A pathway-based classification of human breast cancer

Michael L. Gatza[a], Joseph E. Lucas[a,b], William T. Barry[a,c], Jong Wook Kim[a,d], Quanli Wang[a,b], Matthew D. Crawford[a], Michael B. Datto[e], Michael Kelley[f], Bernard Mathey-Prevot[a,g], Anil Potti[a,f], and Joseph R. Nevins[a,d,1]

[a]Duke Institute for Genome Sciences and Policy, [b]Department of Statistical Science, [c]Department of Biostatistics and Bioinformatics, [d]Department of Molecular Genetics and Microbiology, [e]Department of Pathology, [f]Department of Medicine, and [g]Department of Pediatrics, Duke University Medical Center, Durham, NC 27710

Edited by Joan S. Brugge, Harvard Medical School, Boston, MA, and approved March 2, 2010 (received for review November 5, 2009)

The hallmark of human cancer is heterogeneity, reflecting the complexity and variability of the vast array of somatic mutations acquired during oncogenesis. An ability to dissect this heterogeneity, to identify subgroups that represent common mechanisms of disease, will be critical to understanding the complexities of genetic alterations and to provide a framework to develop rational therapeutic strategies. Here, we describe a classification scheme for human breast cancer making use of patterns of pathway activity to build on previous subtype characterizations using intrinsic gene expression signatures, to provide a functional interpretation of the gene expression data that can be linked to therapeutic options. We show that the identified subgroups provide a robust mechanism for classifying independent samples, identifying tumors that share patterns of pathway activity and exhibit similar clinical and biological properties, including distinct patterns of chromosomal alterations that were not evident in the heterogeneous total population of tumors. We propose that this classification scheme provides a basis for understanding the complex mechanisms of oncogenesis that give rise to these tumors and to identify rational opportunities for combination therapies.

cancer genomics | tumor subgroup

Breast cancer, like most cancers, represents a heterogeneous collection of distinct diseases that arise as a consequence of varied somatic mutations acquired during tumorigenesis (1). This heterogeneity is apparent in tumor ER or HER2 status or in the molecular classification schemes based on gene expression patterns that reflect the cellular origin of the tumor such as basal or luminal (2, 3). In short, breast cancer is a nonspecific description representing many distinct entities.

An ability to dissect breast cancer heterogeneity is critically important for two reasons. First, the ability to understand the significance of the genome alterations in breast cancer, which represent the underlying mechanisms of disease, requires a knowledge of distinct disease states rather than simply examining the heterogeneous population of tumors. Second, the development of therapeutic regimens that will be most effective for individual patients will depend on an ability to define the unique characteristics of the patient's tumor. This is particularly critical in addressing the challenge of treating advanced stage disease in which the standard chemotherapies are largely ineffective. Therefore, the challenge of effectively treating breast cancer patients is to identify subpopulations of individuals who are most likely to respond to a given therapy.

We have previously described the development of gene expression signatures that predict activation of various oncogenic signaling pathways, demonstrating a capacity to profile collections of tumor samples for patterns of pathway activity (4, 5). We have further described the use of these pathway signatures to reveal complexity in the intrinsic breast cancer subtypes (6). In light of this, we now have focused on the development of a strategy to classify human breast tumors on the basis of oncogenic and tumor suppressor pathway deregulation. Using this approach, we have identified 17 breast cancer subgroups that exhibit distinct patterns of pathway activation as well as clinical and biological characteristics. The distinctions between subgroups goes beyond a descriptive classification but rather is based on a predictive classification scheme that reflects the status of important signaling pathways. Additionally, each subgroup exhibits distinct patterns of chromosomal alterations suggesting that the classification scheme can serve as a framework for understanding the complex patterns of DNA aberrations within tumors. Finally, we show that this classification strategy enables the integration of patterns of predicted pathway activity, which correlate with sensitivity to pathway-specific drugs, and predicted response to cytotoxic agents that could aid in the development of potential therapeutic opportunities for breast cancer patients.

Results

Assessing the Molecular Heterogeneity of Human Breast Cancer. Previous work has used genome-scale gene expression measures, coupled with hierarchical clustering, to identify breast cancer subtypes based on distinct patterns of expression and that exhibit specific histological properties and clinical outcomes, suggesting that gene expression patterns can be used as a direct reflection of underlying genomic alterations (2, 3). These initial studies, however, used relative few samples and although this work has been extended in subsequent studies, each remains an analysis of a number of samples that might not capture the full complexity of the disease (7–9).

To investigate the phenotypic complexity of breast cancer, a large collection of breast tumor gene expression data (n = 1,143), derived from 10 independent studies and normalized using Bayesian Factor Regression Modeling (*SI Appendix*), was analyzed by unsupervised hierarchical clustering to reveal complex patterns of gene expression (Fig. 1*A*). The previously defined molecular subtypes (2, 3) were apparent from this analysis; in particular the basal and a large fraction of the luminal B subtypes are clearly delineated. The luminal A and Erbb2 subtypes are more diverse, potentially due to additional complexity revealed as a consequence of analyzing a larger number of samples.

