Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives

Eytan Domany

Abstract

Accurate prognosis and prediction of response to therapy are essential for personalized treatment of cancer. Even though many prognostic gene lists and predictors have been proposed, especially for breast cancer, high-throughput "omic" methods have so far not revolutionized clinical practice, and their clinical utility has not been satisfactorily established. Different prognostic gene lists have very few shared genes, the biological meaning of most signatures is unclear, and the published success rates are considered to be overoptimistic. This review examines critically the manner in which prognostic classifiers are derived using machine-learning methods and suggests reasons for the shortcomings and problems listed above. Two approaches that may hold hope for obtaining improved prognosis are presented. Both are based on using existing prior knowledge; one proposes combining molecular "omic" predictors with established clinical ones, and the second infers biologically relevant pathway deregulation scores for each tumor from expression data, and uses this representation to study and stratify individual tumors. Approaches such as the second one are referred to in the physics literature as "phenomenology": they will, hopefully, play a significant role in future studies of cancer.

See all articles in this Cancer Research section, "Physics in Cancer Research."

Cancer Res; 74(17); 4612–21. ©2014 AACR.

Introduction

Technological advances made during the past 15 years have allowed measurement of massive amounts of molecular data from tumor tissue resected from a patient. These advances were expected to have significant impact on one of the central challenges of modern cancer research—providing patients with personalized prognosis and prediction of response to therapy. Frustratingly, the plethora of existing "omic" methods (see reviews in refs. 1–5) have not provided significant advance over prognosis based on classical clinical indicators (6), even though the acute need for improved prognostic tools was fully recognized and extensively addressed with several approved and commercially available products (see reviews for breast cancer in ref. 7, and for glioblastoma in refs. 8, 9).

Prognosis is highly relevant for early-stage breast cancer, because it affects the decision whether to subject a patient to chemotherapy, which is extremely important for the individual as well as for society for three main reasons. First, nearly all available chemotherapy adversely affects healthy tissue as well as the malignant one. Second, some side effects, even if they do not have a direct impact on the patient’s physical well-being, may cause considerable psychological damage and hardship. Finally, chemotherapy is very expensive. Indeed, the biomedical, clinical, and computational research communities joined forces and responded to the challenge, generating over the years a very large number of genomics-based prognostic signatures. Two questions come to mind: first, how come that, despite the enormous effort invested, no breakthrough has yet been achieved? And second, if there is hope for progress, in which direction does it lie?

The aim of this review is to answer these two questions. About the first, I provide the reader with a simple intuitive step-by-step description of the manner in which many prognostic signatures are derived, using machine-learning approaches. This description is presented in a highly critical manner, emphasizing the potential failures and pitfalls of each of the steps taken. To be specific, I use one of the first and most acclaimed prognostic gene lists, the 70-gene “Amsterdam signature” of van ’t Veer and colleagues (10), as a concrete example. This signature received U.S. Food and Drug Administration (FDA) approval in 2007; a commercial product (MammaPrint) is based on it, and a large-scale validation effort was initiated (11). The methodology that was used to derive it characterizes an entire class of machine-learning approaches, to which I refer as “ignorance based.” A list of prognostic genes is selected from high-throughput data (e.g., expression), with very little of the available biological knowledge used in the process. Because the number of available samples is much less than the number of variables (genes) measured for each sample, these methods encounter a technical difficulty called “the curse of dimensionality” (12).
As to the second question, I focus on two directions that hold, in my opinion, promise. The guiding principle of both is to build on existing knowledge (13).

The first direction incorporates biological understanding to produce a signature in terms of pathways and biological processes (14–16). I will explain the method (17) and some of its results, but the main message should be one of principle. This study is an example of an approach to which physicists refer to as “phenomenology.” The general idea is to

a. make use of existing knowledge (as opposed to “ignorance based” methods);
b. try to “separate scales”—identify and retain “coarse-grained” variables that are essential for the problem and that are experimentally accessible (as opposed to trying to incorporate as many molecular details and entities as possible); and
c. formulate, in terms of these variables, a “minimal,” e.g., simplest, mathematical model that does capture the complexity of the problem.

