1-11-2007

OPTIMIZED CROSS-STUDY ANALYSIS OF MICROARRAY-BASED PREDICTORS

Xiaogang Zhong
Department of Applied Mathematics and Statistics, Johns Hopkins University

Luigi Marchionni
Department of Oncology, Johns Hopkins University

Leslie Cope
Departments of Oncology and Biostatistics, Johns Hopkins University

Edwin S. Iversen
Institute of Statistics and Decision Sciences, Duke University

Elizabeth S. Garrett-Mayer
Departments of Oncology and Biostatistics, Johns Hopkins University

See next page for additional authors

Authors
Xiaogang Zhong, Luigi Marchionni, Leslie Cope, Edwin S. Iversen, Elizabeth S. Garrett-Mayer, Edward Gabrielson, and Giovanni Parmigiani

Optimized cross-study analysis of microarray-based predictors

Xiaogang Zhong[1], Luigi Marchionni[2] , Leslie Cope[2,3] , Edwin S. Iversen[4] , Elizabeth S. Garrett-Mayer[2,3] , Edward Gabrielson[2,5]and Giovanni Parmigiani[*2,3,5]

[1]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Oncology, Johns Hopkins University, Baltimore, MD, USA
[3]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA
[4]Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA
[5]Department of Pathology, Johns Hopkins University, Johns Hopkins University, Baltimore, MD, USA

Email: Xiaogang Zhong - zhong@ams.jhu.edu; Luigi Marchionni - marchion@jhu.edu; Leslie Cope - cope@jhu.edu; Edwin S. Iversen - iversen@stat.duke.edu; Elizabeth S. Garrett-Mayer - esg@jhu.edu; Edward Gabrielson - egabriel@jhmi.edu; Giovanni Parmigiani*- gp@jhu.edu;

*Corresponding author

Abstract

Background: Microarray-based gene expression analysis is widely used in cancer research to discover molecular signatures for cancer classification and prediction. In addition to numerous independent profiling projects, a number of investigators have analyzed multiple published data sets for purposes of cross-study validation. However, the diverse microarray platforms and technical approaches make direct comparisons across studies difficult, and without means to identify aberrant data patterns, less than optimal. To address this issue, we previously developed an integrative correlation approach to systematically address agreement of gene expression measurements across studies, providing a basis for cross-study validation analysis. Here we generalize this methodology to provide a metric for evaluating the overall efficacy of preprocessing and cross-referencing, and explore optimal combinations of filtering and cross-referencing strategies. We operate in the context of validating prognostic breast cancer gene expression signatures on data reported by three different groups, each using a different platform.

Results: To evaluate overall cross-platform reproducibility in the context of a specific prediction problem, we suggest integrative association, that is the the cross-study correlation of gene-specific measure of association with the phenotype predicted. Specifically, in this paper we use the correlation among the Cox proportional

hazard coefficients for association of gene expression to relapse free survival (RFS). Gene filtering by integrative correlation to select reproducible genes emerged as the key factor to increase the integrative association, while alternative methods of gene cross-referencing and gene filtering proved only to modestly improve the overall reproducibility. Patient selection was another major factor affecting the validation process. In particular, in one of the studies considered, gene expression association with RFS varied across subsets of patients that differ by their ascertainment criteria. One of the subsets proved to be highly consistent with other studies, while others showed significantly lower consistency. Third, as expected, use of cluster-specific mean expression profiles in the Cox model yielded more generalizable results than expression data from individual genes. Finally, by using our approach we were able to validate the association between the breast cancer molecular classes proposed by Sorlie $et\ al$ and RFS.

Conclusions: This paper provides a simple, practical and comprehensive technique for measuring consistency of molecular classification results across microarray platforms, without requiring subjective judgments about membership of samples in putative clusters. This methodology will be of value in consistently typing breast and other cancers across different studies and platforms in the future. Although the tumor subtypes considered here have been previously validated by their proponents, this is the first independent validation, and the first to include the Affymetrix platform.

Background

Microarrays have been extensively used in cancer research, and led to the identification of several gene expression signatures involved in various aspects of cancer pathogenesis. Individual studies have typically investigated relatively small numbers of samples, making cross-study validation a crucial step for the scientific community. Use of gene expression data from public repositories has proved difficult due to inherent differences in microarray platforms, protocols used in independent laboratories, experimental designs, and annotations for both genes and samples. Several methodologies have been proposed to address these issues, that depend on the experimental strategies and on the biological and clinical questions. When samples phenotypes are known, statistical methods which handle data sets separately and then apply gene-wise meta-analytic approaches, have proven successful, allowing the identification of the statistically relevant intersections of molecular signatures from different studies [1–4]. As an alternative, the

assimilation of gene expression measurements, achieved by merging the datasets has also been used to evaluate molecular signatures obtained from different studies [5–8]. Finally, we previously developed a method to evaluate cross-platform consistency of expression patterns, using integrative correlation (ICOR). This technique enables the quantification of cross-study reproducibility without relying on direct assimilation of expression data across the platforms considered [9, 10].

In this paper we systematically investigate the principal decisions involved in the comparison of studies conducted using different microarray platforms and evaluate their impact on the overall reproducibility across studies. We specifically consider gene cross-referencing, expression data processing, gene filtering and patient selection. To evaluate overall cross-platform reproducibility in the context of a specific prediction problem, we propose integrative association, that is the cross-study correlation of gene-specific measure of association with the phenotype predicted. We evaluate cross-study reproducibility both in terms of individual genes and in terms of profiles or clusters, by considering their centroids [8].

To demonstrate this strategy in a challenging application, we consider the case of predict survival in breast cancer patients. Microarrays have been extensively used in breast cancer research to identify gene expression-based predictors for survival and response to therapy, as well as for molecular classification. However, due to the costs involved with this type of analysis, the need for fresh frozen tumor specimens with associated clinical information, and other factors, only two genes expression predictors have reached a prospective clinical trial [11–13]. The analysis of multiple published data sets emerged as the main option to independently evaluate arrays study results. Here, we evaluate three breast cancer data sets from three distinct groups, using three different platforms: Sorlie et al. [14], van De Vijver et al. [12] and Huang et al. [15] (hereafter referred to as the "Sorlie", "VanDeVijver" and "Huang" studies). To implement integrative association, we compare the Cox coefficients for the association between gene expression and relapse free survival (RFS). We apply the ICOR approach to select the genes that are consistently correlated across the platforms (hereafter referred as the reproducible genes) to validate the breast cancer taxonomy proposed by Sorlie et al. [5, 11, 14], that defines the basal-like, ERBB2, luminal A, luminal B and normal-breast like molecular subtypes. The authors used hierarchical clustering of tumor samples based on a panel of 534 "intrinsic" genes to identify specific molecular signatures that identified groups of patients with different clinical outcomes [5, 11, 14]. Tumor subgroups were found in two studies [16, 17] and the association with survival was confirmed in an additional study by the same group [5].

The "intrinsic" molecular taxonomy has been recently revised [18], and an expanded gene list (hereafter referred to as the "new" panel) was obtained from the analysis of a combination of previously published data sets [14, 17, 19]. In the same publication the association of the breast cancer subtypes with different clinical outcomes was also confirmed by using two additional data sets [20, 21], and recently the same group successfully applied the "new" classifier also to the VanDeVijver data set [22]. An important aspect of the present study is an independent validation of this association by comparing the Cox coefficients for both "intrinsic" gene lists across the three studies. We consider the study originally used to develop the "intrinsic" gene set, as well as two additional data sets that were originally used for other purposes. The first (Huang et al. [15]) reported two gene expression signatures associated with the estrogen receptor (ER) and lymph node (LN) status in breast cancer patients, while the second (VanDeVijver et al. [12]) proposed a prognostic signature on a cohort of 295 breast cancer patients. This is the first study to include methods to rigorously test both individual genes and gene sets for reproducibility across data sets. Finally we are not aware of any published independent evaluation of the compatibility across studies that included the Huang data set.

Results and Discussion

Results

In this section we present an ICOR-based approach to cross-study analysis of the Sorlie, VanDeVijver and Huang datasets. We also address major issues related to cross-referencing and gene expression data processing, by evaluating their impact on concordance of estimates of association across studies. We evaluated these associations using expression data and RFS or relapse status as response variables. Overall survival was not used in the analysis, since it was not available for the Huang study.

We collected transcript profiles of 487 primary breast tumors as follows: 104 from the Sorlie study (cDNA microarray), 295 from the VanDeVijver study (custom Agilent oligo microarray) and 88 from the Huang study (Affymetrix hgu95av2 oligo microarrays [23]). Normal samples and benign tumors were excluded, and only cancer patients with complete clinical information were considered in the analysis. The Cox coefficients were separately computed for each data set, and the correlations in the three possible pairs of studies (Sorlie versus VanDeVijver, Sorlie versus Huang, and VanDeVijver versus Huang, hereafter respectively referred to as "SV", "SH" and "VH") were used to assess the degree of agreement.

Impact of alternative cross-referencing procedures to agreement among platforms

A cross-platform comparison of independent expression data sets requires cross-referencing annotated microarray features. This can be accomplished by either mapping all features from each platform to a common reference set of identifiers, or by direct comparison of sequence alignments. Both approaches can be gene or transcript oriented, depending on the cross-referencing identifiers used, or the BLAST [24] database considered for the alignments (i.e. Entrez Gene identifiers [25], RefSeq [26, 27], and so forth). We applied different mapping strategies and subsequently evaluated their effect on the integrative association across studies. We re-annotated each platform by mapping the original identifiers to Unigene clusters (UGC) [28], to Entrez Gene identifiers, and to gene symbols, using two web-based tools, MatchMiner and SOURCE [29, 30]. In addition, we applied a cross-referencing strategy based on direct BLAST alignments of the array sequences to the RefSeq transcripts. Table 1 summarizes annotations results obtained for all the common genes, the "intrinsic" gene panels described by Perou et al. Sorlie et al. and Hu et al. [5, 11, 14, 18], and for the 70-genes recurrence signature by van't Veer et al. [12, 17]. Overlaps among the sets obtained are shown in Figure 1.

