

# Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction

Joseph R. Nevins<sup>1,2,6,\*</sup>, Erich S. Huang<sup>1,6</sup>, Holly Dressman<sup>1,6</sup>, Jennifer Pittman<sup>5,6</sup>, Andrew T. Huang<sup>3,4,6</sup> and Mike West<sup>5,6</sup>

<sup>1</sup>Department of Molecular Genetics and Microbiology, <sup>2</sup>Howard Hughes Medical Institute and <sup>3</sup>Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA, <sup>4</sup>Koo Foundation-Sun Yat Sen Cancer Center, Taipei, Taiwan, <sup>5</sup>Institute of Statistics and Decision Sciences, Duke University and <sup>6</sup>Computational and Applied Genomics Program, Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

Received August 10, 2003; Revised and Accepted August 11, 2003

---

Genomic data, particularly genome-scale measures of gene expression derived from DNA microarray studies, has the potential for adding enormous information to the analysis of biological phenotypes. Perhaps the most successful application of this data has been in the characterization of human cancers, including the ability to predict clinical outcomes. Nevertheless, most analyses have used gene expression profiles to define broad group distinctions, similar to the use of traditional clinical risk factors. As a result, there remains considerable heterogeneity within the broadly defined groups and thus predictions fall short of providing accurate predictions for individual patients. One strategy to resolve this heterogeneity is to make use of multiple gene expression patterns that are more powerful in defining individual characteristics and predicting outcomes than any single gene expression pattern. Statistical tree-based classification systems provide a framework for assessing multiple patterns, that we term metagenes, selecting those that are most capable of resolving the biological heterogeneity. Moreover, this framework provides a mechanism to combine multiple forms of data, both genomic and clinical, to most effectively characterize individual patients and achieve the goal of personalized predictions of clinical outcomes.

---

## INTRODUCTION

Recent advances in genome science and technology define the potential for increasingly complex biomedical and molecular information to underlie a coherent system of personalized medicine—health planning, treatment strategies and drugs customized to the individual patient rather than broader population cohorts. The value in genomic data is its scale and complexity; when combined with clinical and demographic factors, multiple forms of molecular data provide information that has the potential to identify unique characteristics of the individual and so lead to customized health care strategies. Thus, rather than having access to a limited number of data inputs that only broadly define individual characteristics, it is conceivable that the entire genetic profile and time-sensitive genomic characteristics of an individual will be available to aid

in substantially improved determinations of individual disease susceptibilities, likely responses to therapies and other clinical outcomes. Near-term advances in genome science will be key in advancing us towards this goal, providing access to increasingly comprehensive and precise molecular data. A key challenge lies in the need to define analytic methods and tools to synthesize, integrate and interpret such increasingly complex data in order to bring it to bear in personalized prognostic and diagnostic settings.

One good example arises in the treatment of cardiovascular disease. Current strategies rely on a cocktail of drugs of proven efficacy; however, most patients benefit from only a few of the five or so drugs that are commonly used, and may indeed have negative side effects from several of the drugs. Thus, an ability to identify which drug or combinations of drugs is most effective for any individual, and to effectively also predict the

---

\*To whom correspondence should be addressed. E-mail: j.nevins@duke.edu

side-effects responses of that individual, will have significant impact on the treatment decisions and strategies.

Cancer is another disease in which individualized treatment is key. A woman diagnosed with early-stage breast cancer will undergo surgery for removal of the tumor and then, typically, be treated with adjuvant chemotherapy. Nevertheless, many such women—inherently lower risk cases—unnecessarily undergo the harsh reality of chemotherapy; such women might be spared this experience were reliable and precise predictors of their longer-term relapse-free condition to be available. Traditional clinical risk factors—such as tumor size, patient age, tumor involvement in lymph nodes and hormone receptor status—certainly have prognostic value in connection with disease progression and the prospects for recurrence. However, the information arising from such factors is nowhere near as precise and accurate as is needed to reliably identify those individuals who will require and benefit from therapy from those who will not. It is largely as a result of this lack of our ability to focus in on individuals with customized predictions, and the resulting high degree of uncertainty about outcome at the individual level, that many otherwise lower risk women currently undergo aggressive therapy, with its concomitant morbidity.