This approach is deeply rooted in all branches of physics; one of its most eloquent advocates, Y. Frenkel, was quoted by I. Tamm (ref. 18, p. 183):

“...the more complicated the system considered, the more simplified must its theoretical description be. One cannot demand that the theoretical description of a complicated atom, and all the more of a molecule or a crystal, have the same degree of accuracy as the theory of the simplest hydrogen atom. Incidentally, such a requirement is not only impossible to fulfill, but also essentially useless.” (emphasis added)

The second direction is more straightforward and advocates combining “omic” signatures with existing clinical ones, rather than replacing them (19); see also ref. 3.

Machine-Learning Derivation of Prognostic Signatures

The “Amsterdam signature” was introduced in a pioneering study (10), which measured, using Rosetta microarrays, the expression levels of thousands of genes for $N = 97$ fresh-frozen, small (<5 cm) lymph-node negative sporadic (without BRCA1/2 germline mutations) tumors resected from young (age <55 years) patients for whom 5-year clinical follow-up was available. All such studies strive for genetic and clinical homogeneity, for example, all patients are sporadic and treated similarly, to eliminate confounding factors that may affect prognosis; in the van ’t Veer cohort only 3(2) patients received chemo (hormonal) therapy. The main difficulty and bottleneck, then and now, is to collect samples and to ensure clinical follow-up from a large enough cohort of patients that satisfy these criteria.

Expression data were used to perform an unbiased search for genes with prognostic value. Each patient was assigned an outcome indicator—1 for good outcome (no distant metastasis during the 5-year follow-up period) and 0 for bad outcome. First, 70 genes were selected on the basis of the correlation of their expression levels with the outcome indicators. Next, a classifier was constructed. van ’t Veer and colleagues (10) calculated for each of the 70 selected genes its average expression level over the good-outcome patients, which formed their “good outcome signature.” A new patient (of unknown outcome) is classified on the basis of the correlation of its expression profile (on the same 70 genes) with this signature. If the correlation exceeds a threshold, the patient is predicted to have good prognosis. I examine, in the form of questions and answers, the way in which the signature was derived and its prognostic and scientific significance.

1. **How was the number of prognostic genes derived and why did the number 70 come out?** Out of the 97 patients, van ’t Veer and colleagues selected at random 78 (44 good and 34 bad outcome), which were assigned to a training set, and another group, of 19 patients (7 good, 12 bad outcome), formed a test set. The test set is now “hidden”—the data from these tumors should not be used until the very end of the process. The first step is “feature selection” (20)—out of thousands of genes represented on the microarray, choose a subset whose expression levels are to be used for prognosis. Feature selection is a must whenever the number of parameters (genes) exceeds the number of datapoints (samples). For every gene $g$, the correlation $C(g)$ of its expression with the outcome indicators was calculated, and the genes were ranked according to the absolute values $|C(g)|$. The top-ranked $N_g$ genes of this list are used for prognosis. To determine $N_g$ van ’t Veer and colleagues constructed a sequence of classifiers; first, the top 10 genes were used, and then, the next 5 high-ranked ones were added and so on. At each step, they performed a leave-one-out cross-validation procedure (12); use 77 training samples to construct the prognostic profile, leaving out one sample for which outcome is predicted. This was done 78 times, and from the 78 predictions, the error rate (fraction of wrongly classified patients, out of those “left out”) was measured. Initially, the error rate decreased as genes were added, but when their number increased beyond $N_g = 70$, the error rate fluctuated and even increased. This is how the number of prognostic genes was derived. As to why did $N_g$ come out to be 70—this value was most probably dictated by the number of samples (77) that were used for training. Had van ’t Veer and colleagues had more available training samples, they probably would have produced a predictive list of more than 70 genes; hence, the number of genes that form the “Amsterdam signature” is neither optimal nor biomedically justified.

2. **How robust is this gene list?** This question came up as a result of the negligibly small overlap between these 70 and previously (21, 22) or subsequently (23) identified lists of prognostic genes. The most prevalent explanation offered (24) was that the small overlap [e.g., three genes between the Amsterdam and Rotterdam (23) signatures] was not surprising as the studies used different microarray platforms and cohorts of somewhat different clinical

---

**Transcriptomic Data for Prognosis**

First, 70 genes were selected on the basis of the correlation of their expression levels with the outcome indicators. Next, a classifier was constructed. van ’t Veer and colleagues (10) calculated for each of the 70 selected genes its average expression level over the good-outcome patients, which formed their "good outcome signature." A new patient (of unknown outcome) is classified on the basis of the correlation of its expression profile (on the same 70 genes) with this signature. If the correlation exceeds a threshold, the patient is predicted to have good prognosis. I examine, in the form of questions and answers, the way in which the signature was derived and its prognostic and scientific significance.