The largest overlap across the three data sets was obtained by MatchMiner with UGC identifiers as the common cross-mapping reference. This approach allowed the identification of a total of 4125 common genes, containing 354 genes from the original "intrinsic" gene list, 382 from the "new intrinsic" gene panel and 22 genes from the 70-gene signature. In selecting the overlapping gene set, mappings of a single original to multiple common identifiers were not allowed, and, in the BLAST-based analysis, the Affymetrix probe sets that had conflicting individual probe matches were excluded. Unambiguous mapping to RefSeq transcripts by BLAST strongly reduced the total number of features included in the common set to 1016. This was mainly due to the two oligonucleotide platforms (Affymetrix and Agilent) for which, in many cases, probes were not able to discriminate between different transcripts of the same gene, and were discarded to avoid multiple matching (3351 RefSeq transcripts would have been included in the common set if multiple matching to RefSeq had been allowed).

We assessed the impact of the different annotation strategies by evaluating reproducibility —defined in terms of correlation of Cox coefficients across studies— by comparing the largest set obtained (UGC by MatchMiner) with the smallest one (RefSeq by BLAST). The following correlations were obtained using the UGC/MatchMiner mappings: SV = 0.115, SH = -0.011 and VH = 0.038; while the results for the

BLAST alignments were as follows: SV = 0.132, SH = -0.036 and VH = 0.083 (see Table 2 for details). Overall the agreement among the three studies, irrespective to the mapping strategy used, appeared poor, and additional data processing was clearly required to increase concordance.

Impact of data standardization and filtering to agreement among platforms

Because the low integrative association may depend on gene expression discrepancies due to platform differences or to the protocols originally used in each study, we evaluated the effect of alternative data handling and filtering procedures on the agreement across the platforms. First, we checked whether the standardization of gene expression data before fitting the Cox models would increase the overall agreement. As expected, this approach proved useful, since it corrected for the different scales of measurement represented by the gene expression data used as predictors. After standardization, the following Cox coefficient correlations were obtained for all the genes in the UCG/MatchMiner set: SV = 0.165, SH = 0.005 and VH = 0.127; while in the case of BLAST alignments the results were as follows: SV = 0.178, SH = 0.001 and VH = 0.168 (see Table 2 for details). Although standardization increased the overall agreement across studies, the results were still not satisfying, and we considered gene filtering as an additional step to increase reproducibility. We decided to perform this type of analysis only on the largest common set available (UGC by MatchMiner), so that enough genes would be available to fit the Cox model after filtering. We therefore filtered data by gene variance to discard uninformative genes, and we used the ICOR method [9] to select the reproducible genes. The two approaches do not exclude each another and are described below.

Selection of reproducible genes was accomplished by comparing the observed and the null distributions of the integrative correlations, and by applying several cutoffs corresponding to various false discovery rates (FDR) [31]. In particular, 3359 UGC genes were deemed reproducible at FDR = 0.1. Of these 2771 UGC genes were retained at an FDR = 0.01. As expected, when only the reproducible genes were considered, integrative association increased (see Table 2). However, agreement increased in the SV and VH comparisons (SV = 0.202 and VH = 0.197), while the SH correlation remained low (SH = -0.002, see Table 2). Alternatively, gene filtering by ICOR could also use fixed cutoffs, rather than FDR (see below and Table 2).

Additionally, we filtered genes with low variance. This was performed by discarding the same proportion of

genes in each study, by applying a threshold corresponding to the 30th percentile of the variance distribution in each data set. This cutoff was determined empirically, in order to balance the increase of reproducibility with the loss of genes. This approach was combined with the ICOR based gene selection and resulted in an additional gain in overall agreement between studies. However, the reproducibility observed in the comparisons involving the Huang study was worse then was observed for the SV pair (SH = 0.013, VH = 0.2 and SV = 0.228, see Table 2).

Investigation of the Huang data set and impact of sample selection on agreement among studies

Since pairwise comparisons involving the Huang data set consistently showed lower reproducibility than the SV comparison, we investigated whether this was due to features associated specifically with this study or the Affymetrix platform in general. Evaluation of the available information, including headers of the CEL files corresponding to the raw data, indicated that there were three major hybridization batches, based on the experiment dates and the chip serial number. We thus evaluated these apparent batches for cross-batch reproducibility, to investigate potential artifacts. We also explored whether batches included patients with different clinical phenotypes, and consequently whether consistency with other studies was batch-specific.

We first evaluated gene expression data correlations for every pair of samples in the Huang study, using all the genes on the chip. A heatmap of the pair-wise correlation matrix suggested that all the hybridizations were fairly homogeneous and comparable across the three batches; pairwise correlations ranged from 0.749 to 0.976 (see Supplementary Materials for details). We furthermore calculated the observed and the null distributions of the ICORs for each pair-wise comparison between batches. Their density plots confirmed that the three batches were highly consistent with each other. We therefore concluded that the three batches were similar in terms of quality of gene expression measurements.

We subsequently investigated the Cox coefficient correlations across the three batches after standardization of expression. These were 0.167 for the combination of batch 1 and batch 3 versus batch 2, and -0.151 for batch 1 versus batch 2. We did not compare batch 2 against batch 3 since there were no relapsing patient in batch 3. All specimens in the third batch corresponded to patients who were LN positive and who showed longer RFS time and no recurrence of the disease. Collectively these analysis suggested that the three groups of patients were phenotypically distinct. Additional details on the distributions of clinical characteristics across the batches are provided in the Supplementary Materials.

For this reason, we decided to keep the three batches separate and explored whether they were different in terms of technical consistency with the other two studies. We examined the Cox coefficients correlations between each batch and the other studies. The second batch was found to have highest correlation. In particular, when all the genes were used, the comparison with the Sorlie study (hereafter referred to as "SH2") showed a Cox coefficients correlation of 0.152, while the comparison with VanDeVijver study (hereafter referred to as "VH2") showed an overall agreement of 0.3. We also re-evaluated the overall agreement after selection of the reproducible genes (as obtained by applying the false discovery rate approach previously described), and as expected, the correlation of the Cox coefficients substantially increased: SV = 0.202, SH2 = 0.204, and VH2 = 0.425 with FDR = 0.1; SV = 0.355, SH2 = 0.437, and VH2 = 0.616 with an ICOR score bigger than 0.25; (see Table 2).

Intrinsic genes signatures validation

We extended our analysis to comparing the agreement of the studies using the prognostic signatures by Sorlie and colleagues [5, 14], since previously published cross-study comparisons have not assessed the performance of the intrinsic genes classification on the Affymetrix platform. Of the 534 intrinsic genes, 93 were formally assigned in the original paper [5] to the different clusters associated with the tumor types; 56 genes from this set were present in our UGC common set and were used in this investigation. When this subset of genes was considered, the Cox coefficient correlation consistently increased across all comparisons (SV = 0.639, SH2 = 0.686, VH2 = 0.623; see Table 2). As before, the use of ICOR to select the reproducible genes allowed to further increase the agreement across the studies (SV = 0.763, SH2 = 0.716, and VH2 = 0.698 with ICOR threshold of 0.25; see Table 2 and Figure 4).

We then investigated the performance of the "centroid gene" for each "intrinsic" cluster, calculated as the mean expression profile for the genes within each group. Similarly to individual genes, centroids were used as predictors, and their Cox coefficients for RFS were compared across studies by calculating the correlation of these coefficients. Overall, centroids proved more powerful and stable than individual genes, with higher Cox coefficients correlations (SV = 0.851, SH2 = 0.977, and VH2 = 0.93; see Table 2 and Figure 4). In this case as well ICOR-based genes filtering increased reproducibility across studies. In all pair-wise comparisons across studies the centroids were ordered in the same way by the Cox coefficients (see Table 2 and Figure 4). Overall, this analysis was consistent with previously published reports, since tumor samples expressing luminal A genes displayed reproducible negative Cox coefficients, while ERBB2

and luminal B tumors were consistently associated with positive coefficients. Although their Cox coefficients were near zero, normal-like and basal tumors appeared to be associated with a worse prognosis compared to the luminal A subtype (Figure 4).

We also evaluated the performance of the revised 1300 "intrinsic" genes set recently proposed by Hu et al. [18]. Mappings for this second list of genes are summarized in Table 1. The Cox coefficient correlations computed in all the UGC in the common set was higher than that of the old set (SV = 0.313; SH2 = 0.25, and VH2 = 0.454, see Table 2). Next we evaluated the performance of gene clusters obtained by hierarchical clustering of the original expression data used by Hu et al. obtained directly from the Perou lab, since in the original paper [18] the "new" intrinsic genes were not formally assigned to individual clusters (see Supplementary document for details). Using this definition of subtypes we again show reproducible negative Cox coefficients for the luminal A subtype and positive coefficients for the ERBB2 and luminal B groups. Centroids proved again more stable than individual genes (Figure 5). Cluster analysis allowed us to detect additional gene clusters that were used to fit the regression models. Two of them (namely, the cell cycle control and the pseudo-luminal A clusters) reproducibly associated with a better prognosis. Furthermore, similarly to what was observed for the old "intrinsic" gene set, all the clusters were similarly ranked by the Cox coefficients correlation, and the overall concordance across platforms increased when the reproducible genes, selected by the ICOR score, were used (see Table 2 and Figure 5).

Discussion

Genomic data analysis investigates the transcriptional activity of thousands of genes simultaneously. Because of the cost and limited accessibility of biological samples, most genomic investigations use relatively small numbers of biological samples. While this can provide highly valuable insight on gene regulation, important biological and medical correlations require a larger sample size. This issue is of particular relevance in the development and validation of gene expression classifiers for cancer patients, and for this reason microarray studies have been criticized for the lack of rigorous validation [32, 33].

