Genome Medicine

COMMENTARY

Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data?

Takayuki Iwamoto and Lajos Pusztai*

Abstract

A large number of prognostic and predictive signatures have been proposed for breast cancer and a few of these are now available in the clinic as new molecular diagnostic tests. However, several other signatures have not fared well in validation studies. Some investigators continue to be puzzled by the diversity of signatures that are being developed for the same purpose but that share few or no common genes. The history of empirical development of prognostic gene signatures and the unique association between molecular subsets and clinical phenotypes of breast cancer explain many of these apparent contradictions in the literature. Three features of breast cancer gene expression contribute to this: the large number of individually prognostic genes (differentially expressed between good and bad prognosis cases); the unstable rankings of differentially expressed genes between datasets; and the highly correlated expression of informative genes.

Introduction

Gene-expression profiling allows simultaneous, semi-quantitative measurements of thousands of different mRNA species in a single experiment. It was considered logical to assume that different cancers will have distinct gene-expression patterns and that the expression of many genes will be associated with clinically relevant disease outcomes in particular cancer types. Consequently, it was assumed these associations might be exploited to develop a new generation of multi-gene diagnostic tests, in particular prognostic and treatment response predictors.

*Correspondence: lpusztai@mdanderson.org
Department of Breast Medical Oncology, MD Anderson Cancer Center, University of Texas, Houston, TX 77230-1439, USA
Full list of author information is available at the end of the article

It has quickly become apparent that cancers of different organs have very different gene-expression patterns; indeed, this fact led to the development of a novel gene-expression-based molecular diagnostic test to assign a histological origin to metastatic cancers that present as 'cancers of unknown primary' [1]. Gene-expression profiling results also prompted re-evaluation of disease classification for certain tumors, most prominently breast cancer. Breast cancer used to be considered as a single disease with variable histological appearance and variable expression of estrogen receptor (ER) and other molecular markers. Gene-expression profiling studies revealed surprisingly large-scale molecular differences between ER-positive and ER-negative cancers that suggested that these two different types of breast cancers are distinct diseases [2-4]. A new molecular classification schema was proposed, but how many molecular classes there are and what method is best to assign these classes continues to be debated [5]. Currently, there is no standard, readily available, gene-expression-based test to determine the molecular class of breast cancer in the clinic.

Molecular classification emerged through unsupervised analysis of gene-expression data. The goal of this analysis is to identify disease subsets that show similar gene-expression patterns within a larger cohort of cases. During this analysis, the molecular subsets are defined without considering clinical outcome information. Consequently, the emerging molecular subsets may or may not differ in prognosis or response to various therapies. A parallel research effort has focused on developing supervised outcome predictors. This approach relies on comparing cases with known outcome (such as recurrence versus no recurrence). The goal of the analysis is to identify differentially expressed genes between outcome groups and use these genes to develop a multi-gene outcome predictor. Evaluation of the predictive accuracy of the supervised model requires independent validation cases. Investigators who developed the first generation of supervised prognostic and treatment response predictors started with the then prevailing notion that breast cancer is a single disease, and all

subtypes of breast cancer were included in the analysis. This resulted in major limitations in the diagnostic products that emerged from this research [6,7].

The plethora of prognostic gene signatures for breast cancer

Unsupervised molecular classification identified three major and robust groups of breast cancers that differ in the expression of several hundred to a few thousand genes. These include basal-like breast cancers, which are negative for ER, progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2); low histological grade ER-positive breast cancers (also called luminal A); and high grade, highly proliferative ER-positive cancers (luminal B). Several smaller and less stable molecular subsets (such as normal-like, HER-2-positive and claudin-low) have also been proposed but are less consistently seen and are distinguished by substantially smaller molecular differences [4,5]. Importantly, among the various molecular subsets, one group, the luminal A class that includes low grade ER-positive cancers, stands out with a very favorable prognosis with or without adjuvant endocrine therapy. The other groups have worse but rather similar prognosis [4,8].

