F. Press, Studies of the HER-2/*neu* proto-oncogene in human breast and ovarian cancer. *Science* **244**, 707–712 (1989).
20. F. Andre, N. Berrada, C. Desmedt, Implication of tumor microenvironment in the resistance to chemotherapy in breast cancer patients. *Curr. Opin. Oncol.* **22**, 547–551 (2010).
21. A. E. Teschendorff, A. Miremadi, S. E. Pinder, I. O. Ellis, C. Caldas, An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol.* **8**, R157 (2007).
22. A. E. Teschendorff, C. Caldas, A robust classifier of high predictive value to identify good prognosis patients in ER-negative breast cancer. *Breast Cancer Res.* **10**, R73 (2008).
23. T. Mattfeldt, S. Eckel, F. Fleischer, V. Schmidt, Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *J. Microsc.* **235**, 106–118 (2009).
24. B. D. Ripley, *Statistical Inference for Spatial Processes* (Cambridge University Press, Cambridge, 1988).
25. E. S. Radisky, D. C. Radisky, Stromal induction of breast cancer: Inflammation and invasion. *Rev. Endocr. Metab. Disord.* **8**, 279–287 (2007).
26. C. G. Kleer, N. Bloushtain-Qimron, Y. H. Chen, D. Carrasco, M. Hu, J. Yao, S. K. Kraeft, L. C. Collins, M. S. Sabel, P. Argani, R. Gelman, S. J. Schnitt, I. E. Krop, K. Polyak, Epithelial and stromal cathepsin K and CXCL14 expression in breast tumor progression. *Clin. Cancer Res.* **14**, 5357–5367 (2008).
27. S. Krause, M. V. Maffini, A. M. Soto, C. Sonnenschein, The microenvironment determines the breast cancer cells' phenotype: Organization of MCF7 cells in 3D cultures. *BMC Cancer* **10**, 263 (2010).
28. K. Pietras, A. Ostman, Hallmarks of cancer: Interactions with the tumor stroma. *Exp. Cell Res.* **316**, 1324–1331 (2010).
29. S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, A. J. Butte, Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
30. J. Clarke, P. Seo, B. Clarke, Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* **26**, 1043–1049 (2010).

31. L. Ding, M. J. Ellis, S. Li, D. E. Larson, K. Chen, J. W. Wallis, C. C. Harris, M. D. McLellan, R. S. Fulton, L. L. Fulton, R. M. Abbott, J. Hoog, D. J. Dooling, D. C. Koboldt, H. Schmidt, J. Kalicki, Q. Zhang, L. Chen, L. Lin, M. C. Wendl, J. F. McMichael, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, E. Appelbaum, K. Deschryver, S. Davies, T. Guintoli, L. Lin, R. Crowder, Y. Tao, J. E. Snider, S. M. Smith, A. F. Dukes, G. E. Sanderson, C. S. Pohl, K. D. Delehaunty, C. C. Fronick, K. A. Pape, J. S. Reed, J. S. Robinson, J. S. Hodges, W. Schierding, N. D. Dees, D. Shen, D. P. Locke, M. E. Wiechert, J. M. Eldred, J. B. Peck, B. J. Oberkfell, J. T. Lolofie, F. Du, A. E. Hawkins, M. D. O'Laughlin, K. E. Bernard, M. Cunningham, G. Elliott, M. D. Mason, D. M. Thompson Jr., J. L. Ivanovich, P. J. Goodfellow, C. M. Perou, G. M. Weinstock, R. Aft, M. Watson, T. J. Ley, R. K. Wilson, E. R. Mardis, Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
32. F. E. Harrell Jr., K. L. Lee, D. B. Mark, Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
33. J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, Cambridge, 2004).
34. A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, kernlab—An S4 package for kernel methods in R. *J. Statist. Soft.* **11**, 1–20 (2004).
35. D. R. Cox, D. Oakes, *Analysis of Survival Data* (Chapman and Hall, London, 1984).
36. A. R. Jonckheere, A test of significance for the relation between *m* rankings and *k* ranked categories. *Br. J. Stat. Psychol.* **7**, 93–100 (1954).
37. S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van de Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavaré, J. D. Brenton, B. Ylstra, C. Caldas, High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* **8**, R215 (2007).
38. A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
39. H. Willenbrock, J. Fridlyand, A comparison study: Applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091 (2005).
40. J. Kononen, L. Bubendorf, A. Kallioniemi, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, O. P. Kallioniemi, Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).
41. A. Baddeley, R. Turner, spatstat: An R package for analyzing spatial point patterns. *J. Statist. Soft.* **12**, 1–42 (2005).
42. J. Ohser, On estimators for the reduced second moment measure of point processes. *Math. Oper. Statist. Ser. Statist.* **14**, 63–71 (1983).