The use of data sets from independent studies has emerged as an important option to overcome this problem, and our ability to efficiently integrate information from related genomic experiments will be critical to the success of the massive investment made on genomic studies. To date, multi-study analysis is

limited by the inherent variability of the technology. Effort must be made to facilitate the transfer and comparison of results among microarray platforms. Even in the presence of studies that investigate the same phenomenon, cross-platform inconsistency may arise from various sources, including erroneous gene mapping. Additional problems may arise from differences associated with the sample collections used in the various studies, or from the differing clinical annotations. Furthermore, the noise in microarray studies can differentially affect genes on the various platforms, thus reducing the overall power of the technology when data are combined. To address these issues, we have developed a simple, transparent, objective, and cheap analytic approach that selects reproducible genes and allows comparison and validation of microarray results without assimilating gene expression values from the various studies. In the present study, we have used this approach to obtain an independent validation of the "intrinsic" gene breast cancer subtype classification.

Although we have shown here that agreement among microarray platforms can be revealed by thorough investigations, several issues remain of concern when performing a cross-platform comparison. The first is how to accurately annotate the microarray features used in each platform. Because genomic and transcript sequences are continuously updated, mappings become inaccurate with time. In this study we have found, as previously noted, that different annotation tools [29, 30], as well as different annotation methods yield different results [34]. It has been reported that matching genes at the sequence level is more efficient than identifier-based mappings [35, 36]. However, the accurate identification of the common genes by sequence alignments (especially when oligonucleotide-based platform are included) resulted in smaller overlapping sets, when stringent criteria were applied. Moreover, a BLAST-based alignment of a large collection of genes requires powerful computational resources which are not always accessible, while identifier mapping methods are currently accessible as web-based tools and have been the most practical option so far. It seems clear that a balance between accuracy in the mapping procedures and the final number of genes used in the combined analysis is required. Use of the ICOR emerged as an effective approach to deal with these issues, since it enables discarding noisy genes, and does not suffer significantly from the potential false matches that may be present in gene sets obtained by identifiers based cross-referencing.

Second, while standards for microarray annotation [37] have contributed to improved comparability of technological and experimental variables, significant progress is still needed with regard to comparability of clinical variables, both in terms of annotation and measurement methodology. For instance, in the

VanDeVijver study ER status was assessed from microarray data and was encoded as a binary variable, in the Sorlie study it was assessed by a ligand binding assay and reported as a binary variable, while in the Huang study immunohistochemistry (and an immunoblot assay to confirm ER negativity status) was used to define three patients classes (+, ++ and +++). A common definition of ER status information required converting each measure to a common binary variable, independent of the methodology used. Despite these compromises, the overall comparison of the Cox coefficients for RFS demonstrated that there are genes that are reliably associated with clinical phenotypes.

In our analysis of the Huang study, we have identified three distinct batches based on hybridization dates. These also differ in terms of RFS and LN status. In particular, the first batch contained LN positive and negative patients characterized by short RFS times, the third batch contained LN positive patients who did not show recurrence and were observed for longer time intervals, while the second batch was composed of patients with variable RFS, relapse and LN status. The three groups showed expression data of consistent quality. The heterogeneity in phenotype made it difficult to identify consistent patterns of association to RFS, and we could demonstrate good agreement only between the second batch and the other two tumor collections. This example makes a strong case for giving careful consideration to study design and ascertainment in integrating microarray studies. The batches that do not show reproducibility are not necessarily incorrect, but most likely reflect a different empirical association between phenotype and transcription levels, as a result of the different mix of severity of disease.

The third issue we address is how to combine expression data from various platforms and how to carry out a cross-study validation of the results from various studies. In previous work, comparative meta-profiling has been successfully used to examine the similarity of significance values for each gene across various prostate cancer gene expression data sets, demonstrating a reasonably consistent pattern of gene dysregulation in prostate cancer compared with normal prostate [1, 2]. The same approach has been applied to other cancer data sets, to identify a common transcriptional profile consistently activated in most cancers compared to normal tissues [3]. We have developed the ICOR to assess the reproducibility of gene expression patterns across studies in both supervised and unsupervised settings. This method has been applied to select a subset of genes that ultimately show more consistent associations with histological classification and outcome in human lung carcinomas [8, 9]. These studies indicated the potential for combined analysis in microarray data. We here use the ICOR approach in breast cancer to identify the

genes that are reproducibly and consistently associated with RFS in three different studies. The implementation of this method allows the experimental determination of alternative cutoffs to pick up "reproducible genes", with the more stringent criteria significantly increasing the agreement among the three studies considered. In addition, this method, in conjunction with the cross-study correlation among Cox coefficients, provided a tool to select the gene mappings that not only resulted in consistent expression but showed similar association with survival across platforms. Furthermore, our approach could be used to deal with multiplicity in cross-referencing procedures, enabling the selection of the "best" mappings when there is more than one possibility.

Finally, we successfully validated the "intrinsic genes" from both the Sorlie and Hu studies [5, 18] in our analysis. It is of note that we have presented here the first independent validation of the molecular "intrinsic" breast cancer taxonomy, reporting how these sets of genes show higher Cox coefficient correlations across studies than the complete set of common genes. The ICOR-based gene filtering proved effective at increasing the agreement across studies. In addition, we have shown that the "centroids" of the "intrinsic" gene clusters' are highly correlated with each other across platforms, and that they characterized the corresponding cancer subtypes more reproducibly than individual genes. In particular, the luminal A subtype confirmed to be reproducibly associated with a better prognosis than the other subgroups, while the luminal B and ERBB2 groups showed consistent associations with worse RFS outcomes. The Cox coefficients for the basal-like phenotype yielded intermediate values. Finally, our analysis used all the tumors available, while in the original studies and in their validations, not all the patients could be classified into subtypes and used in the subsequent Kaplan-Meier analysis. This, as well as the difference in the statistical approach used, could explain the partial divergence that we observed with previous results, for instance, the lack of a clear association with a worst prognosis for the basal-like tumor subtype.

Conclusions

In conclusion, gene expression data often contains a large amount of noise from various experimental factors that make it difficult to combine data from various platforms for validation purposes. We have shown here that our analyses approaches have the potential to provide a robust foundation for the exploration of microarray data from different platforms, thus being a valuable tool for both the development of gene expression classifiers and their validation.

Methods

Data Preparation

Expression data were gathered from public repositories and analyzed using the statistical computing software R [38] with specific add-on packages from the Bioconductor suite [39]. 122 Stanford cDNA arrays [40] from the Sorlie study were obtained from the Stanford Microarray Database [41]. These arrays were from five different print-runs. Within-array print-tip loess normalization [42] without background subtraction [43] was performed separately for each hybridization, while cross-array scale normalization with limma [44–46] was performed separately for each batch. 8280 IMAGE clones were found to be in common across the five batches and thus used for all further analyses. Since there were no raw data available for the VanDeVijver study, 295 pre-processed Agilent long-oligo arrays representing 24479 transcripts were downloaded from Rosetta Inpharmatics [47] and used without further pre-processing. 89 Affymetrix human U95Av2 arrays were collected from the Duke Institute Genome Sciences & Policy [48], the expression values were obtained after background correction and quantile normalization at the probe levels with gcrma [49] as implemented in the `affy` package [50].

Patients Selection

We selected 104 patients from the Sorlie study, 295 from the VanDeVijver study and 88 from the Huang study. Normal samples, benign tumors and specimens corresponding to patients without evidence of disease or to patients having incomplete clinical information were not used in the analysis. Of the 487 patients included, 201 with evidence of local recurrence or distant metastasis were counted as failures in our analysis of RFS. Other clinical parameters, such as tumor size, LN status, ER status, and whether the patient was under chemotherapy, were also considered as possible cross-platform validation factors.

Microarray Features Annotation

Since three different platforms were considered, the following strategies were used to obtain the overlapping set of genes across the studies. The original identifiers were collected for each platform: IMAGE clone identifiers were used for the Sorlie study, GeneBank accession numbers for the VanDeVijver study and Affymetrix probe sets identifiers for the Huang study. These identifiers were subsequently mapped to a different common identifier. This task was accomplished by using the two web-based annotation tools, MatchMiner and SOURCE [29, 30]. Different unifying identifiers were then applied to obtain cross-referencing: UGC identifiers [28], Gene Symbols and Entrez Gene identifiers [25–27]. We also

performed cross-referencing across platforms using BLAST [24] alignments to RefSeq transcripts: IMAGE clone alignments to RefSeq were obtained from the IMAGE consortium [51], alignments of individual Affymetrix probes were obtained from the Lung transcriptome database [35, 36, 52], while the VanDeVijver sequences were aligned by BLAST to the last RefSeq release available in March 2006. Only probes showing a perfect match to the target sequence were further considered in the analysis.

Regression Models

Several regression models were used to explore the relationship between the patients' clinical characteristics and gene expression. Logistic regression was used to investigate the association between gene expression and ER, LN status and tumor size. Gene expression data, relapse status and RFS time were taken together to fit a Cox model [53] predicting the recurrence of cancer. The following covariates, representing different patient characteristics, were also added to the Cox models: tumor size, LN status, ER status, chemotherapy treatment. The regression models were fitted by using the original expression values and the normalized data, since different scales in expression in the three studies yielded coefficients on different scales. The genes were standardized by subtracting the mean expression value across patients and by dividing them by their standard deviation before the fitting.

The correlations among the vectors of gene-specific regression coefficients were used to evaluate the concordance between each pair of studies. Higher correlations of the coefficients indicate stronger concordance across studies in how the transcripts associated with the clinical phenotypes. This approach also allowed the evaluation of the procedures and settings applied in the various steps of our validation analysis, enabling the evaluation of their impact on ultimate consistency across studies. When the "intrinsic" gene sets were considered, the Cox coefficients were computed for all the individual genes, for the "intrinsic" clusters, and for the "centroids" of the clusters, which were obtained as the mean expression value of the genes within each cluster.