To address the extent to which the full phenotypic diversity of breast cancer has been captured, a series of clustering analyses were performed using random subsets of an increasing number of tumor samples. As shown in Fig. 1*B* and in *SI Appendix*, the number of clusters at a given level in the dendrogram increases as a function of the sample size and then plateaus when the sample size reached approximately 700 tumors. This result suggests that the complexity of breast cancer is considerable and emphasizes the importance of a metaanalysis, such as presented here, to be able to capture the full extent of breast cancer heterogeneity. At the same time, this analysis

Author contributions: M.L.G., J.E.L., B.M.-P., A.P., and J.R.N. designed research; M.L.G. and M.D.C. performed research; M.L.G., J.E.L., W.T.B., J.W.K., Q.W., M.B.D., and M.K. contributed new reagents/analytic tools; M.L.G. and J.E.L. analyzed data; and M.L.G. and J.R.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. E-mail: j.nevins@duke.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0912708107/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0912708107

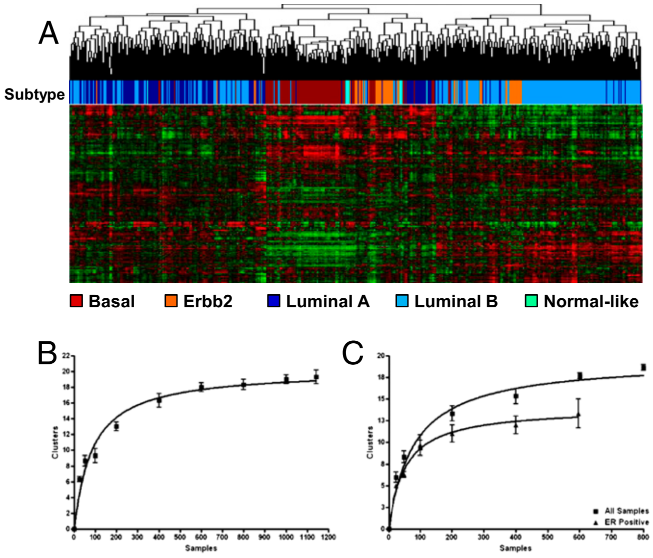

Fig. 1. **Meta-analysis of breast cancer gene expression patterns.** (*A*) A breast cancer dataset comprise of 1,143 samples derived from 10 independently generated datasets was clustered by complete linkage hierarchical clustering based on the gene expression patterns of Affymetrix U133A probes. The intrinsic subtype of each sample is reported. (*B*) The number of clusters identified in random subsets of the 1,143 samples demonstrates saturation of the complexity of expression patterns as a function of dataset size. (*C*) Analysis of the number of clusters identified in random subsets of tumors with known ER status ($n = 828$) compared to the number of clusters identified in subsets of ER+ tumors ($n = 596$).

suggests that a finite number of breast cancer subgroups can be identified.

To further validate that this analysis has captured the true biological complexity of breast cancer, we analyzed a subset of tumors from which a well-defined subcategory of disease, the ER negative tumors, was first removed. From the original dataset of 1,143 samples, a total of 828 had known ER status. Of these, 596 were ER positive and 232 were ER negative. The clustering analysis using random subsets was then repeated with the dataset of 828 for which ER status was known and with the dataset of samples from which the ER negative samples were removed. As shown in Fig. 1*C*, the analysis with the full set of 828 samples again demonstrated an increase number of clusters as the sample size increased with a plateau at approximately 18 clusters. In contrast, the analysis of the ER positive samples reached a plateau at 12 clusters, consistent with a reduction in the biological complexity of this subset.

Patterns of Pathway Activity Characterize the Diversity of Breast Cancer. Although patterns of gene expression can provide a basis to characterize the diversity of breast cancer, it is limited by the inability to interpret the underlying biological significance of these clusters. We previously described an alternative strategy to assess patterns of gene expression while concurrently providing biological insight through the use of expression signatures of pathway activation (4, 5). We have now extended this initial work by developing a large collection of pathway signatures (*SI Appendix*) that were used to predict the probability of pathway activity for each sample within the normalized breast tumor dataset (table S1 at http://data.duke.genome.edu/breast_subgroups). Because each signature has been validated with independent biochemical or genetic analyses (*SI Appendix* and table S11 at http://data.duke.genome.edu/breast_subgroups.), the predicted probability of pathway activity can be considered a correlative measure of in vivo pathway activity; high predicted pathway activity correlates with high in vivo pathway activity, whereas a low predicted probability of pathway activity correlates with low levels of in vivo activity. Therefore, the predicted pathway status based on these signatures provides a measure of pathway function based on a common assay (gene expression). This strategy enables the integration of measurements to reveal patterns of pathway dysregulation not possible when using data from disparate forms of pathway analysis. Thus, similar to the use of gene expression data for clustering based on probe-level hybridization intensities, hierarchical clustering of the predicted probabilities of pathway activation reveals distinct patterns of pathway deregulation (Fig. 2*A*).

In addition to clusters of samples, patterns of pathway coregulation can be identified from two-way hierarchical clustering providing insight into the nature of pathway associations across the spectrum of the disease. Two-way hierarchical clustering was first used to identify clusters of pathway that are statistically coactivated (Fig. 2*A*) and Pearson Correlation (Fig. 2*B* and table S2 at http://data.duke.genome.edu/breast_subgroups) was used to validate the statistical correlation between clustered pathways. These analyses identified a clear relationship between the ER, PR, and p53 pathways as expected from past studies and likewise for IFNα and IFNγ (10). Additionally, MYC and RAS exhibit a strong coactivation across the breast tumor datasets consistent with previous studies suggesting a genetic relationship between MYC and RAS in oncogenesis (11, 12). Other relationships, not necessarily anticipated from past work, are also evident in this analysis including coactivation of E2F1 and PI3K pathways together with β-catenin. Interestingly, E2F1 is known to act as a signal for p53-dependent apoptosis, which is negated by PI3K activity (13). Finally, other patterns are evident including AKT/p63/SRC as well as EGFR/TGFβ and STAT3/TNFα.