Genomic information, in the form of massive profiles of gene activity (gene expression) within tumor samples, has in recent years demonstrated the capacity to identify characteristics that reflect tumor behavior and that relate to disease progression and outcomes, including cancer recurrence. Tumor-based gene expression data from DNA microarrays adds immense detail and complexity to the information available from traditional clinical and pathological sources; it is a snapshot of the total gene activity of the tumor, providing complex and detailed data on both the inherent genetic state of the patient and on the current characteristics of the tumor and disease state. The potential is then for this information to substantially improve the accuracy with which we can predict the likely development of disease process for *this* patient from this point on; such improved predictions will critically aid clinical decision making at a level of individualization that is currently unachievable.

## GENE EXPRESSION PROFILES AND PREDICTING BREAST CANCER OUTCOMES

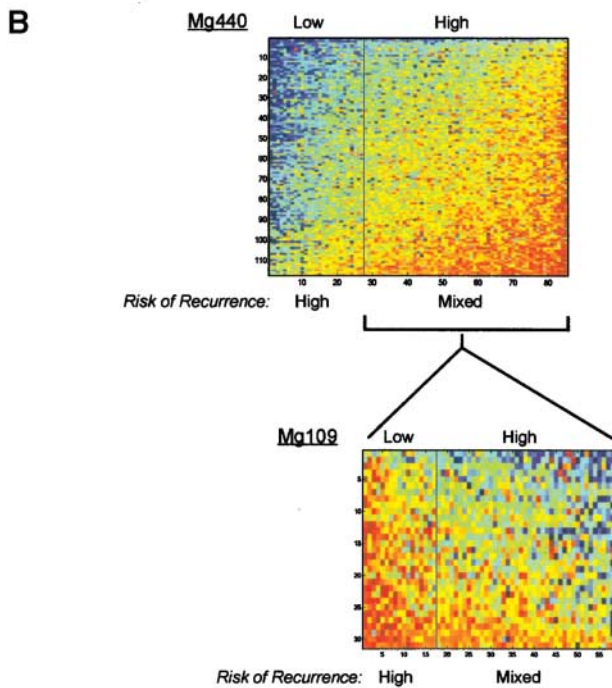
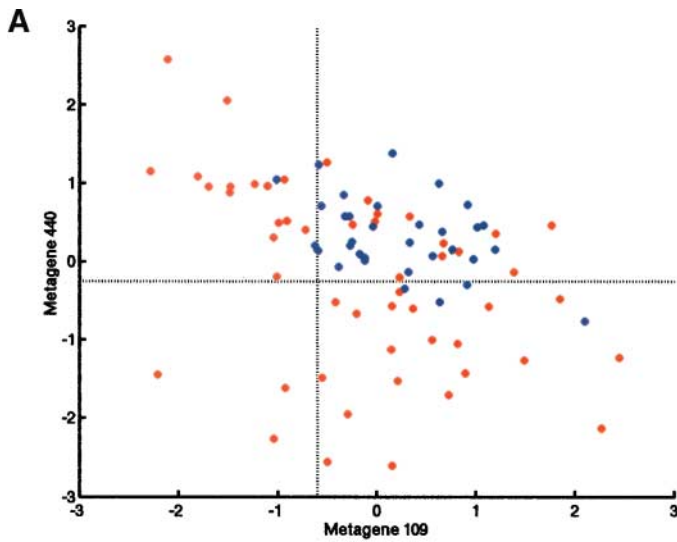
Several studies have now reported the ability to make use of gene expression patterns to classify and sometimes to predict disease outcomes in cancer patients and other disease contexts (1–8). Most of these studies do not, however, go much beyond the traditional classification of patient populations into broad risk groups based on the values of a defined gene expression predictor. One such study (6,7), for example, uses a gene expression predictor, based on a collection of 70 genes, to classify breast cancer patients into high-risk and low-risk categories relative to long-term recurrence. Individuals classified as high-risk are, however, simply assigned a 50% recurrence-free survival probability. For a woman in this group, much more is needed to refine and customize the prognosis based on other individual clinical, genetic and genomic factors. This initial grouping displays the power of gene expression data to achieve a broad patient stratification that points to