1. **How was the number of prognostic genes derived and why did the number 70 come out?** Out of the 97 patients, van ’t Veer and colleagues selected at random 78 (44 good and 34 bad outcome), which were assigned to a training set, and another group, of 19 patients (7 good, 12 bad outcome), formed a test set. The test set is now "hidden”—the data from these tumors should not be used until the very end of the process. The first step is "feature selection" (20)—out of thousands of genes represented on the microarray, choose a subset whose expression levels are to be used for prognosis. Feature selection is a must whenever the number of parameters (genes) exceeds the number of datapoints (samples). For every gene $g$, the correlation $C(g)$ of its expression with the outcome indicators was calculated, and the genes were ranked according to the absolute values $|C(g)|$. The top-ranked $N_g$ genes of this list are used for prognosis. To determine $N_g$ van ’t Veer and colleagues constructed a sequence of classifiers; first, the top 10 genes were used, and then, the next 5 high-ranked ones were added and so on. At each step, they performed a leave-one-out cross-validation procedure (12); use 77 training samples to construct the prognostic profile, leaving out one sample for which outcome is predicted. This was done 78 times, and from the 78 predictions, the error rate (fraction of wrongly classified patients, out of those "left out") was measured. Initially, the error rate decreased as genes were added, but when their number increased beyond $N_g = 70$, the error rate fluctuated and even increased. This is how the number of prognostic genes was derived. As to why did $N_g$ come out to be 70—this value was most probably dictated by the number of samples (77) that were used for training. Had van ’t Veer and colleagues had more available training samples, they probably would have produced a predictive list of more than 70 genes; hence, the number of genes that form the "Amsterdam signature" is neither optimal nor biomedically justified.

2. **How robust is this gene list?** This question came up as a result of the negligibly small overlap between these 70 and previously (21, 22) or subsequently (23) identified lists of prognostic genes. The most prevalent explanation offered (24) was that the small overlap [e.g., three genes between the Amsterdam and Rotterdam (23) signatures] was not surprising as the studies used different microarray platforms and cohorts of somewhat different clinical
characteristics. This explanation was proven to be incorrect (25, 26). Ein-Dor and colleagues (25) repeated the van 't Veer and colleagues' procedure on the van 't Veer and colleagues data by repeatedly selecting at random different training sets of 78 samples (out of the available 97). For each training set, the genes were ranked according to their correlations with outcome, and a corresponding list of 70 top-ranked genes was created. The typical number of genes shared by two such lists was 12. When two groups of 78 are selected out of 97 patients (even with bootstrap ref. 27), many patients will appear in both groups; hence for truly different samples, the overlap is smaller (28). Because the different lists of top-ranked genes were now derived using data from the same platform and patients, one cannot attribute the low overlap between lists to these differences! The procedure used by van 't Veer and colleagues to generate their list of 70 prognostic genes suffers from an inherent instability and lack of robustness.

3. **Do different prognostic gene lists reflect the same biology?** The criticism voiced above was fended off initially by some (29), claiming that even though the specific genes on the different lists were not the same, they did represent the same pathways and biological processes. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (15) assigns 88 genes to the apoptotic pathway. Deregulation of the pathway is reflected in the expression levels of some of these genes, and it is possible that one prognostic signature contains one group of apoptosis-related genes and the other signature contains a different group. To check whether two gene lists indeed capture the same pathway, one has to determine whether the genes of this pathway are overrepresented, in a statistically significant manner, in both signatures. Such an analysis (30), which took into account the false discovery rate (31) associated with testing multiple genes and pathways, as well as the nested nature (32) of the gene ontologies used, and avoided other pitfalls (33), has shown that the Amsterdam and Rotterdam signatures share only a single overrepresented biological process—cell proliferation. On the one hand, this is good news—proliferation is one of the most well-established prognostic indicators; the mitotic index (34) and other proliferation-sensitive biomarkers (35) have been measured and used for prognosis for many years, and comfortingly (36, 37), it is captured by both signatures. On the other hand, do we need complex and expensive gene expression signatures to measure cell proliferation (38, 34)?