Integrative Correlation Method

Evaluation of the consistency of gene expression across different platforms was performed with the integrative correlation method [9]. Starting from the expression matrices for the common set of genes in different studies, we first computed the correlation matrices for all the studies, every row of which measures the linear relation between the corresponding gene with all the others. For every pair of matrices, we then

calculated the correlation between corresponding rows, or "correlation of correlations". We further averaged the correlation of correlation scores for each gene across all possible pairs of studies, and refer to the result as the "integrative correlation score". Since the different platforms had different identifiers, the cross-referencing process involved redundancy. For instance, multiple IMAGE clones, or probe set identifiers, could be mapped to the same UGC identifier. For this reason several features on one platform could refer to multiple features on a different platform, increasing the total number of pairs involved in the comparison. In this case, we saved all the possible pairing without averaging the gene expression levels. For example, we found 4125 common UGC across the studies, which corresponded to 11531 possible cross-referencing mappings. Such multiple mappings were not used to calculate the within-study correlation matrix. All the analyses above were performed with the add-on R package MergeMaid [10].

Gene Screening Method

In order to reduce noise by irrelevant genes, two approaches of gene selection were applied. One approach —variance filtering with pre-determined cutoff— was used to remove the genes that were not differentially expressed within each study. The other approach was based on the integrative correlation calculation above. To set a threshold, the observed "integrative correlation score" obtained from the original expression matrices was compared with the "null" integrative correlation score, as obtained by randomly labeling every row in the original expression matrices. The two distributions were compared using approximate density functions. Genes that were concordant across platforms were selected based on a bound on the FDR [31], given by the ratio of the tail probabilities in the empirical and null distributions. By using alternative FDR cutoffs we could detect the genes with high integrative correlation, the so-called "reproducible genes". The cutoffs in both approaches were experimentally determined in our analysis, by empirically balancing the gain in the reproducibility among data sets with the loss of genes.

Authors contributions

XZ, LM, EG and GP conceived and designed the study and participated in its coordination. XZ carried out most of the statistical analysis and the computer implementation; LM collaborated in the statistical analysis, performed the platforms cross-referencing procedures and the clustering analysis; ESI identified the Huang batches and performed exploratory analysis of the phenotypes in the Huang data set; LC, EG, ESGM and ESI critically evaluated the analysis plan and earlier draft, and provided implementation

suggestions; all authors read and approved the final manuscript.

Acknowledgments

This work has been supported by NSF Grant DMS034211 and by the Johns Hopkins SPORE in Breast Cancer P50CA88843. We thank Dr. Charles M. Perou for kindly providing additional unpublished information about the Hu study.

References

1. D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*, 62(15):4427–33, 2002. 0008-5472 (Print) Journal Article Meta-Analysis.
2. D. Ghosh, T. R. Barette, D. Rhodes, and A. M. Chinnaiyan. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct Integr Genomics*, 3(4):180–8, 2003. 1438-793X (Print) Journal Article Meta-Analysis.
3. D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101(25):9309–14, 2004. 0027-8424 (Print) Journal Article Meta-Analysis.
4. J. Wang, K. R. Coombes, W. E. Highsmith, M. J. Keating, and L. V. Abruzzo. Differences in gene expression between b-cell chronic lymphocytic leukemia and normal b cells: a meta-analysis of three microarray studies. *Bioinformatics*, 20(17):3166–3178, Nov 2004.
5. T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–23, 2003.
6. Z. Hu, M. Troester, and C. M. Perou. High reproducibility using sodium hydroxide-stripped long oligonucleotide dna microarrays. *Biotechniques*, 38(1):121–4, 2005. 0736-6205 (Print) Evaluation Studies Journal Article Validation Studies.
7. A. V. Kapp, S. S. Jeffrey, A. Langerød, A. Børresen-Dale, H. Wonshik, D. Noh, I. K. Bukholm, M. P. Nicolau, P. O. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(231):1–15, 2006.
8. D. Neil Hayes, Stefano Monti, Giovanni Parmigiani, C. Blake Gilks, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A Socinski, Charles Perou, and Matthew Meyerson. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol*, 24(31):5079–5090, Nov 2006.
9. G. Parmigiani, E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res*, 10(9):2922–7, 2004.
10. L. Cope, X. Zhong, E. Garrett, and G. Parmigiani. Mergemaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol*, 3(1):Article29, 2004.
11. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.
12. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

13. S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351(27):2817–26, 2004. 1533-4406 (Electronic) Clinical Trial Journal Article Randomized Controlled Trial.

14. T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74, 2001.

15. E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene expression predictors of breast cancer outcomes. *Lancet*, 361(9369):1590–6, 2003.

16. M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, Jr. Olson, J. A., J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98(20):11462–7, 2001. 0027-8424 (Print) Journal Article.

17. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002. 0028-0836 (Print) Journal Article.

18. Z. Hu, C. Fan, D. S. Oh, J. S. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. G. Ewend, L. R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Ruiz Orrico, D. Dreher, J. P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. F. Quackenbush, M. J. Ellis, O. I. Olopade, P. S. Bernard, and C. M. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7:96, 2006.

19. C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–8, 2003. 0027-8424 (Print) Journal Article.

20. X. J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. T. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5(6):607–16, 2004. 1535-6108 (Print) Journal Article.

21. H. Y. Chang, D. S. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, H. Dai, Y. D. He, L. J. van't Veer, H. Bartelink, M. van de Rijn, P. O. Brown, and M. J. van de Vijver. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A*, 102(10):3738–43, 2005. 0027-8424 (Print) Journal Article.

22. C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med*, 355(6):560–9, 2006. 1533-4406 (Electronic) Journal Article.

23. D. J. , H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, 1996. 1087-0156 (Print) Journal Article.

24. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997. 0305-1048 (Print) Journal Article Review.

25. D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue):D54–8, 2005.

26. K. D. Pruitt, K. S. Katz, H. Sicotte, and D. R. Maglott. Introducing refseq and locuslink: curated human genome resources at the ncbi. *Trends Genet*, 16(1):44–7, 2000.

27. K. D. Pruitt and D. R. Maglott. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Res*, 29(1):137–40, 2001.

28. JU Pontius, L Wagner, and GD Schuler. Unigene: A unified view of the transcriptome. In NCBI, editor, *The NCBI Handbook*, volume Part 3 - Section 21 - Querying and Linking the Data. 2003.

29. K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay, and J. N. Weinstein. Matchminer: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*, 4(4):R27, 2003.

30. M. Diehn, G. Sherlock, G. Binkley, H. Jin, J. C. Matese, T. Hernandez-Boussard, C. A. Rees, J. M. Cherry, D. Botstein, P. O. Brown, and A. A. Alizadeh. Source: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res*, 31(1):219–23, 2003.

31. B Efron and R Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

32. T. K. Jenssen and E. Hovig. Gene-expression profiling in breast cancer. *Lancet*, 365(9460):634–5, 2005. 1474-547X (Electronic) Comment Journal Article.

33. R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1):14–8, 2003. 0027-8874 (Print) Journal Article Review.

34. Yuan Ji, Kevin Coombes, Jiexin Zhang, Sijin Wen, James Mitchell, Lajos Pusztai, W. Fraser Symmans, and Jing Wang. Refseq refinements of unigene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Appl Bioinformatics*, 5(2):89–98, 2006.

35. B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res*, 32(9):e74, 2004. 1362-4962 (Electronic) Journal Article.

36. B. H. Mecham, D. Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T. J. Mariani. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics*, 18(3):308–15, 2004. 1531-2267 (Electronic) Journal Article.

37. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001. 1061-4036 (Print) Journal Article.

38. R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

39. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

40. C. A. Ball, I. A. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, and G. Sherlock. The stanford microarray database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*, 33(Database issue):D580–2, 2005. 1362-4962 (Electronic) Journal Article.

41. Stanford microarray database.

42. Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. spie. Technical report, SPIE BiOS 2001, San Jose, California, 2001.

43. Robert B. Scharpf, Christine A. Iacobuzio-Donahue, Julie B. Sneddon, and Giovanni Parmigiani. When should one substract background fluorescence in two color microarrays? Technical report, Johns Hopkins University, Dept. of Biostatistics, 2005.

44. G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(Article 3), 2004.

45. G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, R. V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

46. G. K. Smyth, J. Michaud, and H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–75, 2005. 1367-4803 (Print) Evaluation Studies Journal Article Validation Studies.

47. Rosetta inpharmatics.

48. Duke Institute of Genome Sciences & Policy.

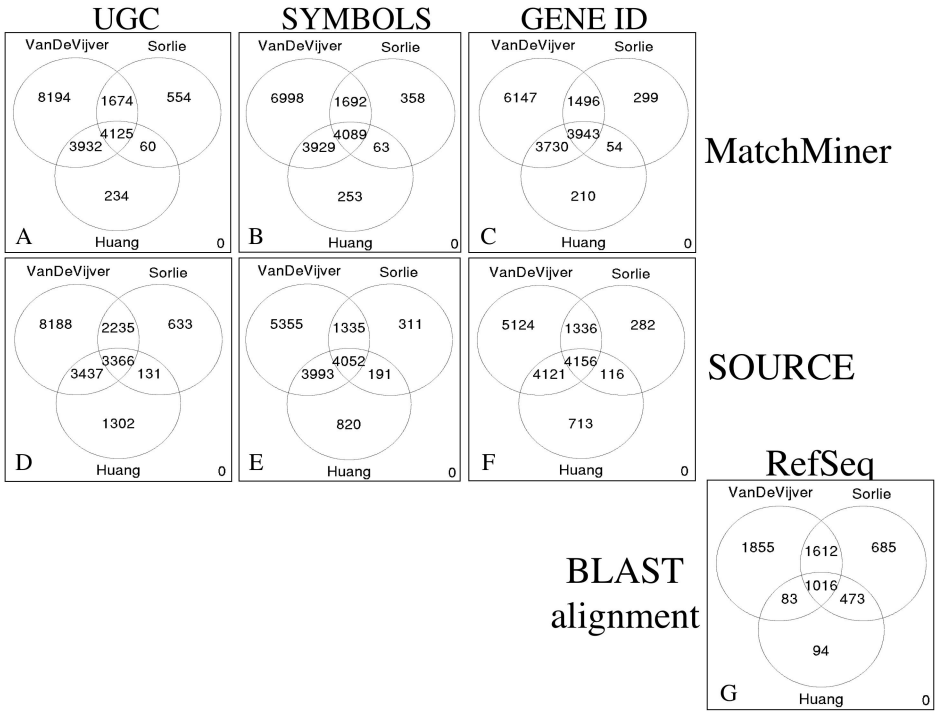
49. Zhijin Wu, Rafael Irizarry, Robert Gentlemen, F. Martinez-Murillo, and Forest Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

50. L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–15, 2004. 1367-4803 (Print) Evaluation Studies Journal Article.