Identification of Breast Tumor Subgroups Based on Predictive Models of Pathway Activity. Although hierarchical clustering reveals structure in the data that can form a basis for classification, this method is largely descriptive. To serve as a framework for future studies, it is

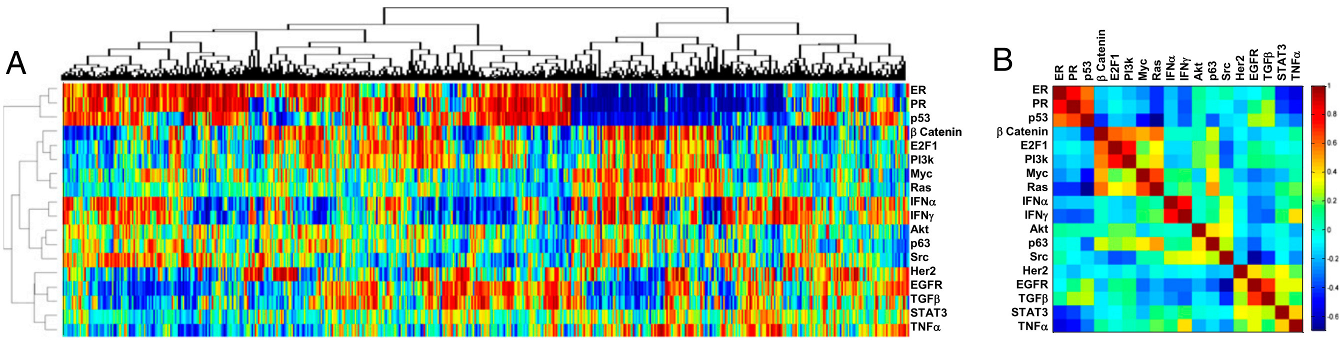

Fig. 2. **Patterns of pathway activity that characterize breast cancer.** (*A*) Heat map depicting the two-way hierarchical clustering of the predicted probability of 1,143 breast tumor samples and 18 pathways. Low (blue) and high (red) pathway activity and predicted probabilities are shown. (*B*) Heat map depicting the correlation coefficient of pathway coregulation (red indicates a positive correlation; blue, a negative correlation).

essential that classifications are based on predictive models. To address this challenge, we developed a tumor classification strategy that utilizes an initial affinity propagation scheme together with mixture modeling to define breast tumor subtypes on the basis of patterns of pathway activity (Fig. 3*A*). The mixture model can then be used to assign new samples to subgroups based on the relative likelihood of each of the mixture components. From this analysis, 17 subgroups were identified in which closely related samples, as measured by Euclidean distance (*SI Appendix*), could be optimally assigned (Fig. 3*B* and table S3 at http://data.duke.genome.edu/breast_subgroups) based on patterns of pathway activity.

Previous work has delineated a series of breast cancer subtypes based on gene expression patterns (2, 3). In light of this established framework for understanding breast cancer heterogeneity, we have now evaluated the pathway-defined subgroups in relation to these previously identified intrinsic subtypes of breast cancer, making use of the dataset of 1,143 tumors (Fig. 3*C*). Several conclusions can be drawn from this analysis. First, a clear relationship exists between pathway-defined subgroups and the intrinsic subtypes including basal (subgroups 2, 5, and 8), luminal A (subgroups 11 and 17), luminal B (subgroups 3, 4, 6, 9, and 16), and Erbb2 (subgroups 7 and 10) subtypes (*SI Appendix*). Secondly, it is evident that overall, the previous defined intrinsic subtypes exhibit distinct patterns of pathway activity. For example, the basal subgroups (2, 5, 8) exhibit low ER and PR activity and elevated Myc and Ras activity, whereas the luminal subgroups 1, 3, 4, 6, 9, 11, 16, and 17 generally exhibit an inverse pattern for these pathways. Thirdly, the pathway patterns also provide a basis for further subdivision of the intrinsic subtypes. For the basal-like tumors, subgroups 2 and 5 have low EGFR activity, whereas subgroup 8 has high EGFR expression. The inverse is true for SRC activity where subgroup 8 has low activity while subgroups 2 and 5 are high. Similar observations explain the division of luminal B tumors between several subgroups based on EGFR, β-catenin, and IFN activity. Finally, it is also evident from this analysis that pathway-defined subgroups can be composed primarily of a single intrinsic subtype or include multiple subtypes. For instance, subgroups 1, 12, 13, and 15 contain a mixture of luminal A and B tumors, suggesting both common and unique aspects of luminal tumors.