accurate predictions of disease outcome, but the resulting subgroups remain quite heterogeneous. What is needed is more precise delineations of patients into subgroups that are more homogeneous with respect to disease outcomes, so that a future patient may be matched much more accurately with past patients with closely related risk profiles. This notion was a key motivation for our approach (8,9) to defining statistical models for personalized prognosis: continuing the stratification of patients to define finer sub-categorization by collections of risk factors. We also hold the view that, to aid this refined analysis, *multiple* summaries of gene expression profile should be brought to bear—array-based expression profiles from tumors carry information on multiple aspects of tumor biology that may have prognostic value. In our breast cancer studies, we do indeed find that multiple gene expression signatures—weighted average measures of expression of defined groups of genes that we term *metagenes*—are capable of defining such refined patient stratification, and that the combination of such signatures with traditional clinical factors can lead to more accurate predictions. Importantly, we define predictions via outcome probabilities that are specific to individuals in finer patient subgroups defined by interacting patterns of risk factors—a key step towards personalized predictions.

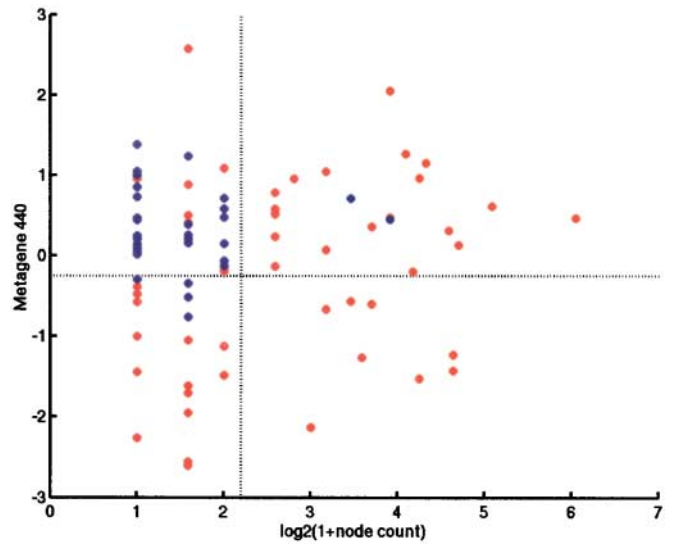
An illustration of the importance of going beyond traditional single stratification of patients, and the value of using multiple predictors of patient outcome—in this case multiple metagene expression patterns—is depicted in Figure 1. Gene expression measures were generated from tumors of 86 lymph node positive breast cancer patients. Metagenes were created from the data and used to stratify the patients. Among 500 metagenes created in this analysis, a small number are implicated as key predictors of survival; two are used for this display (metagenes 440 and 109). A first stratification on metagene 440 achieves only a partial separation of the patients based on risk of recurrence—those with a low value on metagene 440 (below the horizontal line) are generally high-risk cases, depicted as red symbols (Fig. 1A). In contrast, those with a high value on this metagene (above the line) are heterogeneous with respect to recurrence. A second metagene resolves some of the heterogeneity of this group—those with a low value on metagene 109 (and high 440) are largely high-risk while those with a high value on metagene 109 (and high 440) are still heterogeneous but now enriched for low-risk cases. This remaining heterogeneity can be further evaluated based on other factors. The actual gene expression profiles that underlie these metagenes are shown in Figure 1B. These graphics thus highlight the importance of multiple predictors in resolving patient heterogeneity.

## CLINICO-GENOMIC MODELS FOR PERSONALIZED DISEASE OUTCOME PREDICTION

The use of multiple predictors of clinical outcome need not be limited to any one form of data, whether genomic or otherwise. Indeed, our recent work has demonstrated that, beyond the relevance of multiple interacting gene expression signatures, it is the combination of genomic data with some traditional clinical risk factors that currently define the most accurate predictions of breast cancer recurrence. The extent of lymph node metastasis,