51. IMAGE consortium.

52. Lung transcriptome database.

53. D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.



Figures

Figure 1 - Venn diagrams representing genes in common among the three studies

This figure summarized the number of genes in common in the three data sets considered in the present study. Mappings were obtained by the use of the two web-based tools MatchMiner and Source, and by direct sequence alignment with the RefSeq database by BLAST. Cross-referencing by identifiers was accomplished by using alternative mapping systems. Panel A: mapping by UGC obtained with MatchMiner; Panel B: mapping by gene symbols obtained with MatchMiner; Panel C: mapping by Entrez Gene identifiers obtained with MatchMiner; Panel D: mapping by UGC obtained with SOURCE; Panel E: mapping by gene symbols obtained with SOURCE; Panel F: mapping by Entrez Gene identifiers obtained with SOURCE; Panel G: mapping obtained by BLAST alignment with RefSeq transcripts.

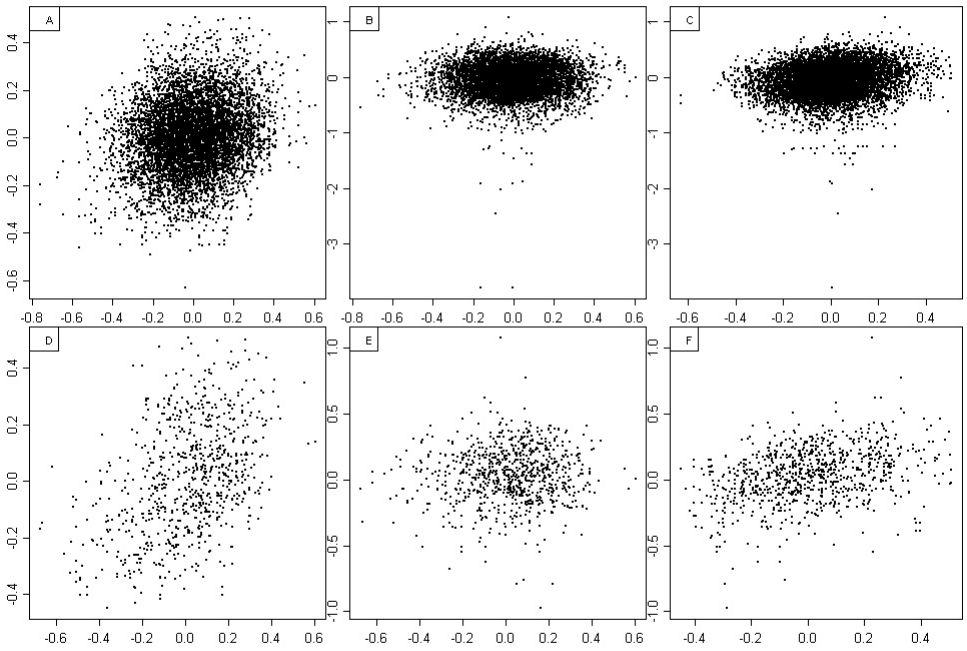


Figure 2 - All the genes with 88 patients from the Huang study

Two-way comparison of Cox coefficients with the 88 patients from the Huang study. 2(A-C) are plots for all the "common genes" across three studies, whereas 2(D-E) are plots for the highly reproducible "intrinsic gene clusters"(with integrative correlation more than 0.25): 2(A,D), Sorlie vs VanDeVijver; 2(B,E), Sorlie vs Huang; 2(C,F), VanDeVijver vs Huang.

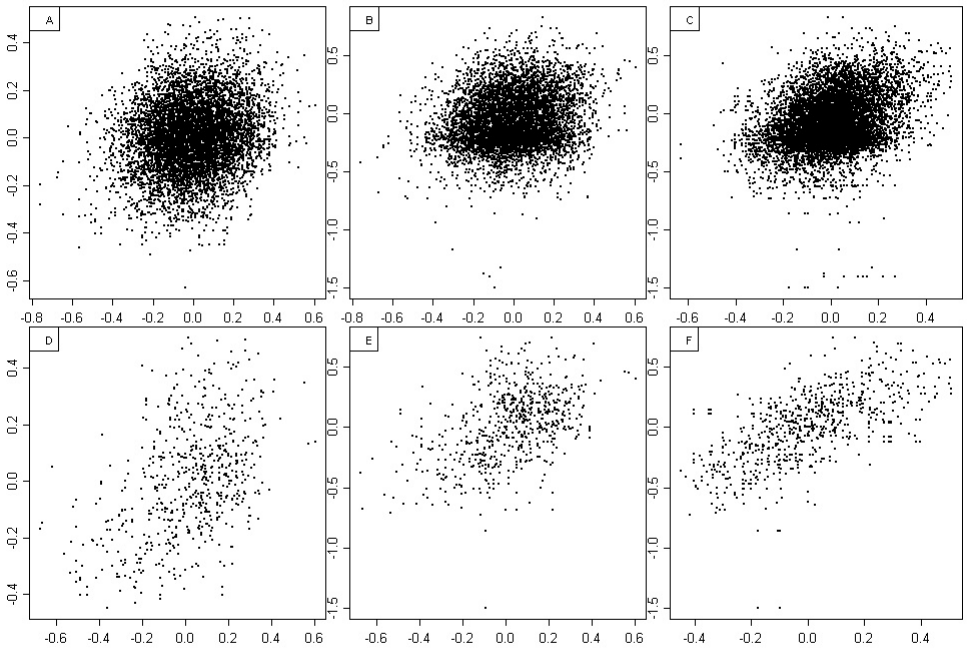


Figure 3 - All the genes with Huang batch2

Two-way comparison of Cox coefficients for all the gene mappings with Huang batch2. 3(A-C) are plots for all the "common genes" across three studies, whereas 3(D-E) are plots for the highly reproducible "intrinsic gene clusters"(with integrative correlation more than 0.25): 3(A,D), Sorlie vs VanDeVijver; 3(B,E), Sorlie vs Huang batch2; 3(C,F), VanDeVijver vs Huang batch2.

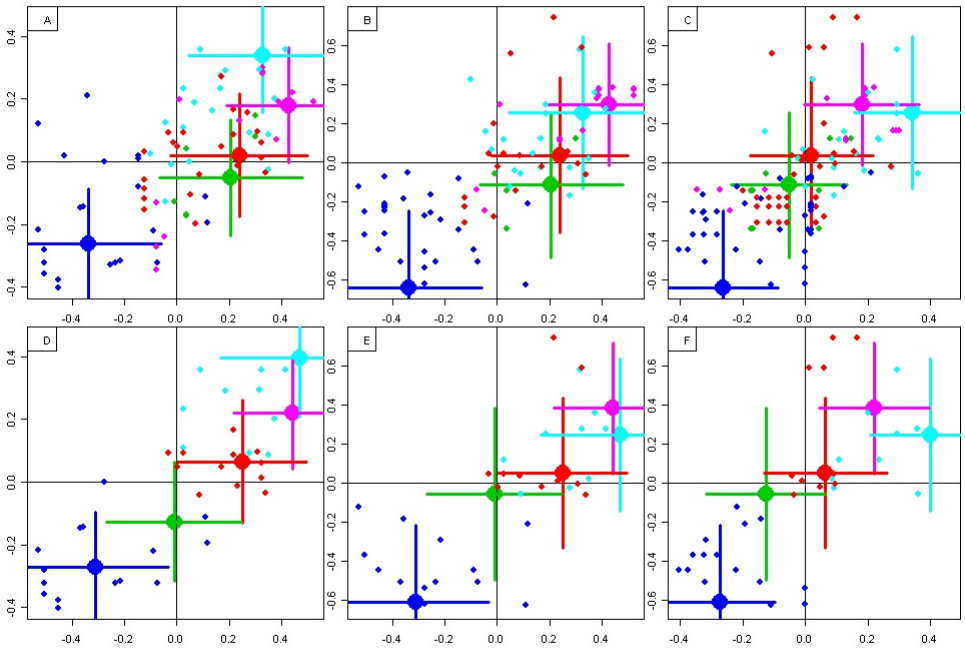


Figure 4 - Intrinsic gene clusters

Two-way comparison of Cox coefficients for the "intrinsic gene clusters". We calculated the mean expression value of the genes in each cluster as the corresponding "centroid genes". The bigger dots represent Cox coefficients of the "centroid genes", and the bars represent their confidence intervals. Twice the standard deviance of the coefficients was computed to measure the range of the confidence intervals. The small dots represent Cox coefficients of the gene mappings for the "intrinsic gene clusters". The five different colors identify the five associated breast cancer subtypes; the color scheme for the plot is as follows: light blue, luminal B; dark blue, luminal A; green, normal breast like; red, basal-like; pink, ERBB2. 4(A-C) are plots for all the "intrinsic gene clusters", whereas 4(D-E) are plots for the highly reproducible "intrinsic gene clusters"(with integrative correlation more than 0.25): 4(A,D), Sorlie vs VanDeVijver; 4(B,E), Sorlie vs Huang batch2; 4(C,F), VanDeVijver vs Huang batch2.