If one understands these close associations between clinical phenotype, molecular class and prognosis, it is no longer surprising that comparing gene-expression profiles of breast cancers that recurred (mostly the ER-negative and the high grade, ER-positive cancers) and those that did not (low grade, ER-positive cancers) in the absence of any systemic therapy (or after anti-estrogen therapy alone in the case of ER-positive cancers) yields a very large number of differentially expressed genes. The relative position of individual genes in a rank-ordered gene list varies greatly, but the consistency of the gene list membership is fairly high across various datasets [9]. Functional annotation indicates that the majority of these prognostic genes are proliferation-related genes and the remainder are mostly ER-associated and, to a lesser extent, immune-related genes [10-12]. Because these genes function together in a coordinated manner in the regulation and execution of complex biological processes, such as cell proliferation, or originate from a particular cell type, such as immune cell infiltrate, many of these prognostic genes are also highly co-expressed with one another. It is therefore expected that a large number of nominally different prognostic signatures can be constructed that all perform equally well.

For example, a particular gene may be highly significantly discriminating in two datasets but it is ranked 5th among the most discriminating genes in one dataset (based on *P*-value or fold difference) but only 35th in another dataset (which is still very high considering the thousands of comparisons!). In multivariate prediction model building, the top few informative features are usually combined and genes are added incrementally to increase the predictive performance. However, because many of the genes are highly correlated with each other, adding genes lower on the list yields less and less improvement in the model as a result of lack of independence. Therefore, the gene in question will be included in a predictor developed from the first dataset (because it is ranked as 5th) and will work well on validation in the second dataset; but if a new predictor were to be developed from the second dataset, this gene may not be included in the predictor (because it is ranked 35th). These three features of the breast cancer prognostic gene space – the large number of individually prognostic features, the unstable rankings, and the highly correlated expression of informative genes – explain why it is easy to construct many different prognostic predictors that perform equally well even if they rely on nominally different genes in the model. However, this does not mean that all published prognostic gene signatures are equally ready for clinical use.

Before adoption in the clinic, a molecular diagnostic assay has to be standardized, the reproducibility within and between laboratories and stability of results over time have to be demonstrated, and its predictive accuracy has to be validated in the right clinical context, preferably in multiple independent cohorts of patients. Most importantly, clinical utility implies that the assay improves clinical decision making and complements or replaces older standard methods, which in turn leads to better patient outcomes. Few published prognostic predictors have met these criteria [13,14].

Why signatures work less well than expected

The predictive performance of a multivariate model largely depends on the number of independent informative genes included in the model, the magnitude of differential expression of the informative genes and the complexity of the background. Different clinical prediction problems show different degrees of difficulty. From the discussion above it should be apparent that prediction of ER status, histological grade of breast cancer, or better or worse prognosis associated with these clinical phenotypes should be relatively easy when considering all breast cancers together, and that such predictions can therefore yield predictors with good overall accuracy. Indeed, prognostic gene signatures developed for breast cancer in general or for ER-positive cancers tend to have good performance characteristics [12,15-17].

However, the first-generation prognostic signatures share some limitations. Because these were invariably developed by analyzing all subtypes of breast cancers

together, they tend to assign high risk category to almost all ER-negative cancers (which are almost always high grade), even though a substantial majority of these cancers have good prognosis [18,19]. Similarly, the good- and poor-prognosis ER-positive cancers, as assigned by gene profiling, tend to correspond to the clinically low grade/low proliferation versus high grade/high proliferation subsets, respectively. This strong correlation between prognostic risk as predicted by gene signatures and routine clinical variables, such as histological grade, proliferation rate and ER status, limits the practical value of these tests. Efforts are under way to develop simple multivariate prognostic models that use routine pathological variables (such as ER, histologic grade and HER2 status), and these could eventually rival the performance of the first-generation prognostic gene signatures [20,21]. However, standardization of the pathological assessment of breast cancer and reducing the inter-observer variability remains an important challenge.