4. **Perhaps the 70-gene Amsterdam signature luckily does provide the best prognostic results?** Ein-Dor and colleagues (25) addressed this question first by using the same ranking of the genes as van 't Veer and colleagues (the top 70 constitute the van 't Veer signature) and checking the prognostic value of different consecutive gene lists; those ranked 71–140, then 141–210, and so on, until 701–770. Kaplan–Meier analysis of seven such lists, together with the van 't Veer list, is presented in Fig. 1. The separation (into good and bad outcome) of the classifier based on the top 70 is no better than those of much lower ranked 70-gene based classifiers. Classifiers 1 and 3 (25) had the same error rates on the test set as the van 't Veer list, and the others were only slightly worse. There is a simple reason for this; the samples contain estrogen receptor–positive (ER”) and ER-negative (ER”) tumors, and the expression profiles of these two cancer types are so different that most random sets of 70 expressed genes will capture it. Because these two diseases also have very different outcome, such a random gene list provides a prognostic signature as well. Indeed, subsequently, it has been shown (39) that randomly selecting 70 expressed genes performed no worse than most published prognostic predictors (40). Michiels and colleagues (26) calculated the success rates (on the test set) for a large number of “top-ranked” gene lists, generated in the manner described above, and determined the distribution of these success rates. They did this for many prognostic signatures that were published for different types of cancer and have shown that the reported success rates of nearly all studies corresponded to very unlikely tails of these distributions; the probability to obtain the reported success rate was significantly lower than 0.05. The Amsterdam signature has just made the 5% chance of being generated. Hence, the prognostic signatures were described (26) as “overoptimistic.” This has been attributed to “information leak”—the situation when the test set (or part of it; ref. 41) is used in some manner in the course of the training process. The most common case is when the researcher tests his predictor and is unsatisfied with the error rate (on the test set), he realizes that something could have been done differently—low-level processing of the raw data, for example—and he repeats the whole process, improving the performance on the test set (which no longer is a bona fide test set, as it has been already used!). Dupuy and Simon (42) looked carefully at 40 such studies and discovered some kind of information leak in about half of them.

5. **Does the prognostic predictor provide significant improvement over existing methods?** This is the last question posed by defenders of the various prognostic lists—implying that if the answer is affirmative, why should one care about robustness, biological interpretation, etc.? To get a fair answer to this question, one must use the clinically accepted requirements from the prognostic predictor. There are two types of errors—the first one is the case when a bad-outcome patient, who could benefit from chemotherapy, is erroneously classified as good outcome. The thresholds of the classical criteria (43–45) are set to keep this error at less than 10% (sensitivity >0.90). Given this constraint, one measures the specificity, i.e., the percentage of correctly identified good-outcome patients. The initial van 't Veer classifier reported (at sensitivity >0.90) a specificity of 0.75 for all 97 patients (and 0.72 for the test set only), far better than the NIH criteria (0.12). A validation study (46) involved 295 patients (that, however, included 61 of the
original van 't Veer cohort, which is an "information leak"; ref. 41). The specificity obtained for 180 new patients (with known 5-year follow-up of metastasis) dropped to 0.53 (6). From the reported figures of a larger study (on 307 independent patients; ref. 47), one can derive a specificity of 0.42 (6). Specificity decreases as a function of sample size, indicating that for a large population, performance is likely to be below that of the non-omic methods. Even the current figure (0.42) is only slightly better than that obtained for the same patients by Adjuvant! (48), a software that uses clinical parameters (age, general health, ER status, histologic grade, tumor size, and lymph node involvement) to calculate the chance of survival under various modalities of therapy.