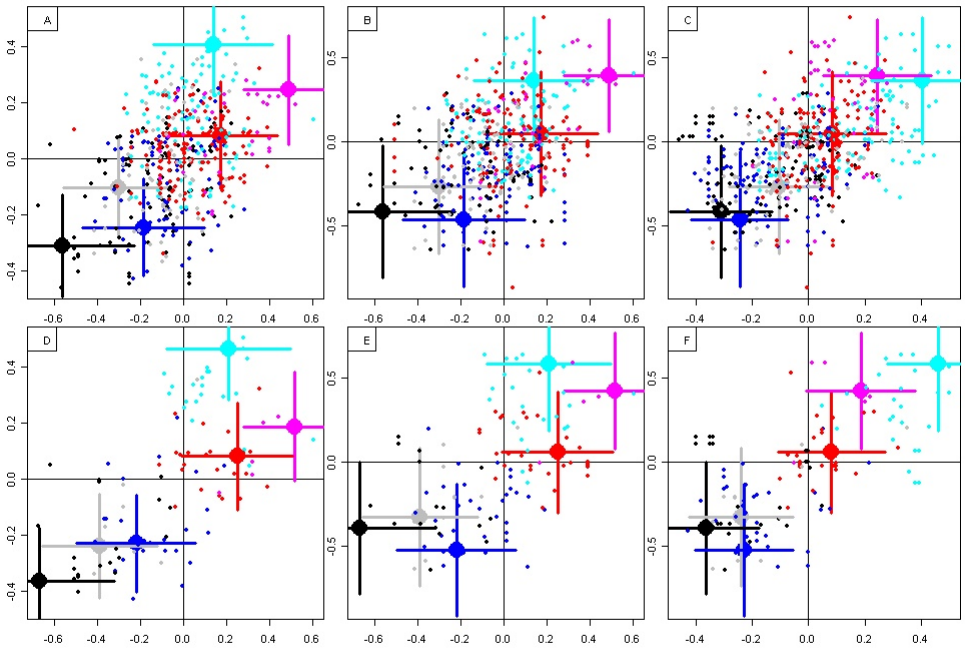


Figure 5 - New intrinsic gene clusters

Two-way comparison of Cox coefficients for the new "intrinsic gene clusters". We used the same plot scheme as Figure 4, and we added two more gene clusters created from the new "intrinsic gene list" by hierarchical clustering of the original data from Hu et al [18]: black with pseudo-luminal A and gray with cell-cycle control. 5(A-C) are plots for all the new "intrinsic gene clusters", whereas 5(D-E) are plots for the highly reproducible "intrinsic gene clusters"(with integrative correlation more than 0.25): 5(A,D), Sorlie vs VanDeVijver; 5(B,E), Sorlie vs Huang batch2; 5(C,F), VanDeVijver vs Huang batch2.

Tables

Table 1 - Genes in common among the considered studies

Table 1 summarizes the genes that were found to be in common across the three data sets, according to the various annotation methods that were used. Data are reported for the complete common sets and for the "old" and the "new" intrinsic gene lists.

	MatchMiner			SOURCE			RefSeqBLAST
	UGC	SYMBOLS	EGID	UGC	SYMBOLS	EGID	RefSeqNR
Sorlie et al.	6413	6202	5972	6365	5889	5890	3786
VanDeVijver et al.	17925	16708	15316	17226	14734	14737	4566
Huang et al.	8351	8334	7937	7066	9056	9106	1666
Common	4125	4089	3943	3366	4052	4156	1016
Intrinsic Genes	354	350	342	287	340	348	86
New Intrinsic Genes	382	378	372	154	353	375	87
70-genes -van't Veer et al.	22	19	19	14	18	19	6

	Procedures	Num of Gene Mappings	Num of UGIs	SV	SH	SH2	VH	VH2
Common Refseqs	all the common genes	1226	1016	0.132	-0.036	NA	0.083	NA
	all the common genes, Std	1226	1016	0.178	0.001	NA	0.168	NA
Common UGCs	all the common genes	11531	4125	0.115	-0.011	0.061	0.038	0.1
	all the common genes, Std	11531	4125	0.165	0.005	0.152	0.127	0.3
	reproducible genes, variance filtering, FDR .1, Std	3912	1935	0.228	0.013	NA	0.2	NA
	reproducible genes, FDR .1, Std	7419	3359	0.202	-0.002	NA	0.197	NA
	reproducible genes, Huang Batch 2, FDR .1, Std	7215	3305	0.202	NA	0.204	NA	0.425
	reproducible genes, Huang Batch 2, FDR .1, Std	5065	2612	0.234	NA	0.239	NA	0.498
	reproducible genes, Huang Batch 2, ICOR > .25, Std	865	513	0.355	NA	0.437	NA	0.616
Common intrinsic genes, UGCs	intrinsic genes, Std	1087	354	0.198	0	0.217	0.246	0.423
	reproducible genes, FDR .1, Std	889	334	0.237	NA	0.244	NA	0.453
	reproducible genes, FDR .01, Std	760	295	0.359	NA	0.311	NA	0.551
	reproducible genes, ICOR>.25, Std	224	100	0.382	NA	0.425	NA	0.545
Intrinsic gene clusters, UGCs	intrinsic gene clusters, Std	119	56	0.639	0.357	0.686	0.305	0.623
	reproducible gene clusters, FDR .1, Std	104	52	0.731	NA	0.706	NA	0.652
	reproducible gene clusters, FDR .01, Std	93	50	0.744	NA	0.712	NA	0.657
	reproducible gene clusters, ICOR>.25, Std	40	30	0.763	NA	0.716	NA	0.698
	gene clusters centroids, Std	5 centroids	5 centroids	0.851	0.459	0.977	0.361	0.93
New intrinsic gene clusters, UGCs	new intrinsic gene clusters, Std	796	296	0.337	0.064	0.241	0.113	0.41
	reproducible gene clusters, FDR .1, Std	547	265	0.39	NA	0.323	NA	0.526
	reproducible gene clusters, FDR .01, Std	462	243	0.418	NA	0.34	NA	0.563
	reproducible gene clusters, ICOR>.25, Std	129	89	0.592	NA	0.472	NA	0.661
	gene clusters centroids, Std	5 centroids	5 centroids	0.831	0.706	0.88	0.496	0.97

Table 2 - Correlation of the Cox coefficients under various conditions

Table 2 summarized the correlation of the Cox coefficients for different subsets of genes with various groups of patients(see details in the method section). Here, "Std" means genes were standardized before fitting the Cox model; "FDR x" means genes were filtered by FDR cutoff x; "Variance filtering" means low-variance genes were filtered before the analysis; "NA" means the correlation value is not available; UGI represents unique gene identifiers, whereas UGCs and Refseqs represents unique gene clusters and Refseqs transcripts; ICOR is the integrative correlation; "Intrinsic gene clusters" are those genes used to build the gene clusters; There are five pairs of comparison "SV"(Sorlie versus VanDeVijver), "SH"(Sorlie versus Huang), "SH2"(Sorlie versus Huang Batch 2), "VH"(VanDeVijver versus Huang) and "VH2"(VanDeVijver

versus Huang Batch 2).

Additional Files

Supplementary Information

Original paper: ``Optimized integrative analysis of gene expression patterns for independent cross-platform validation''.

Authors: Xiaogang Zhong, Luigi Marchionni, Leslie Cope, Edwin S. Iversen, Elizabeth S. Garrett-Mayer, Giovanni Parmigiani and Edward Gabrielson.

Analysis of the ``new'' intrinsic genes list from Hu et al [1].

Hu and colleagues recently published [1] a reviewed ``intrinsic'' gene list obtained from the analysis of a completely new patients collection. To develop the new gene set, 105 breast tumor samples and 9 normal breast samples, which contained 26 sample pairs, were assayed, using various Agilent oligonucleotide microarrays. The same methodology applied by Sorlie et al. [2-4] was used to select a list of 1410 features representing 1300 UGC. The main differences between the ``old'' and the ``new'' gene list were the number of included genes and the use of pre-treatment tumor pairs, rather then pre- and post- chemotherapy pairs. Since in the original paper by Hu and colleagues genes were not formally assigned to the different clusters, we performed hierarchical clustering analysis to obtain the groups of genes to be used in our analysis. Pre-normalized expression data used by He et al. were kindly provided by Dr. Charles M. Perou and were not further processed. However, in order to be consistent with the mapping strategy we applied in the present work to cross-reference the different platforms, we re-annotated the microarrays features contained in the ``new'' intrinsic gene set by using their GenBank accession numbers, as the input for the web-based tools MatchMiner and SOURCE [5, 6].
382 genes were present in the common set obtained by using UGC and MatchMiner and were further used in the analysis. Hierarchical clustering was performed using the Pearson uncentered distance and the complete linkage method. The gene tree was cut at a distance equal to 0.31, obtaining 9 different clusters of genes (see Figure S1)