Predicting clinical outcome, such as prognosis or response to chemotherapy, within clinically and molecularly more homogeneous subsets (such as triple-negative breast cancers or high grade, ER-positive cancers) would be highly desirable. Unfortunately, these prediction problems seem to be more difficult [22,23]. It seems that fewer genes are associated with outcome in homogeneous disease subsets and the magnitude of association is modest when currently available datasets are analyzed. This leads to predictors that are specific for a particular dataset from which they were developed. These prediction models are fitted to the dataset and rely on features that have no or limited generalizability. This means that they fail to validate when applied to independent data or may demonstrate only nominally significant predictive value (that is, they may predict outcome slightly better than chance). Also, the discriminating value may not be substantial enough to be clinically useful [24,25]. For example, if the good-prognosis group has a recurrence rate of 30% compared with 50% in the poor-risk group, these may be significantly different but the risk of recurrence in the good-risk group is still too high to safely forego adjuvant chemotherapy.

Can we improve prediction through new technology platforms and improved bioinformatics tools?

It seems that for certain clinical prediction problems, the currently available breast cancer gene-expression datasets may not contain enough information to be able to develop highly accurate predictors [22,23]. This may reflect limitations of the sample sizes for the subsets of interest and, as more data become available, the empirically developed models may improve. However, it is also possible that major advances will need to take place in our understanding of how the 10,000 to 12,000 genes expressed in breast cancer interact before we can construct more accurate prediction models. Current statistical methods cannot readily adjust for different levels of gene-expression change that may be required for a functional effect. The level of expression change that results in a functional change may be different from gene to gene: for some genes a 15 to 20% increase in mRNA expression level may lead to functional consequences, whereas for others a 100 to 150% change may be needed.

New bioinformatics approaches, such as examining the information content of the correlation matrix of gene-expression values or applying network analysis tools to the data, may also reveal additional prognostic information that is not readily revealed by studying gene-expression levels alone. New analytical platforms, such as next generation sequencing, will generate more comprehensive expression data than the current array-based methods and will also yield extensive nucleotide sequence information. The information content of these currently nascent datasets may be highly relevant to prognosis or treatment response of cancers and certainly warrants further exploration.

Conclusions

The predictive performance of multi-gene signatures depends on the number and robustness of informative genes that are associated with the outcome to be predicted. Some clinically important prediction problems are easier to solve than others. For example, it is possible to predict the prognosis of ER-positive breast cancers relatively accurately because prognosis is closely related to the proliferative status of these cancers and proliferation affects the expression of several hundreds of genes that regulate and execute cell division. Not surprisingly, several different models that use different genes and different algorithms can be built with each performing similarly. On the other hand, predicting response to individual drugs based on gene-expression signatures has proved substantially more difficult. Fewer genes are significantly associated with these outcomes, measured on current analytical platforms (gene-expression arrays), and therefore prediction models invariably contain substantial amounts of 'noise' (predictive features that are specific to the dataset, not the actual outcome) and have poorer predictive performance on independent datasets. Larger datasets and new analytical platforms (such as next generation sequencing) that broaden the portfolio of variables that can be used for model building are expected to lead to improved predictors for these currently difficult classification problems.

Abbreviations
ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TI drafted the manuscript; LP reviewed and revised the manuscript. Both authors read and approved the final version of the article.