**Figure 1.** (A) Scatter plot of 86 lymph node positive breast cancer samples on two metagenes related to cancer recurrence, showing 5-year recurrence-free survivors (blue) and cases of recurrence within 5 years (red). Note that a first stratification by a threshold on metagene 440, as indicated, partitions samples into a clearly high-risk group (low metagene 440) and a group of mixed cases (high metagene 440). Subsequent partitioning of this latter group, according to a threshold on metagene 109, defines a further high-risk subset (high metagene 440 coupled with low metagene 109), leaving the (high metagene 440, low metagene 109) group remaining to be further evaluated based on other metagenes and clinical factors. This illustrates the ability of tree-like partitioning methods to capture non-linear interactions between potential predictors of clinical outcomes, and highlights the importance of multiple predictors in resolving patient heterogeneity. (B) Gene expression intensity images for genes defining metagenes 440 (consisting of 117 genes) and 109 (31 genes) in the node-positive breast cancer samples. Each column represents one patient, ordered by the value of the metagene. In the case of metagene 440, a threshold as indicated in (A) corresponds to a partition of patients as indicated here. A refined sample partition can then be defined on metagene 109 for those cases of 'high metagene 440' as also indicated.



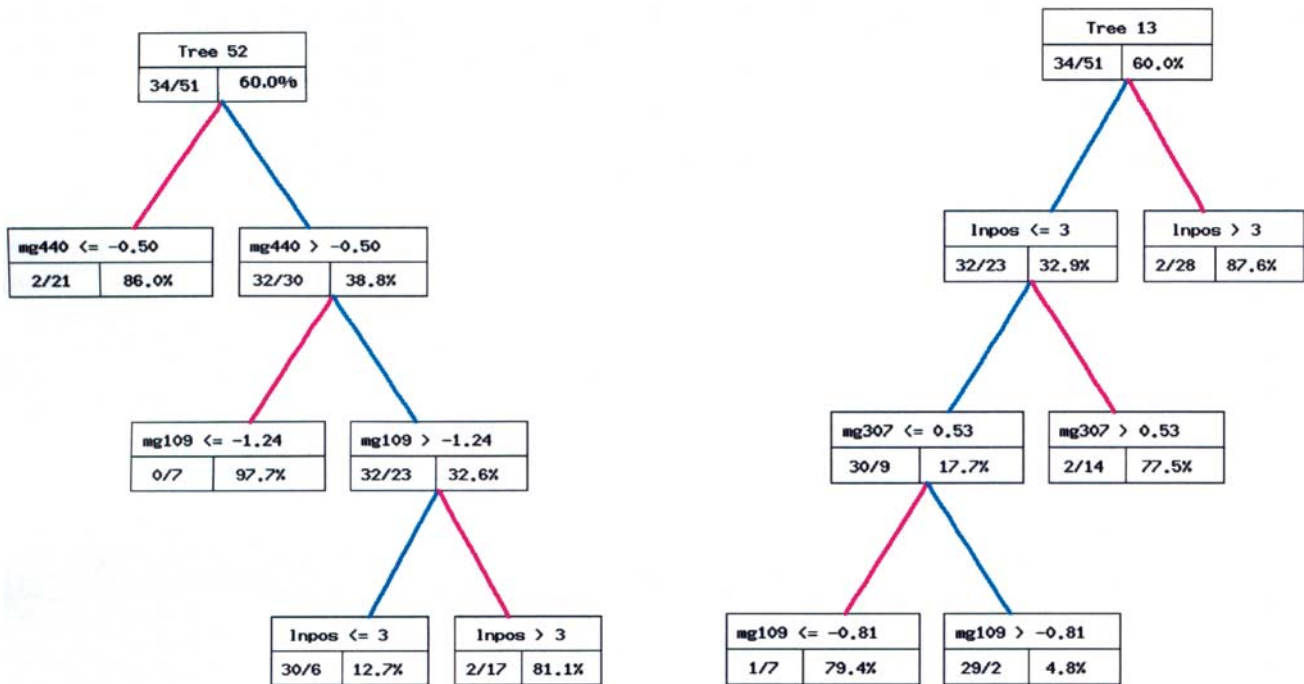
**Figure 2.** Scatter plot of node positive breast cancer samples on metagene 440 and a transform to log scale of the axillary lymph node positive count, showing 5-year recurrence free survivors (blue) and cases of recurrence within 5 years (red). As in Fig. 1, the utility of successive partitioning of samples according to thresholds on these two predictors of recurrence is highlighted, now indicating the relevance and importance of an integrative clinico-genomic approach.

for example, is currently the key clinical predictor of tumor state and aggressiveness, and recurrence risk; so, as long as this risk factor is available, ignoring its predictive value would be both scientifically and professionally inappropriate.