Figure S1 - Hierarchical clustering of the ``new'' intrinsic gene list

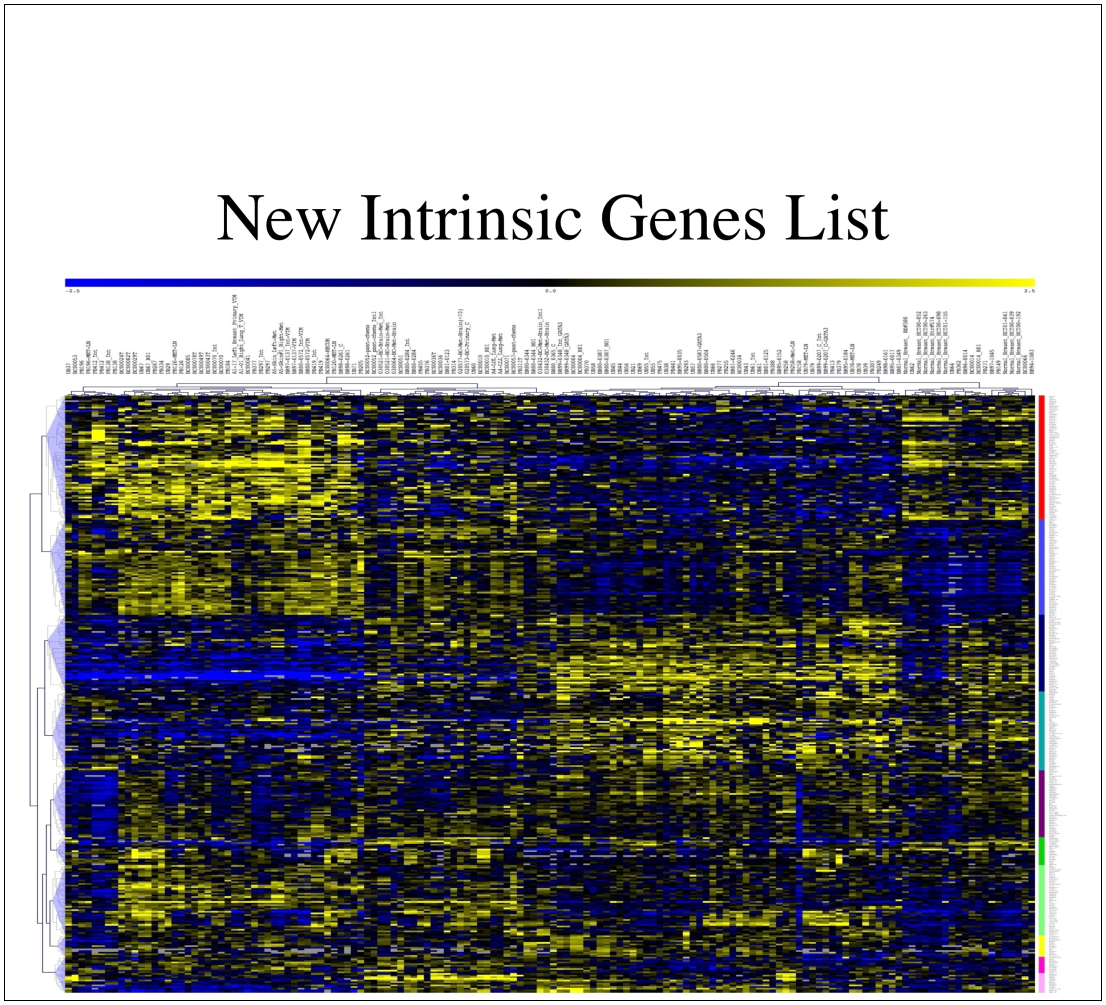


Figure S1: Hierarchical clustering of the 382 genes of the ``new'' intrinsic list, that could be mapped in the set of genes in common among the Huang [7], the Sorlie [3, 4] and the VanDeVijver [8] studies, as obtained by UGC and MatchMiner. The Pearson uncentered correlation and the complete linkage method were used; 9 clusters were find by cutting the three at a distance equal to 0.31.

Gene clusters were subsequently manually reviewed and genes present in the old intrinsic gene list, as obtained from the original paper by Sorlie and colleagues, were used to label them. The basal-like, luminal A, luminal B, and ERBB2 clusters were readily identified, while none of the genes from the normal-like cluster could be retrieved. Several additional clusters were also obtained, which were labeled accordingly to the biological processed in which the contained genes were involved. The following additional gene clusters were identified:

- Lymphocyte B/ signal transduction cluster;
- Lymphocyte T/ Interferon response cluster;
- Cell cycle control genes cluster;
- Myst3/Wisp1 cluster;
- Pseudo-luminal A cluster

Identification and evaluation of the Huang study batches.

Our analysis on reproducibility across the three considered studies [3, 4, 7, 8] showed that all the pair wise comparisons involving the Huang data set did not reveal good reproducibility with respect to the other two data sets. For this reason we deeply investigated this study, by looking at both expression and phenotype data, to understand whether this was due to any specific feature associated with this study or platform. Evaluation of the CEL files headers showed that there were three major hybridization batches, if the experiments' date was considered (see Figure S2).

Figure S2 – Huang batches

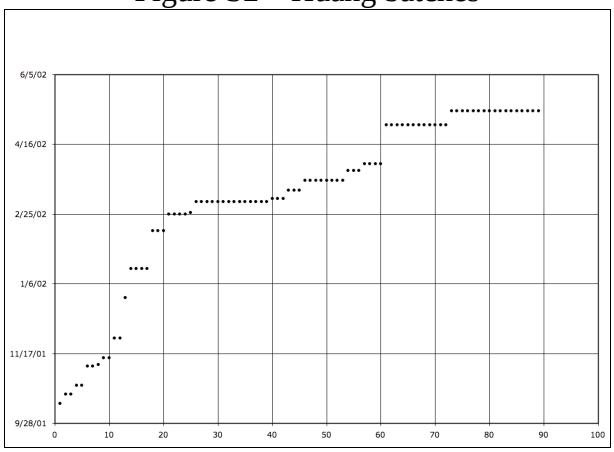


Figure S2 – Huang batches: arrays are ordered chronologically the hybridization dates are reported on the y-axis. Two major splits are visible after the 17th and the 60th hybridizations.

We subsequently evaluated if the identified batches corresponded to subgroup of hybridizations with distinct features or to patients with different clinical characteristics. The correlation of expression data for every pair of samples, using all the genes in the Affymetrix hgu95av2 platform was calculated and a heatmap of such pair-wise correlation matrix was drawn (see Figure S3), which indicated that the expression of genes was fairly homogenous and comparable across all the samples, with a range of the correlations between 0.749 to 0.976.

Figure S3 – Huang batches

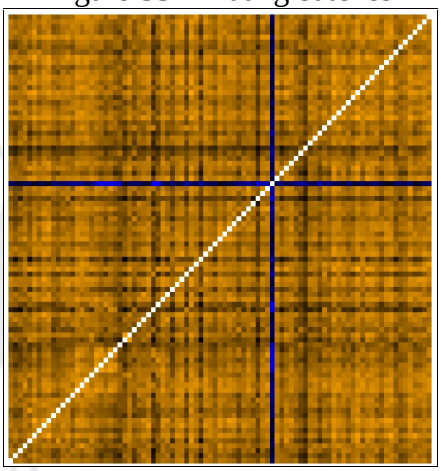


Figure S3 – Huang batches

We also calculated the observed and the null distributions of the integrative correlations of each pair wise comparison between batches and the approximate density plots confirmed that the three subgroups were highly correlated with each other in terms of expression (see Figure S4).

Figure S4a

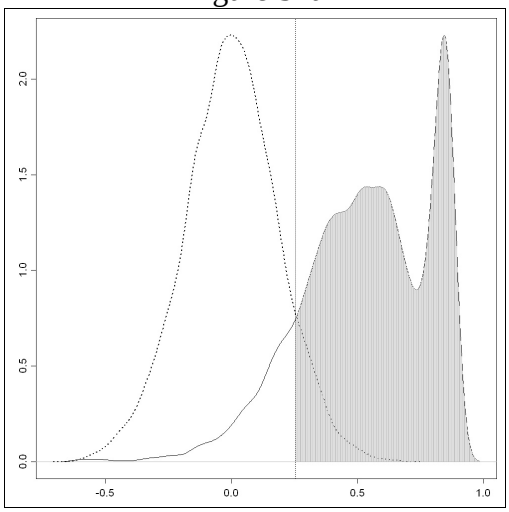


Figure S4a: integrative correlation distributions for the Huang study batches; highly correlated genes corresponded to low intensity and saturated genes. This picture shows how the three batches are concordant in terms of gene expression. Batch1 and batch 3 are aggregated and compared with batch 2.

Considering the hump with high integrative correlation, we filtered those low-variance genes (we cut the lowest 20 percentiles), and the same plot can be made as follows, Figure S4b,

Figure S4b

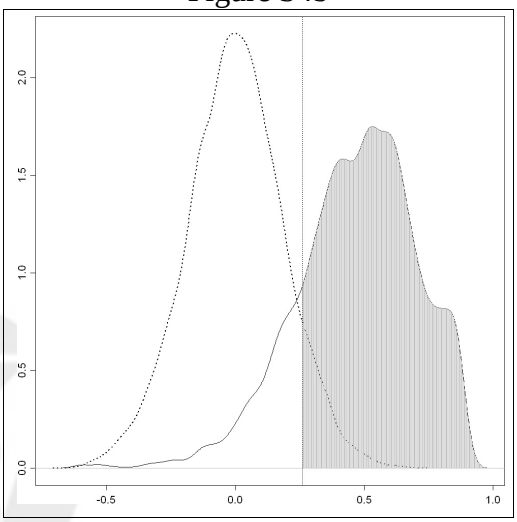


Figure S4b: integrative correlation distributions for the Huang study batches; highly correlated genes corresponding to low intensity and saturated genes were removed. This picture shows how the three batches are concordant in terms of gene expression. Batch1 and batch 3 are aggregated and compared with batch 2.

Since we could conclude that the three batches were similar in terms of gene expression, we subsequently evaluated if they were homogeneous in term of clinical phenotypic data. We simply plot the RFS survival time as a function of the date of the experiment and a clear increasing trend was evident (see Figure S5).

Figure S5 - RFS in the Huang study batches

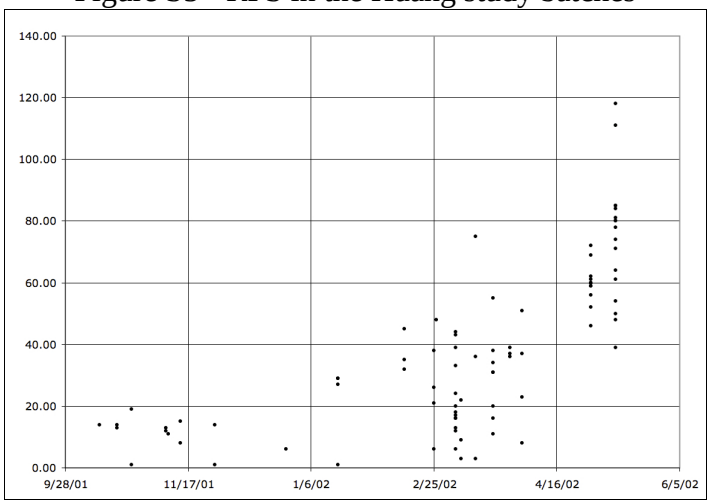


Figure S5: RFS time in months (y-axis) plotted against the experiments date (x-axis).