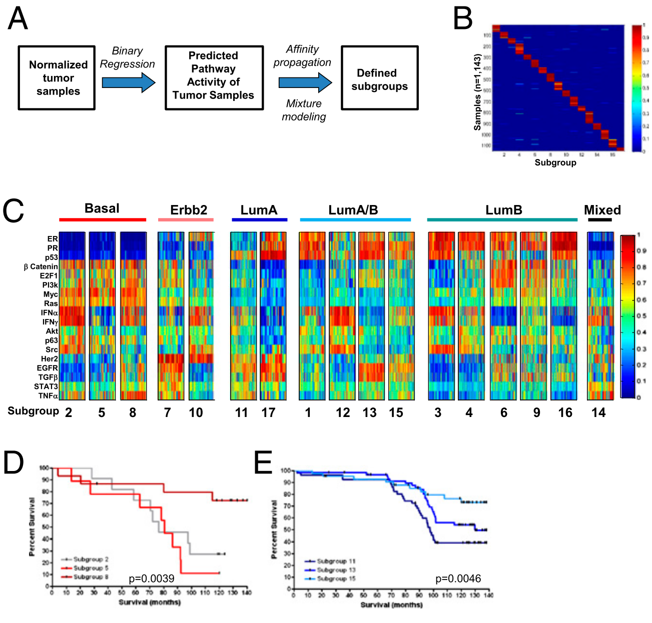

Fig. 3. **Identification of breast tumor subtypes using patterns of pathway activity.** (*A*) Scheme for the development of pathway-derived breast tumor subgroups. (*B*) Predicted probability of subgroup membership for 1,143 breast tumor samples where each row represents a sample; each column, a subgroup (samples are organized by subgroup). (*C*) Heat map depicting patterns of pathway activity in the 17 identified breast tumor subgroups organized by the relationship with intrinsic subtypes. Red indicates a high predicted probability, blue a low probability. Overall survival differences between pathway-derived subgroups classified as (*D*) basal-like ($P = 0.0039$) and (*E*) luminal A-dominant ($P = 0.0046$) were analyzed by a Kaplan-Meier survival curve and demonstrate a statistically significant difference in survival (log-rank test).

The biological significance of further resolving tumor subtypes is evident by an examination of the Kaplan-Meier analyses where survival differences exist in the basal and luminal A pathway-derived subgroups, despite previous studies reporting that basal-like tumors generally have a poor prognosis whereas luminal A tumors have a favorable prognosis (2, 14, 15). Overall survival was examined in the three subgroups with the greatest percentage of basal and luminal A samples, respectively, for which a sufficient number of samples had reported survival data (*SI Appendix*). Within the basal-like subgroups, a statistically significant difference ($P = 0.0039$, log-rank test) exists in overall survival between subgroups 8 (median survival >130 months) and 5 (median survival: 80.6 months) (Fig. 3*D*). Likewise, a statistically significant difference ($P = 0.0046$, log-rank test) in overall survival exists between luminal A-dominant subgroups 15 and 11 (median survival >140 months and 97.6 months, respectively) (Fig. 3*E*).

A Predictive Framework for Breast Cancer Classification.

In order for a classification strategy to be effective in future studies, it is necessary that the described classification scheme represents a predictive framework by which new tumor samples can be quantitatively assigned to a subgroup based on patterns of pathway activation. To evaluate the extent to which such classifications are indeed robust, an independent breast cancer dataset ($n = 547$) was investigated. Based on the pattern of predicted pathway activity (See table S5 at http://data.duke genome.edu/breast_subgroups), each sample was assigned to one of the 17 subgroups (Fig. 4*A* and table S6 at http://data.duke.genome.edu/breast_subgroups). The clinical properties of samples assigned to each subgroup in both the original and validation datasets were found to be highly coincident. For instance, tumors assigned to subgroups 2, 5, and 8 are basal-like (*SI Appendix* and table S4 at http://data. duke.genome.edu/breast_subgroups).

Because the established framework enables the classification of new samples, it also provides a mechanism to classify cancer cell lines as experimental models of a given subgroup. Fifty breast cancer cell lines (8) were assigned to subgroups (Fig. 4*B*) on the basis of pathway predictions (tables S7 and S8 at http://data.duke.genome.edu/breast_subgroups). In total, 12 of the 17 tumor subgroups were represented by cancer cell lines that exhibited a predicted probability greater than 0.80, and the assignment of a cell line to a subgroup also coincided with the intrinsic subtype identity of the cell line, whether basal or luminal (*SI Appendix*). Although several cell lines in this dataset (14/50) did not significantly (>0.80) associate with a single subgroup, the majority (9/14) of these cell lines have a probability of membership in multiple subgroups that are highly correlative by Euclidean distance (*SI Appendix*) and are comprised of tumors with a similar intrinsic subtype. These data suggest that these cells may have diverged from their original state during the course of establishing the cell line or during the subsequent years of growth in culture. Nevertheless, because the majority (72%) of the breast cancer cell lines examined in the current study can be assigned to a single subgroup with a high probability, our analyses suggest these particular breast cancer cell lines may serve as a good model system for the in vitro and in vivo studies of each subgroup; the remaining cell lines may be good model systems for a given pathway but do not represent a specific subgroup.

Pathway-Defined Breast Tumor Subtypes Exhibit Unique Patterns of DNA Copy Number Changes.