Acknowledgments: We acknowledge the support of the University of Cambridge, Cancer Research UK, Hutchison Whampoa Limited, and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre and the Cambridge Experimental Cancer Medicine Centre. We thank L. Blackburn for editorial support, and A. Purushotham, S. Pinder, and C. Gillett for contributing a subset of the samples. **Funding:** H.F. acknowledges support by an ERASMUS fellowship. H.R.A. is supported by a fellowship funded by Addenbrooke's Charitable Trust and the NIHR Cambridge Biomedical Research Centre. **Author contributions:** Y.Y. led the analysis, designed the experiments, and co-wrote the manuscript. H.F. wrote the R package and contributed to manuscript preparation. O.M.R. led the copy number correction algorithm design and experiment. H.R.A. provided pathological expertise and contributed to the study design. S.G., R.F.S., and C. Curtis contributed to the study design. S.-F.C. generated data and performed the experiment. M.J.D. performed the analysis. H.B. contributed to image generation. N.J. and S.D. performed the FISH experiment. G.T. and E.P. provided histopathology expertise. S.A. and C. Caldas provided data, contributed to overall study design, and are the METABRIC project leaders. F.M. designed the study and co-wrote the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** Corresponding images, copy numbers, and expression data are deposited in the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAS00000000098. Requests for materials should be addressed to C. Caldas (carlos.caldas@cancer.org.uk) or S.A. (saparicio@bccrc.ca).

Submitted 17 May 2012
Accepted 18 September 2012
Published 24 October 2012
10.1126/scitranslmed.3004330

Citation: Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas, F. Markowetz, Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143 (2012).

CANCER

A Correction to the Research Article Titled: "Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling" by Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas, F. Markowetz

There was an error in the text on page 5 stating that "Kaplan-Meier survival curves revealed distinctly different outcomes for these two groups in both cohorts, where higher lymphocyte proportion is associated with poor outcome." The correct statement is: "Kaplan-Meier survival curves revealed distinctly different outcomes for these two groups in both cohorts, where higher lymphocyte proportion is associated with good outcome." The corrected online version of the article is at http://stm.sciencemag.org/content/4/157/157ra143.full, and the corrected PDF version of the article is at http://stm.sciencemag.org/content/4/157/157ra143.full.pdf.

10.1126/scitranslmed.3005298

Citation: A correction to the Research Article titled: "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling" by Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas, F. Markowetz. *Sci. Transl. Med.* **4**, 161er6 (2012).

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Science Translational Medicine
AAAS

Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling

Yinyin Yuan *et al.*

Sci Transl Med **4**, 157ra143 (2012);
DOI: 10.1126/scitranslmed.3004330

Editor's Summary

Digitizing Pathology for Genomics

The tumor microenvironment is a complex milieu that includes not only the cancer cells but also the stromal cells, immune cells, and even normal, healthy cells. Molecular analysis of tumor tissue is therefore a challenging task because all this "extra" genomic information can muddle the results. Conversely, biopsy tissue staining can provide a spatial and cellular readout (architecture and content), but it is mostly qualitative information. In response, Yuan and colleagues have developed a quantitative, computational approach to pathology. When combined with molecular analyses, the authors were able to uncover new knowledge about breast tumor biology and, in turn, predict patient survival.

Yuan *et al.* first collected histopathology images, gene expression data, and DNA copy number variation data for 564 breast cancer patients. Using a portion of the images (the "discovery set"), they developed an image processing approach that automatically classified cells as cancer, lymphocyte, or stroma on the basis of their size and shape. This approach was validated on the remaining samples, and any errors in this analysis were digitally corrected before obtaining a plot of tumor cellular heterogeneity. With exact knowledge of the tumor's cellular composition, the authors were able to correct copy number data to more accurately reflect *HER2* status compared with uncorrected data.

Yuan and colleagues combined their digital pathology with genomic information to devise an integrated predictor of survival for estrogen receptor (ER) –negative patients. Higher number of infiltrating lymphocytes (immune cells) as quantified by their image analysis platform were found in a subset of patients with better clinical outcome than the rest of ER-negative patients, and this outcome difference was significantly enhanced with the addition of gene expression. The quantitative and objective nature of this integrated predictor could benefit diagnosis and prognosis in many areas of cancer by using the rich combination of tumor cellular content and genomic data.

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.