Logistic regression analyses using serial number as the predictor variable indicate the presence of a temporal trends in the data (see Figure S6). Note that samples for the recurrence analysis where collected later in the study while those used in the lymph node study where arrayed earlier. Early arrays tended to be of ER+, PR+ tumors while those later in the study reflected a mix of subtypes. These features are also evident in the tabular analysis that follows.

Figure S6: Logistic regression.

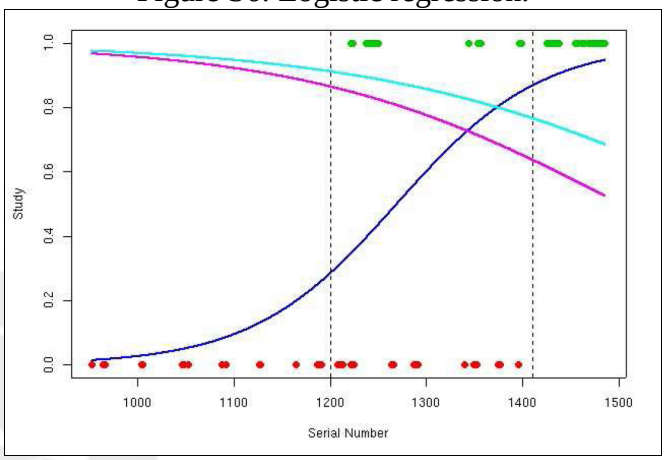


Figure S6: Plot of sub-study (lymph node = 0; recurrence = 1) by array serial number. The dark blue line is the logistic regression line of the recurrence samples on serial number, the light blue and the purple one represents the logistic regression line of ER positive and PR positive samples versus serial number. The vertical dashed lines delineate our partitions. The red points and green points are lymph node samples (LN) and recurrence samples (Re), respectively.

For this reason we decided to evaluated the three batches by fitting a Cox model with RFS, the relapse status and expression data and to compared the obtained Cox coefficients as estimate of the agreement among batches. The correlation of the coefficients for all the genes in the three subgroups resulted to be -0.151 for batch1 vs batch2, 0.167 for batch1 vs batch2+batch3. We could not compare batch2 with batch3 alone, since no patients in this latter group relapsed.

The tables below report the distributions of other phenotypic variables known for the Huang study by sub-study (Lymph Node Positivity or Relapse Status) and/or batch, as we've defined them. These tables detail differences in cases used in the two analyses and temporal trends in recruitment. Table 1 tabulates samples by sub-study and temporal batch. Samples in the LN sub-study were arrayed early in the study while those for the Relapse sub-study where arrayed later.

Table S1. LN and relapse status

Samples Included	LN	Relapse
Batch one	17	0
Batch two	20	23
Batch three	0	29

Tables S2 and S3 tabulate ER and PR status by sub-study. Note that there are no ER negative tumors and only 1 PR negative tumor in the lymph node sub-study. Batch 1 is comprised solely of ER positive tumors and has only one PR negative case. Of the remaining PR negatives, 10 are in batch 2, and 12 are in batch 3. Seven of the 15 ER negatives are found in batch 2.

Table S2. ER status by sub-study.

Samples Included	LN	Relapse
ER +	11	14
ER ++	8	12
ER +++	18	10
ER –	0	15

Table S3. PR status by sub-study.

Samples Included	LN	Relapse
PR +	17	14
PR ++	7	10
PR +++	12	5
PR –	1	12

Table S4 summarizes the distribution of tumor size by batch. While tumor size does not vary appreciably by batch, there appears to be a general trend to smaller tumors as the study progressed.

Table S4. Distribution of tumor size by batch.

Samples Included	Min	First Quarter	Median	Mean	Third Quarter	Max	NA
Batch 1	1.1	1.8	2.5	3.282	4.2	7.5	NA
Batch 2	0.5	1.725	2.25	2.643	3.075	8.5	1

Batch 3	0.2	2	2.3	2.565	3	5	NA

Tables S5 and S6 tabulate samples by their nodal status (positive or negative) and sub-study (S5) or batch (S6). Note that all relapse sub-study patients are node positive and, consequently, that all batch 3 cases are as well.

Table S5. Table of lymph node status by sub-study.

Samples Included	Negative	Positive
LN sub-study	19	18
Relapse sub-study	0	51

Table S6. Table of lymph node status by batch

Samples Included	LN Negative	LN Positive
Batch one	6	11
Batch two	13	29
Batch three	0	29

Tables S7 through S10 tabulate Batch 2 samples according to their sub-study, ER status, PR status and relapse status. Note that in Batch Two, which is the Huang subset that we used in our validation, all data on individuals that don't relapse comes from the LN sub-study and they are all ER+/PR+.

Table S7. ER versus PR status among lymph node sub-study samples with no relapse.

Samples Included	ER Negative	ER Positive
PR Negative	0	0
PR Positive	0	11

Table S8. ER versus PR status among recurrence sub-study samples with no relapse.

Samples Included	ER Negative	ER Positive
PR Negative	0	0
PR Positive	0	0

Table S9. ER versus PR status among lymph node sub-study samples with relapse.

Samples Included	ER Negative	ER Positive
PR Negative	0	0
PR Positive	0	9

Table S10. ER versus PR status among recurrence sub-study samples with relapse.

Samples Included	ER Negative	ER Positive
PR Negative	6	1
PR Positive	4	11

Finally, using the subset of matched intrinsic genes, we compared the coefficients of standardized expression measures in logistic regressions of Relapse(1/0) on expression estimated from (1) all Sorlie data, (2) the Huang Recurrence sub-study only, (3) Huang batch one only, (4) Huang batch two only, (5) Huang batch three only, (6) Huang batch two only, conditioning on a binary indicator for sub-study, (7) all Huang data given binary variables for ER and PR positivity and (8) all Huang data given ER and PR positivity and serial number. We fit a separate logistic regression model for each matched intrinsic gene under each of these 8 scenarios. For each regression fit, we saved the estimate of the coefficient of the expression variable and collected these estimates into eight vectors, each corresponding to a scenario. Table S10 tabulates correlation coefficients for each pair of scenarios. The more highly correlated two scenarios are, the more reproducible the relationship between expression and relapse. Note that the scenario most highly correlated with the Sorlie data is (4) unadjusted Huang batch 2.

Table S10. Correlations of study to study and sub-study expression estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	1.000	0.014	0.043	0.263	-0.042	-0.109	0.019	-0.016
(2)	0.014	1.000	-0.019	0.565	0.070	0.107	0.733	0.633
(3)	0.043	-0.019	1.000	0.002	0.068	-0.010	0.015	0.009
(4)	0.263	0.565	0.002	1.000	0.006	0.430	0.629	0.621
(5)	-0.042	0.070	0.068	0.006	1.000	-0.071	0.028	0.069
(6)	-0.109	0.107	-0.010	0.430	-0.071	1.000	0.383	0.363
(7)	0.019	0.733	0.015	0.629	0.028	0.383	1.000	0.929
(8)	-0.016	0.633	0.009	0.621	0.069	0.363	0.929	1.000

These results suggest a significant level of sub-study to sub-study variability within the Huang study. This appears to be due to the fact that the three 'batches' differed with respect to patient RFS time, relapse status and LN status. All specimens in the third batch, indeed, corresponded to patients who were all LN positive and who showed longer RFS time and no recurrence of the disease. Collectively these evaluations showed that the three groups of patients, although not different in terms of gene expression data, were distinct in terms of relapse free survival and LN status, possibly being the result of sampling from different patient populations.

References

1. Z Hu, C Fan, DS Oh, JS Marron, X He, BF Qaqish, C Livasy, LA Carey, E Reynolds, L Dressler, et al: **The molecular portraits of breast tumors are conserved across microarray platforms**. *BMC Genomics* 2006, **7**:96.
2. CM Perou, T Sorlie, MB Eisen, M van de Rijn, SS Jeffrey, CA Rees, JR Pollack, DT Ross, H Johnsen, LA Akslen, et al: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**:747-52.
3. T Sorlie, CM Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, MB Eisen, M van de Rijn, SS Jeffrey, et al: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proc Natl Acad Sci U S A* 2001, **98**:10869-74.
4. T Sorlie, R Tibshirani, J Parker, T Hastie, JS Marron, A Nobel, S Deng, H Johnsen, R Pesich, S Geisler, et al: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proc Natl Acad Sci U S A* 2003, **100**:8418-23.
5. KJ Bussey, D Kane, M Sunshine, S Narasimhan, S Nishizuka, WC Reinhold, B Zeeberg, W Ajay, JN Weinstein: **MatchMiner: a tool for batch navigation among gene and gene product identifiers**. *Genome Biol* 2003, **4**:R27.
6. M Diehn, G Sherlock, G Binkley, H Jin, JC Matese, T Hernandez-Boussard, CA Rees, JM Cherry, D Botstein, PO Brown, et al: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data**. *Nucleic Acids Res* 2003, **31**:219-23.
7. E Huang, SH Cheng, H Dressman, J Pittman, MH Tsou, CF Horng, A Bild, ES Iversen, M Liao, CM Chen, et al: **Gene expression predictors of breast cancer outcomes**. *Lancet* 2003, **361**:1590-6.
8. MJ van de Vijver, YD He, LJ van't Veer, H Dai, AA Hart, DW Voskuil, GJ Schreiber, JL Peterse, C Roberts, MJ Marton, et al: **A gene-expression signature as a predictor of survival in breast cancer**. *N Engl J Med* 2002, **347**:1999-2009.