Previous reports of cancer genome sequencing efforts, including in breast cancer, reveal a very large number of gene mutations that fall into two general categories—frequent mutations seen in the majority of samples (sometimes

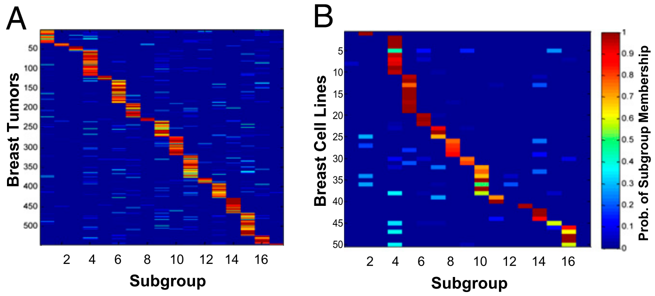

Fig. 4. **Prediction of subgroup membership.** (*A*) Breast tumors in the validation dataset (*n* = 547) were classified into 17 pathway-derived subgroups and the probability of subgroup assignment was plotted for each sample (red indicates a high probability of subtype membership; blue, low probability). (*B*). 50 breast cancer cell lines were classified into 13 of 17 pathway-derived subgroups on the basis of patterns of pathway activity, and the predicted probability of subgroup membership is shown; 36 of 50 (72%) samples had a predicted probability of subgroup membership greater than 0.80.

referred to as gene mountains in a cancer genome landscape) and gene mutations that are seen infrequently across the population of tumors (referred to as gene hills) (16). Although such analyses provide a starting description of the mutation landscape of breast cancer, they also present a challenge in understanding the significance of the infrequent mutations. Given the evidence that breast cancer is in fact multiple distinct disease entities, it is entirely possible that the so-called gene hills become gene mountains in the context of a defined subgroup; however, current classification schemes are unable to generate sufficiently homogeneous classes of tumors to identify these changes. Therefore, a primary goal of the classification strategy we describe here is to identify subgroups that exhibit common molecular mechanisms of disease that can then serve as a framework by which to investigate relevant genetic alterations in a homogeneous population of tumors.

To investigate whether pathway-derived subgroups are characterized by common genomic alterations, a collection of breast tumors for which combined expression data and copy number variation (CNV) data were available (9) were assigned to subgroups and chromosomal abnormalities associated with each subgroup were analyzed by comparative genomic hybridization analysis. Consistent with previous studies (8, 9, 17), various chromosomal regions that demonstrate CNV could be identified when all samples were analyzed as a group (Fig. 5). However, when these tumors were assigned to the pathway-derived subgroups, increasingly homogeneous patterns of CNV became evident in chromosomal regions that showed little CNV when viewed in the context of all breast tumors. For example, 75% of subgroup 5 tumors exhibit uniform losses at 3p14.3 (Fig. 5*B*) but only 18% of all other tumors were characterized by this change ($P = 0.0009$). Further, 100% of patients in subgroup 7 have losses at 4p15.1, and 60–80% of patients in subgroup 6 have losses at 11q21-24 (Fig. 5*B*), whereas only 8% ($P = 0.0106$) and 9–23% ($P = 0.0093$) of all other tumors have copy number losses at these chromosomal positions.

Similar results are seen with copy number gains that occur at a relatively low frequency in total breast cancer (Fig. 5*C*). For instance, copy number gains at 3q25.1 are present in 75% of subgroup 5 patients but only 11% of all other patients ($P = 0.0211$). Similarly 60–80% of patients in subgroup 11 also show copy number gains at 20p12-13, whereas only 6–14% of all other tumors show copy number gains at these chromosomal bands ($P < 0.0001$).

As detailed in previous sections, one important aspect of the use of the pathway signatures is the capacity to reveal further complexity in the previously-defined intrinsic subtypes. As seen in Fig. 5*D*, this subdivision also revealed distinct CNV patterns. For instance, whereas the basal subgroups 2, 5, and 8 all exhibited copy number gains at chromosome 8q24 ($P = 0.4575$, ANOVA), only the subgroup 5 tumors exhibited deletion of 3p14 ($P < 0.0001$, ANOVA).

Although these analyses are limited by the small numbers of samples for which both expression data and copy number data are available, it is nevertheless clear from the results in Fig. 5 that the ability to identify homogeneous subgroups of disease, based on patterns of pathway activity that reflect underlying biology, does provide an opportunity to reveal chromosomal alterations that might be overlooked by an analysis of the total population of tumors. As such, we believe this provides a framework for future studies that will attempt to identify the fully complexity of genome alterations, including DNA sequence changes, that characterize not just breast cancer but the particular subgroup of breast cancer.

Discussion

Previous studies have detailed the analysis of genome-scale gene expression data to characterize tumor heterogeneity including the identification of tumor subtypes not recognized by other methods (2, 14, 18–21). Current clinical-based classification relies on parameters