Published: 12 November 2010

References

1. Varadhachary GR, Talantov D, Raber MN, Meng C, Hess KR, Jatkoe T, Lenzi R, Spigel DR, Wang Y, Greco FA, Abbruzzese JL, Hainsworth JD: **Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation.** *J Clin Oncol* 2008, **26:**4442-4448.
2. Pusztai L, Ayers M, Stec J, Clark E, Hess K, Stivers D, Damokosh A, Sneige N, Buchholz TA, Esteva FJ, Arun B, Cristofanilli M, Booser D, Rosales M, Valero V, Adams C, Hortobagyi GN, Symmans WF: **Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors.** *Clin Cancer Res* 2003, **9:**2406-2415.
3. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS: **Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns.** *Cancer Res* 2001, **61:**5979-5984.
4. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
5. Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS: **Breast cancer molecular profiling with single sample predictors: a retrospective analysis.** *Lancet Oncol* 2010, **11:**339-349.
6. Sotiriou C, Pusztai L: **Gene-expression signatures in breast cancer.** *N Engl J Med* 2009, **360:**790-800.
7. Marchionni L, Wilson RF, Wolff AC, Marinopoulos S, Parmigiani G, Bass EB, Goodman SN: **Systematic review: gene expression profiling assays in early-stage breast cancer.** *Ann Intern Med* 2008, **148:**358-369.
8. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160-1167.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545-15550.
10. Reyal F, van Vliet MH, Armstrong NJ, Horlings HM, de Visser KE, Kok M, Teschendorff AE, Mook S, van 't Veer L, Caldas C, Salmon RJ, van de Vijver MJ, Wessels LF: **A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer.** *Breast Cancer Res* 2008, **10:**R93.
11. Bianchini G, Qi Y, Alvarez RH, Iwamoto T, Coutant C, Ibrahim NK, Valero V, Cristofanilli M, Green MC, Radvanyi L, Hatzis C, Hortobagyi GN, Andre F, Gianni L, Symmans WF, Pusztai L: **Molecular anatomy of breast cancer stroma and its prognostic value in estrogen receptor-positive and -negative cancers.** *J Clin Oncol* 2010, **28:**4316-4323.
12. Symmans WF, Hatzis C, Sotiriou C, Andre F, Peintinger F, Regitnig P, Daxenbichler G, Desmedt C, Domont J, Marth C, Delaloge S, Bauernhofer T, Valero V, Booser DJ, Hortobagyi GN, Pusztai L: **Genomic index of sensitivity to endocrine therapy for breast cancer.** *J Clin Oncol* 2010, **28:**4111-4119.
13. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr: **American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer.** *J Clin Oncol* 2007, **25:**5287-5312.
14. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thurlimann B, Senn HJ: **Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009.** *Ann Oncol* 2009, **20:**1319-1329.
15. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347:**1999-2009.
16. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365:**671-679.
17. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351:**2817-2826.
18. Bueno-de-Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, Wu TS, Atsma D, Rutgers EJ, Brink G, Floore AN, Glas AM, Roumen RM, Bellot FE, van Krimpen C, Rodenhuis S, van de Vijver MJ, Linn SC: **Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER).** *Lancet Oncol* 2007, **8:**1079-1087.
19. Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE, Sledge GW Jr, Perez EA, Shulman LN, Martino S, Sparano JA: **Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features.** *J Clin Oncol* 2008, **26:**4063-4071.
20. Cuzick J, Dowsett M, Wale C, Salter J, Quinn E, Zabaglo L, Howell A, Buzdar A, Forbes JF: **Prognostic value of a combined ER, PgR, Ki67, HER2 immunohistochemical (IHC4) score and comparison with the GHI recurrence score – results from TransATAC [abstract].** *Cancer Res* 2009, **69 Suppl:**74.
21. Viale G, Regan MM, Dell'Orto P, Mastropasqua MG, Rasmussen BB, MacGrogan G, Braye S, Orosz Z, Giobbie-Hurder A, Neven P, Knox F, Oehlschlegel C, Thuerlimann B, Coates AS, Goldhirsch A: **Central review of ER, PGR and HER2 in BIG 1-98 evaluating letrozole vs. letrozole followed by tamoxifen vs. tamoxifen followed by letrozole as adjuvant endocrine therapy for postmenopausal women with hormone receptor-positive breast cancer [abstract].** *Cancer Res* 2009, **69 Suppl:**76.
22. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD Jr, Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, *et al.*: **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nat Biotechnol* 2010, **28:**827-838.
23. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L: **Effect of training-sample size and classification difficulty on the accuracy of genomic predictors.** *Breast Cancer Res* 2010, **12:**R5.
24. Juul N, Szallasi Z, Eklund AC, Li Q, Burrell RA, Gerlinger M, Valero V, Andreopoulou E, Esteva FJ, Symmans WF, Desmedt C, Haibe-Kains B, Sotiriou C, Pusztai L, Swanton C: **Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials.** *Lancet Oncol* 2010, **11:**358-365.
25. Rody A, Holtrich U, Pusztai L, Liedtke C, Gaetje R, Ruckhaeberle E, Solbach C, Hanker L, Ahr A, Metzler D, Engels K, Karn T, Kaufmann M: **T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers.** *Breast Cancer Res* 2009, **11:**R15.

doi:10.1186/gm202
Cite this article as: Iwamoto T, Pusztai L: Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Medicine* 2010, 2(11):81.