Figure 2 indicates the relevance of this factor and also how the partitioning analysis approach will lead trivially to a combined clinico-genomic model. As with the use of a second metagene, the use of a clinical factor, in this case lymph node involvement, resolves much of the heterogeneity in the group of samples with a high value on metagene 440. Ultimately, it may be that molecular data alone will supercede other non-genomic factors in prognosis, based on refined and improved genomic technologies that improve the capacity to characterize complex oncogenic states and the precision with which we can record the resulting data, and that also provide additional forms of molecular data on a genome-wide. Indeed, our own work has already shown that genomic information can predict and therefore begin to replace lymph node status and other clinical factors. However, at this point we believe that the needs and goals of personalized medicine must be addressed within an integrated, clinico-genomic approach that allow for and weigh the contributions of all forms of data.

### CLASSIFICATION TREE MODELS

These analyses involve the successive partitioning of patient sample—and by inference the populations they represent—into more and more homogeneous subgroups, and illustrate the ability of statistical tree-like partitioning methods to capture non-linear interactions between potential predictors of clinical outcome. The analyses rely on application of statistical classification tree models (10,11), developed using Bayesian



**Figure 3.** Two candidate tree models that combine clinical and metagene expression predictors to define personalized predictions of 5-year recurrence. The boxes at nodes indicate (i) the risk factors chosen to define the partition at the node; (ii) the numbers  $x/y$  of patients from the sample, indicating the number ( $x$ ) of cases recurrence-free at 5 years, and the number ( $y$ ) of those that recurred within 5 years; and (iii) the corresponding model-based prediction of the probability of recurrence within 5 years for future patients whose clinical and genomic characteristics would place them at the node. A future patient is assigned to a leaf of a given tree based on her combination of clinical and genomic factors, and that tree implies the prediction based on the recurrence probability at that leaf. The integrative clinico-genomic model fits multiple such tree models, weights them according to how well they fit the data, and then combines them according to the resulting weights to derive overall probabilistic predictions for future cases.

statistical methods (8,9,12). The partitions illustrated in Figure 1 correspond with the splits of the first two nodes in the upper part of the single tree model in Figure 3. At each node of a tree, the collection of metagenes and clinical factors is sampled to determine which functions to optimally divide the patients at the node—a split is made if the significance exceeds a specified level. The growth of trees is terminated when no additional metagene or clinical variable can be selected that allows a significant further split. Multiple possible splits generate collections of trees, and each is then formally evaluated based on statistical fit to the data. Each tree generates predictions for future patients: a new patient is assigned to a unique leaf of any one tree based on her genomic profile and other factors, with the corresponding prediction of recurrence based on the model-based probability at that leaf. Finally, overall predictions are based on averaging across the collection of candidate tree models.

The technical framework of predictive tree models is capable of using any form of data, and evaluates and selects from clinical and metagene factors in defining the most predictively reliable sets of models. This is obviously important in terms of both the scientific goals of customizing analysis to define personalized prognosis, since the models can evaluate and use clinical and demographic patient information as well as genomic data, and also in terms of integrating genomic information into the current professional culture of clinical decision-making. Our recent work with breast cancer bears this out, and a simple example in Figure 3 illustrates the

combination of the key clinical risk factor—extent of invasion of the axillary lymph nodes—with metagene signatures. These two tree models are from a larger set automatically generated in a combined clinico-genomic model for 5-year recurrence among these lymph node-positive patients, extending our previous work with a more refined subgroup (9).

The true predictive accuracy of any class of models can be assessed using cross-validation protocols in which the analysis is repeatedly performed by removing one sample, carrying out the training of the model on the remaining samples, and then predicting the state of the held out sample. Importantly, the entire model building process—including the selection of metagenes and clinical factors as well as the generation of trees properly weighted by the data—is performed each time to provide a true predictive evaluation. In our initial breast cancer studies, we find that the current predictive models are capable of achieving very substantial accuracy in prediction of recurrence and lymph node metastasis, with correct predictive classifications made at around the 85–90% level.