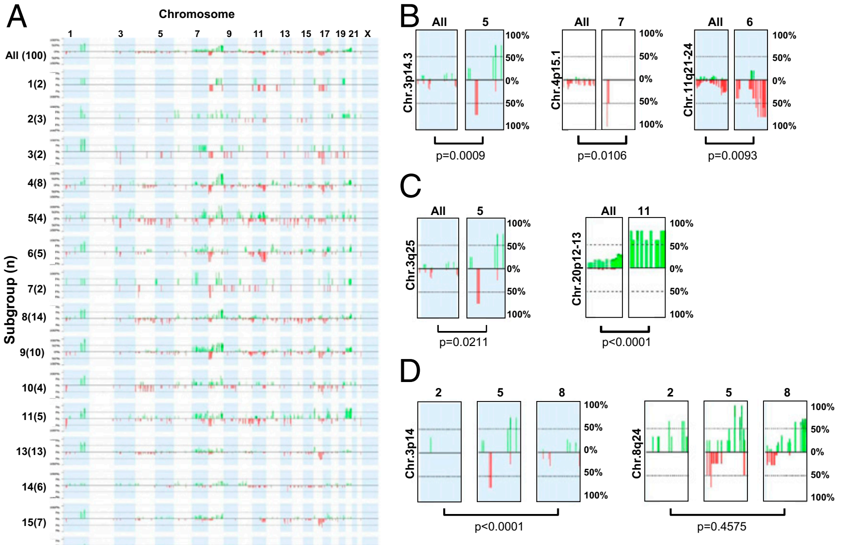

Fig. 5. **Pathway-defined breast tumor subgroups exhibit unique patterns of DNA copy number changes.** Patterns of DNA copy number changes were calculated for each subgroup. (*A*) The percent of samples in each of the 16 subgroups with identified copy number gains and losses are shown. Green indicates a region of amplification and red indicates a region of chromosomal loss; dark green and red indicate the percentage of samples with high copy number gains or homozygous deletion, respectively. Chromosomal borders are delineated by alternating gray and white regions. (*B*) Increasingly homogeneous patterns of copy number losses are evident in pathway-derived subgroups as compared to all breast tumors for subgroup 5 at 3p14.3 (P = 0.0009, unpaired *t* test), subgroup 7 at 4p15.1 (P = 0.0106, unpaired *t* test), and subgroup 6 at 11q21-24 (P = 0.0093, unpaired *t* test). (*C*) Increasingly homogeneous patterns of copy number gains are evident in breast tumor subgroups compared to all other samples. Subgroup 5 shows a amplification at 3q25.1 (P = 0.0211, unpaired *t* test) and subgroup 11 shows an amplification at 20p12-13 (P < 0.0001, unpaired *t* test). (*D*) Basal-like subgroups 2, 5, and 8 show copy number gains at 8q24 (P = 0.4575, ANOVA); only subgroup 5 shows copy number losses at 3p14 (P < 0.0001, ANOVA).

that include visual characteristics, tumor size, and a limited number of histochemical markers. Given that each of these phenotypic characteristics are the result of the expression of a unique complement of genes, the use of genome-scale gene expression analysis differs only in the scale of the data and the consequent ability to add greater precision to these determinations. Studies investigating genomic abnormalities and epigenetic modifications add complexity and detail to the description of cancer phenotypes (22–24). Although these data are important in describing cancer characteristics, it is critical to develop a unifying underlying platform that can accommodate complex data while concurrently reducing the complexity to a form that provides biological insight. We suggest that using experimentally-derived pathway signatures provides an approach to this challenge by organizing the inherent complexity in a form that offers a predictable framework linked to defined biology.

The use of pathway signatures as the basis for tumor classification adds additional value to basic gene expression analysis. First, the pathway signatures provide an immediate biological interpretation because these signatures are based on experimental determinations of pathway activity. Second, given the link between predicted pathway activity and sensitivity to pathway-specific therapeutics, this information can form the basis for the rational design of therapeutic regimens for subgroups of patients. Finally, the relative few pathways, as compared to the total number of gene probes on an array, allows for more rigorous modeling of the observed patterns of pathway activity and enables prediction of subgroup membership.

The inherent heterogeneity in human cancer presents an enormous challenge for predicting therapeutic response and for understanding mechanisms of disease. This challenge is best illustrated by attempts to interpret DNA sequencing information from cancer genome projects where large numbers of mutations are identified but without a capacity to clearly associate which sequence variations are relevant for a cancer phenotype (16, 22, 25). Given that this complexity translates into many forms of breast cancer, the ability to define distinct phenotypes that distinguish heterogeneity and identify tumors with common properties will be critical to the interpretation of DNA sequencing data and the ability to define molecular mechanisms responsible for various classes of breast cancer.

We propose that the classification scheme we describe here, that extends previous work defining breast cancer subtypes by now making use of predictive models generated from pathway probabilities, can provide a framework for future studies that evaluate aspects of breast cancer biology. The ability to assign a new sample to a particular subgroup affords the opportunity to build on and extend the current body of knowledge. Further, we suggest this also provides a framework to relate experimental systems such as cancer cell lines, xenografts, and genetic models, to allow a characterization of the breast cancer subgroups. We also do recognize limitations in this approach as indicated by some uncertainty in the subgroup assignment of tumors from the validation cohort as well as the fact that a fraction of breast cancer cell lines failed to fit within a given subgroup. Although this may reflect divergence in the cell lines as they have been in culture, it is also possible that the subgroup framework is limited by the available pathway signatures and not fully representative of the full scale of breast cancer variation. Nevertheless, given the fact that the majority of tumors or cell lines were assigned with high probability to a unique subgroup, we do believe this represents a foundation for further development.

Identifying breast cancer subgroups on the basis of homogeneous patterns of pathway activation also provides a framework to evaluate and interpret the complex alterations that characterize cancer genomes because deregulation of these pathways reflect the genetic alterations unique to each tumor. Indeed, our initial analyses of DNA copy number alterations support this conclusion as a representation of one such mechanism of oncogenesis. Finally, we suggest that the described pathway signatures provide an opportunity for identifying populations of patients that may benefit from a particular agent by linking a given drug with a pathway based on knowledge of the drug target. We have shown this connection in many in vitro examples and recent data indicate that signatures derived from an initial EGFR pathway signature can be effective in identifying patients that are responsive to cetuximab (26).