6. **What is the reason for the lack of robustness of the prognostic gene lists?** There are two answers to the question, mathematical and biological. The mathematical reason is the inherent instability of the ranking process (25, 26). The rank of a gene \( g \) was set on the basis of the correlation of its expression with outcome, \( C(g) \). This number was calculated using the samples in the training set, which were randomly selected; hence, \( C(g) \) is a random variable. Imagine that we have access to samples from all the patients with node-negative breast cancer in the world, select \( n = 78 \) out of them as our training set, and we repeat this 1,000 times, generating 1,000 sets of 78 samples. For a given gene, we calculate \( C(g) \) for every one of these 1,000 training sets; clearly, we get 1,000 different values, distributed around some mean with variance \( s^2_n(g) \). The value of \( s^2_n(g) \) is \( \sqrt{n} \), where \( n \) is the number of samples used to calculate the correlation. For a heterogeneous disease, \( C(g) \) will change a lot when correlation is recalculated on the basis of a different group of patients (28)—this is the biological reason mentioned above. For the data of van 't Veer and colleagues, for \( n = 78 \) we have \( s^2_n(g) \approx 0.1 \) for all genes (25), so when we change the training set, the \( C(g) \) changes by this much. On the other hand, the top 1,000 genes have \( C(g) \) between 0.25 and 0.45 (25), an interval of 0.2; therefore, a change of 0.1 in \( C(g) \) will typically shift a gene's rank by a few hundred! Hence,
Pathway-Based Analysis Holds Promise for Better Prognosis

I review here our approach (17), which does use existing knowledge, avoiding the “curse of dimensionality” and the need for inherently unstable procedures (such as ranking genes). The general philosophy and approach is phenomenological: use existing knowledge, work with coarse-grained variables, constructing minimal, simple models that capture the essential aspects of the problem, ignoring (or treating in a heuristic manner) a multitude of details. This approach differs from that used in ref. 53, which relies on detailed connectivity information between tens of thousands of entities, including protein abundances and phosphorylation status. For the description of several other pathway-based analyses (53–60), see ref. 17.

We use high-throughput expression data, but focus on small groups (tens) of genes at a time. Such a group contains genes whose expression levels have been found (14–16) to be related to a particular biological process or pathway. \( P \). The number of samples—usually hundreds—exceeds the number of genes in the group (typically, tens), avoiding the curse of dimensionality and generating robust results, much less sensitive to noisy measurements and to heterogeneity of samples than ranking procedures of single genes. Analysis yields for each particular sample \( k \) a score, \( D(P,k) \), which represents the extent to which pathway \( P \) is deregulated or activated in the tumor of patient \( k \). These pathway deregulation scores (PDS) are calculated for all relevant pathways, one at a time. A patient is represented now by a set of such PDSs. On the basis of these scores, patients are assigned to groups of similar pathway deregulation profiles and pathways of prognostic value are identified.

The manner in which the PDS is calculated is outlined in Fig. 3 for the apoptotic pathway (15); 33 of its 88 genes are variably expressed in our colon cancer data (61). The expression value of each of these genes constitutes one of 33 coordinates, and every sample \( k \) is represented as a point \( X_k \) in this 33-dimensional space. Because it is difficult to display points in 33 dimensions, Fig. 3A presents 313 samples (61) projected onto a three-dimensional subspace. The colors of the points identify the corresponding sample as normal colon tissue, polyp, primary or metastatic tumor. A principal curve (62) is passed through the cloud of points—it represents the main (nonlinear) variation of the data. Next (Fig. 3B), each point \( X_k \) is projected onto the principal curve: tumor \( k \) is represented now by \( Y_k \), its image on the curve, Fig. 3C. Because normal samples came out at one end, then polyps, primaries and metastases, for this pathway \( P \), the variation captured by the principal curve clearly corresponds to disease progression. The PDS of each tumor \( k \) is given by the distance, \( D(P,k) \), measured along the principal curve from the “normal endpoint” to \( Y_k \). The larger this distance, the higher is the deregulation of the tumor.

Several assumptions implicitly underlie the approach. First, there is a most important kind of variation due to some process (such as progression) that characterizes the expression changes among the samples. The trajectory of the tumors in expression space of a pathway can be attributed to this process, with added noise. There is no reason to assume that this trajectory is along a straight line; in colon cancer, one set of...
genes may vary as we move from normal tissue to polyps, another group changes from polyps to primary tumors, and yet a third set of genes modifies its expression levels upon transition to metastasis. The resulting nonlinear trajectory in expression space is approximated by the principal curve. Because deregulation of a pathway can be due to increased expression of some genes while others decrease, it makes little sense to look only for significant shifts of the mean expression of the genes of a pathway, or just count the number of the genes that distinguish, say, tumor from normal tissue. Deregulation is estimated by a distance, measured "along the path traveled"—this is a heuristic, phenomenological assumption.