A further critical aspect of prognosis is the need to provide honest assessments of the uncertainty associated with any prediction. A predicted 70% recurrence probability, for example, should be treated quite differently by clinical decision makers if its associated uncertainty is  $\pm 30\%$  than if it were  $\pm 2\%$ . Communication of such uncertainties can be critical to the patient and the treating physician in allowing informed decisions to be made regarding the best choice for further therapeutic

treatment. Uncertainty in the predictions can arise from variability in tissue processing, hybridization measures and the limitations of analyzing relatively small numbers of samples. Perhaps most importantly, substantial uncertainty is inherent in cases for which a patient's characteristics are in conflict, so that different models may suggest different outcomes. Hence, it is important to reflect model uncertainty in prediction. The use of multiple plausible tree models (i.e. multiple statistical models) reflects the fact that there are multiple plausible combinations of interacting clinical and genomic patterns that adequately represent the observed data. Then it is critical to define overall predictions by appropriately averaging across the multiple candidate models rather than selecting one single model. To do otherwise runs the risk of incurring bias in predictions and, often more importantly, underestimating uncertainty about predicted probabilities. A single model produces a predicted probability with an associated uncertainty interval reflecting the precision of the estimated probability; combining such predictions across multiple plausible tree models, with appropriate weights that reflect the relative fit of the trees to observed data, can very substantially increase the uncertainty about the resulting overall prediction, and this is particularly true in cases when various models conflict in their predictions. Communicating this information to the patient and clinician is vital as it reflects inherent ambiguity that must be factored into the personalized decision process.

The two candidate trees in Figure 3 reflect also represent the fact that multiple metagene signatures (metagenes 440 and 307 in this example) may be surrogates for each other due to patterns of between-metagene correlation. In this analysis, a range of additional tree models involves the metagenes displayed here as well as a number of additional, correlated metagenes. This reflects the ability of the multiple metagene model approach to deal properly with the scale and complexity of tumor-based gene expression profiles that naturally leads to multiple measures of extent and aggressiveness of the tumor that show up as potential predictors of recurrence and other outcomes. Resulting, overall predictions, appropriately averaged over plausible models and with accompanying measures of honest uncertainty, then define the relevant summaries of the complex of interactions of predictive variables that are customized to the patient and so feed into the process of personalized prognosis.

## THE NEXT STEPS: TOWARDS REFINED THERAPIES

Advances and success in predicting disease outcomes based on studies involving existing therapeutic strategies are major steps in the process of bringing genomic information to clinical practice. Genomic information now has proven capacity to substantially improve prognosis in critical clinical contexts; our work with multiple metagene signatures in clinico-genomic models in breast cancer exemplifies this, and the near future will surely see application of multiple forms of genomic data in purely prognostic clinical settings to better manage patients based on personalized predictions.

Beyond this, one of the major steps in the application of genomic information in achieving the goals of personalized medicine will be concurrent development of new therapeutic modalities that can be matched to the individual patient's characteristics. This will rely on substantial advances in our understanding of multiple genes and interacting pathways that may eventually represent new therapeutic targets. Hence, in parallel with the move to the clinic of prognostic tests, we see as key the initiation of larger-scale efforts to develop investigations of the gene pathways and interactions that are indicated by the discovery of metagene patterns that are truly predictive of the critical clinical endpoints. Defining focused efforts to understand the biology underlying this predictive value will pave the way for therapeutic advances towards curing cancer, rather than just better personalized disease management.

## REFERENCES

1. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.
2. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S. *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.
3. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.
4. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.
5. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
6. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
7. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *New Engl. J. Med.*, **347**, 1999–2009.
8. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A. Jr, Marks, J.R. and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.
9. Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Hornig, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R. and Huang, A.T. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.
10. Breiman, L., Friedman, J., Olshen, L. and Stone, C. (1984) *Classification and Regression Trees* (Chapman and Hall/CRC, Boca Raton, FL).
11. Chipman, H., George, E. and McCulloch, R.E. (1998) Bayesian CART model search. *J. Am. Stat. Assoc.*, **93**, 935–960.
12. Pittman, J., Liao, M., Huang, E., Nevins, J.R. and West, M. (2002) Binary prediction tree modeling with many predictors. ISDS Discussion paper, submitted for publication.