Although the ability to develop predictive tools for targeted therapeutic agents is important, we believe the power of this approach is the potential to rationally identify drug combinations that can be matched with specific subpopulations of patients based on underlying biological properties. The importance of a rational strategy for combination therapy is highlighted by the limited clinical benefit of single agents; drugs such as cetuximab, erlotinib, avastin, and others achieve approval based on clinical activity that extends overall patient survival by a small margin (27–31). Although this activity is important and does represent a true measure of therapeutic benefit, more must be done to translate this activity into clinical success. One likely basis for the limited clinical benefit is the fact that any single agent, even when combined with a cytotoxic agent or regimen, fails to match the complexity of the tumor. It is reasonable to propose that a therapeutic strategy using multiple drugs, each of which alone demonstrated a small but real clinical benefit in an individual patient, might have a significant and lasting therapeutic benefit when used in combination. It is currently impossible, however, to predict the clinical benefit of novel drug combinations in a specific patient. These questions will only be answered when investigated in a clinical setting, where the therapeutic efficacy of novel drug combinations can be examined in subpopulations of patients with a common disease mechanism. Therefore, the proposed pathway-based classification strategy provides a concrete framework to define potential rational combination regimens that can be tested in clinical studies.

Materials and Methods

Human Breast Tumor Samples and Cancer Cell Lines. A total of 1,143 patient samples from 10 independent datasets (GSE1456, GSE1561, GSE2034, GSE3494, GSE3744, GSE4922, GSE5460, GSE5764, GSE6596, and E-TABM-158) were analyzed (9, 32–40). The validation dataset (n = 547) was derived from two independent datasets (41, 42). Fifty breast cancer cell lines (E-TABM-157) (8) were analyzed.

Pathway Signature Training Data. The training data used to generate the 18 pathway signatures developed are described in *SI Appendix* and signature conditions are detailed in tables S9 and S10 at http://data.duke.genome.edu/breast_subgroups.

Processing of Microarray Data. Microarray data were normalized by RMA or MAS5.0 algorithms using Affymetrix Expression Console Software Version 1.0. All data were filtered to include those probes on the U133A platform. BFRM (Bayesian Factor Regression Modeling) (43, 44) was used to eliminate technical differences between breast tumor samples in multiple datasets by normalizing the data against 69 Affymetrix probes for human maintenance genes using 15 principal components. These methods are described in *SI Appendix*.

Analysis of Expression Data for Predicting Pathway Activity. The statistical methods used here to develop gene expression signatures of pathway activity have been previously described (5) and are described in detail in *SI Appendix*.

Validation of Pathway Signature Accuracy. To validate pathway signatures two types of analyses were performed (*SI Appendix*). First, a leave-one-out cross validation was used to formally confirm the validity and robustness of each signature to distinguish between the two phenotypic states. Secondly, genetic and biochemical analyses were used to validate the correlation between predicted probability of pathway activity and measured in vivo pathway activity.

Analysis of Patterns of Pathway Activity. Two-way hierarchical clustering (Complete Linkage) was performed using Cluster3.0 to analyze patterns of pathway coactivity based on the predicted probabilities of pathway activity for each sample in the breast tumor dataset. To validate the correlation between clustered pathways, a Pearson correlation was performed; r- and P values are reported in *SI Appendix*.

Generation of Breast Cancer Subgroups. Full details on the statistical modeling used to define each subtype are available in *SI Appendix*. Briefly, the predicted activity of 18 cellular pathways was determined for 1,143 breast tumor samples.

Preliminary subgroup characteristics were defined by affinity propagation using a Euclidean distance similarity function based on the pathway predictions. The affinity propagation tunable parameter was set to the default setting of −33. Mixture modeling of pathway predictions was then used to further refine each subgroup. Finally, a log likelihood test was used to validate the identified subgroups. The probability of an independent sample being assigned to each identified subtype is calculated by the relative likelihood that the new sample belongs to each of the components of the mixture model.

Analysis of Breast Cancer Subtypes. Intrinsic subtype membership was determined using previously described methods (7). Briefly, the U133A probe set was filtered to include 684 probes (360 genes), which correlate to the intrinsic gene list (14). The top 66% (451 probes) of variable probes were then used to cluster the BFRM normalized Mas5 formatted gene expression data using complete linkage hierarchical clustering. Previously identified intrinsic subtypes were identified and the expression characteristics of each subtype were found to be consistent with previously published studies (4, 13, 6). Specifically, the HER2+ expression cluster was found to show high expression of genes in the 17q21 amplicon including HER2/ERBB2 and GRB7. The basal expression cluster was found to express KRT5 and KRT17 and have low ESR1 expression. The Luminal A and B clusters were characterized by high expression of ESR1 and GATA3 and the Luminal A cluster was distinguished by high ADH1B expression (*SI Appendix*).

Unsupervised Hierarchical Clustering. Affymetrix U133A expression data for the 1,143 breast tumor samples was MAS5 normalized and probes and samples were mean centered and clustered by complete linkage using Cluster 3.0. In triplicate, 25, 50, 100, 200, 400, 600, 800, and 1,000 random samples were selected and the resulting number of clusters at a given level in a dendrogram (*SI Appendix*).