Caveats, shortcomings, and how they are addressed

The first and most obvious criticism is that only mRNA expression data are used. Many potentially important variables are left out—protein concentration, posttranslational modifications, microRNA abundance; genetic variants such as point mutations, translocations, DNA copy-number changes; epigenetic variables such as histone modifications and DNA methylation. Integration of these signals will improve the precision, reliability, and robustness of the derived PDSs. There are several answers to this criticism. The first is that expression data are the most abundant, i.e., available for the largest numbers of samples. Most of the other data types lag behind, and their inclusion is at the price of working with a much lower number of samples. Second, the main idea is to parametrize a tumor by coarse-grained, biologically meaningful variables (the PDSs), and expression profiling does provide a window on the PDS; deregulation of most pathways in a tumor is reflected in changing expression levels of the associated genes. Whether the PDSs based on expression alone are good enough will be known only a posteriori, once the method is used on real data, and the resulting new predictions and stratifications are tested.

Another critique concerns the noisy, nonrobust nature of mRNA expression data. Our use of coarse-grained (versus single-gene) variables is part of our answer. In addition, we insist on applying the method for every cancer type on more than one dataset and trust only those findings that were seen in more than one.

Finally, assignment of genes to pathways and biological processes is far from being perfect. To (partially) alleviate this concern, we worked with three different pathway databases (14–16).

Selected results

The method was applied (17) to glioblastoma, colon cancer, and (63) breast cancer expression data (64). In all cases, the results made biological sense and, as opposed to single-gene-based prognostication, were robust. Analysis of the same disease on the basis of different datasets reproduced the main findings; for example, two lists of pathways that exhibited correlation with outcome in two distinct datasets [The Cancer Genome Atlas (TCGA; ref. 65) and REMBRANDT (66)] had 50% overlap.

To appreciate the prognostic potential of the approach, note that expression-based stratification (67) divides glioblastoma to one of four classes: classical, mesenchymal, neural, and proneural. Patients with proneural tumors live longer. As opposed to unsupervised analysis of expression data (65), which does not yield prognostically meaningful substratification of the neural/proneural group, our PDS-
based analysis discovered several subgroups of the four known tumor types. The heatmap of Fig. 4 shows the PDSs of 548 pathways in the 455 TCGA samples, reordered by hierarchical clustering. Kaplan–Meier analysis showed (see Fig. 5) that patients from TgS15, one subgroup of proneural tumors, lived for much longer; a second, TgS13, had moderately long survival, and a third, TgS12, had much worse outcome. When the three special subgroups are removed, the survival curves of the remaining neural/proneural patients do not differ from that of the other types of glioblastoma. Unsupervised discovery of these subtypes on the basis of our PDS (but not when "raw" expression data are used) indicates the potential applicability of the method for prognosis in glioblastoma.

Combining Molecular and Clinical Signatures

The second promising direction is to combine prognostic predictors based on classical clinical variables (patient age, tumor size, grade, node involvement, etc.), with biomarkers measured by immunohistochemistry and with expression-based signatures. One of the many publications that described such an approach is the DREAM project, based on a competition between authors of different prognostic algorithms that train their tools on the same set of samples and submit their software to a panel of referees who test them on a (hidden) validation set. According to the report on this competition (3), many of the leading strategies were based on combining clinical and molecular variables.

A second recent study (19) tested combinations of a clinical treatment score (CTS) with IHC4 (68), a score based on four immunohistochemically measured markers, and with two expression-based signatures: the OncotypeDX recurrence score (RS; ref. 69) and the PAM50 (70) risk of RS (ROR).
OncoTypeDX (69) received FDA approval at the same time as MammaPrint and is the most used method in the clinic. It is not an "omic" method; it was derived in a knowledge-based manner. Sixteen genes that take part in biological processes that are involved in cancer were selected on the basis of the prognostic values of their expression levels. These were measured together with five transcripts used for calibration by quantitative real-time PCR (qRT-PCR). An RS, derived from the resulting 21-gene signature, is used to indicate whether the patient needs chemotherapy. Most importantly, the technique uses formalin-fixed paraffin-embedded (FFPE) blocks, which are routinely produced from solid tumors, while obtaining fresh-frozen samples (needed for MammaPrint) requires special effort.

PAM50 (70) builds on one of the first microarray-based classification schemes of breast cancer (21). The basic "intrinsic" stratification of breast cancer to luminal A/B, basal, HER2+, and normal-like subtypes has withstood the tests of time, and the prognostic significance of this classification is clear. Parker and colleagues (70) have selected 50 genes whose expression levels can be reliably measured on FFPE samples by qRT-PCR and constructed a classifier that successfully assigns a tumor to one of the intrinsic subtypes, which is then fed into a simple prognostic predictor (using also tumor size) to yield an ROR score.

IHC4 (68) is based on a score constructed out of four immunohistochemical measurements: ERs and PRs, human epidermal growth factor receptor 2 (HER2), and Ki67, a protein associated with cell proliferation. Dowsett and colleagues (19) tested each of these predictors on their own and in combinations on postmenopausal patients with hormone receptor–positive primary breast cancer treated with Arimidex, tamoxifen alone, or combined. The cohort had a mean age of 54 years, with tumors of mean size 18 mm (equal numbers of node-positive and node-negative patients). Their conclusions were the following:

a. On all samples, the predictor based on clinical information alone (CTS) performed significantly better than ROR, which was better than RS. Combining ROR with CTS outperformed CTS alone; this was the best predictor, better than RS with CTS.

b. On two subgroups (all node-negative patients and node negative/HER2+), the best single predictor was ROR, followed by CTS, and then RS. Combining ROR with CTS was better than CTS and, again, better than RS with CTS.

c. Testing IHC4 required sample splitting, and IHC4 combined with CTS was slightly better than ROR + CTS on all patients and on the node-negative group, while the relative merit was reversed on the node-negative/HER2− group. For all groups, the RS+CTS combination was least favorable.

To summarize, the best overall single predictor is the one based on clinical variables; adding molecular signatures improves performance. In all cases, combination of clinical and molecular signatures outperformed each classifier on its own.

Discussion

This review has two main messages—one, on the clinical utility of the currently available molecular prognostic predictors and, the other, on the potential influence of physics on this field.

About the first message, it is important to realize that clinical usage of prognostic signatures is not governed by their scientific quality and success rate. For example, the possibility to perform measurements on FFPE samples (versus fresh frozen) confers considerable advantage on a method. Introduction of prognostic tests to the clinic is also governed by financial issues, such as cost and the willingness of insurance carriers to pay for them. Proprietary aspects also enter as confounding variables.

Staying within scientific perspective, I strongly believe in the utility of molecular predictors that do use existing knowledge (as opposed to ignorance-based brute force methods). I also believe in combining clinical predictors with molecular ones, and I am hopeful about using coarse-grained representations (such as PDS) for prognosis.

Turning to the role of physics in all this, it should be noted that using molecular signatures for prognosis is not a physics problem. Rather, it is a central issue in clinical cancer research. So what can physics (and physicists) contribute? There was one technical point mentioned above in which methods of theoretical physics were used, but, in my view, the main potential contribution of physicists lies elsewhere in legitimizing phenomenology. The approach that suppresses most "atomistic" details and looks for separating scales and for the "right" coarse-grained variables that describe a problem is deeply rooted in physics, while it is less accepted in biology. I believe that this way of thinking does have considerable potential, and its implementation in different biological and clinical contexts, including prediction of response to therapy in cancer (71), may be the most important contribution of physicists to the field.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The author worked in collaboration with Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, Oly Zuk, Carolyn Fuku, Yotam Drier, Michal Sheffer, and Anna Livshits.

Grant Support

This work was supported by the Leir Charitable Foundation and a Germany–Israel Cooperative Research grant (DIP).

Received November 20, 2013; revised January 21, 2014; accepted February 10, 2014; published online September 2, 2014.


Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives

Eytan Domany


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/74/17/4612

Cited articles
This article cites 67 articles, 27 of which you can access for free at:
http://cancerres.aacrjournals.org/content/74/17/4612.full#ref-list-1

Citing articles
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/74/17/4612.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.