Comparative Genome Hybridization (CGH) Analysis. Array CGH data (E-TABM-158) was processed as previously described (table S12 at http://data.duke.genome.edu/breast_subgroups) (9). Tumor samples were classified into breast tumor subtypes on the basis of gene expression patterns. DNA copy number alterations associated with each subgroup were determined using NEXUS Copy Number 4.0 (BioDiscovery, Inc.) that relies on a Rank Segmentation algorithm, similar to the Circular Binary Segmentation (CBS) algorithm to segment the genome and position probes (45). The significance threshold used to identify chromosomal regions of copy number variation was set to 0% to identify all regions of variation. The identified regions of copy number variation for each subgroup are reported in table S13 at http://data.duke.genome.edu/ breast_subgroup. Each bar in Fig. 5 represents the percentage of samples in a subgroup with copy number variation at a segmented chromosomal band. To validate that copy number changes were statistically significant, probe intensities for BACs in each identified chromosomal band were averaged (table S14 at http://data.duke.genome.edu/ breast_subgroups) and either an unpaired *t* test or one-way ANOVA were used to compare probe intensities between subgroups.

ACKNOWLEDGMENTS. We thank members of our laboratory, specifically Jeffrey Chang, Erich Huang, and Jason Reeves, as well as Simon Gregory and Aaron Towers for helpful discussion. We are grateful to Kaye Culler for her assistance in the preparation of the manuscript. All aspects of the research were supported under the National Cancer Institute Integrative Cancer Biology Program via grant National Institutes of Health 5-U54-CA112952-05 and 5-RO1-CA106520-05 to J.R.N. M.L.G. is supported by National Institutes of Health HL007101-32 and by postdoctoral fellowship National Institutes of Health CA139890-01.

1. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. *Cell* 100:57–70.
2. Perou CM, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406: 747–752.
3. Sorlie T, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 100:8418–8423.
4. Huang E, et al. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet* 34:226–230.
5. Bild AH, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353–357.
6. Bild AH, et al. (2009) An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. *Breast Cancer Res* 11:R55.
7. Smid M, et al. (2008) Subtypes of breast cancer show preferential site of relapse. *Cancer Res* 68:3108–3114.
8. Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527.
9. Chin K, et al. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 10:529–541.
10. Nielsen TO, et al. (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 10:5367–5374.
11. D'Cruz CM, et al. (2001) c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations. *Nat Med* 7:235–239.
12. Sinn E, et al. (1987) Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: Synergistic action of oncogenes in vivo. *Cell* 49:465–475.
13. Hallstrom TC, Mori S, Nevins JR (2008) An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer Cell* 13:11–22.
14. Sørlie T, et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 98:10869–10874.
15. Sotiriou C, et al. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA* 100:10393–10398.
16. Wood LD, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
17. Bergamaschi A, et al. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* 45:1033–1040.
18. Golub TR, et al. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.
19. Alizadeh AA, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.
20. van de Vijver MJ, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999–2009.
21. West M, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98:11462–11467.
22. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812.
23. Leary RJ, et al. (2008) Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci USA* 105:16224–16229.
24. Jones S, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321:1801–1806.
25. Sjöblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
26. Chang JT, et al. (2009) A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol Cell* 34:104–114.
27. Sandler A, et al. (2006) Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med* 355:2542–2550.
28. Van Cutsem E, et al. (2009) Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N Engl J Med* 360:1408–1417.
29. Vermorken JB, et al. (2007) Open-label, uncontrolled, multicenter phase II study to evaluate the efficacy and toxicity of cetuximab as a single agent in patients with recurrent and/or metastatic squamous cell carcinoma of the head and neck who failed to respond to platinum-based therapy. *J Clin Oncol* 25:2171–2177.
30. Shepherd FA, et al.; National Cancer Institute of Canada Clinical Trials Group (2005) Erlotinib in previously treated non-small-cell lung cancer. *N Engl J Med* 353:123–132.
31. Tsao MS, et al. (2005) Erlotinib in lung cancer—molecular and clinical predictors of outcome. *N Engl J Med* 353:133–144.
32. Turashvili G, et al. (2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* 7:55.
33. Carroll JS, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* 38:1289–1297.
34. Farmer P, et al. (2005) Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 24:4660–4671.
35. Ivshina AV, et al. (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* 66:10292–10301.
36. Miller LD, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* 102:13550–13555.
37. Pawitan Y, et al. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Res* 7:R953–R964.
38. Richardson AL, et al. (2006) X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* 9:121–132.
39. Wang Y, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365:671–679.
40. Klein A, et al. (2007) Comparison of gene expression data from human and mouse breast cancers: Identification of a conserved breast tumor gene set. *Int J Cancer* 121:683–688.
41. Loi S, et al. (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 25:1239–1246.
42. Rouzier R, et al. (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* 11:5678–5685.
43. Carvalho C, et al. (2008) High-dimensional sparse factor modelling: Applications in gene expression genomics. *J Am Stat Assoc* 103:1438–1456.
44. Lucas JE, Carvalho C, West M (2009) A Bayesian analysis strategy for cross-study translation of gene expression biomarkers. *Stat Appl Genet Mol Biol* 8:11.
45. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:5l57–5572.