MOLECULAR ONCOLOGY 5 (2011) 190–196

available at www.sciencedirect.com

www.elsevier.com/locate/molonc

Review

Multidimensionality of microarrays: Statistical challenges and (im)possible solutions

Stefan Michiels[a,*], Andrew Kramar[b], Serge Koscielny[c]

[a]Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles, ULB290, Boulevard de Waterloo 121, 1000 Bruxelles, Belgium
[b]Unité de Méthodologie et Biostatistique, Centre Oscar Lambret, Lille, France
[c]Unit of Biostatistics and Epidemiology, Institut Gustave Roussy, France

ARTICLE INFO

Article history:
Received 22 September 2010
Accepted 27 January 2011
Available online 3 February 2011

Keywords:
Biomarker
Microarray
Development
Validation
-omics
Prognostic

ABSTRACT

A typical array experiment yields at least tens of thousands of measurements on often not more than a hundred patients, a situation often denoted as the curse of dimensionality. With a focus on prognostic multi-biomarker scores derived from microarrays, we highlight the multidimensionality of the problem and the issues in the multidimensionality of the data. We go over several statistical challenges raised by this curse occurring in each step of microarray analysis on patient data, from the hypothesis and the experimental design to the analysis methods, interpretation of results and clinical utility. Different analytical tools and solutions to answer these challenges are provided and discussed.

1. Introduction

A typical microarray experiment yields currently some tens of thousands of measurements – or even millions of genotypes – on often not more than a hundred patients, a situation often denoted as the curse of dimensionality. At the end of the road the experimenters try to summarize the huge magnitude of information in a parsimonious equation or multi-biomarker score. It is desirable that this biomarker score will be reproducible, sensitive and specific for its association with a clinical endpoint, with an important impact on treatment decisions from a clinical and economic viewpoint.

In this paper, we focus on biomarkers that affect the outcome or prognosis of individual patients in terms of a clinical endpoint ("prognostic" biomarkers) and to a lesser extent on biomarkers that are related to the effect of a specific treatment on a clinical endpoint ("predictive" biomarkers, more broadly called treatment "effect modifiers" outside the oncology field). Array technologies are expected to obtain reliable information for developing such biomarkers and to open the door to patient-specific personalized medicine. Microarray experts have even hailed the possibility of conducting clinical trials with only a few patients (Liu and Karuturi, 2004). As of September 2010, 31,633 peer-reviewed articles containing the words "gene" and "microarray" can be found in PubMed.

* *Corresponding author*. Tel.: +32 2 541 37 43; fax: +32 2 538 08 58.
E-mail address: stefan.michiels@bordet.be (S. Michiels).
1574-7891/$ – see front matter © 2011 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.molonc.2011.01.002

A closer look at the literature reveals many conflicting results. When different analysis teams start from the same raw microarray data, completely opposite results in terms of prognostic value can be obtained (Coombes et al., 2007). On the other hand, when the analysis strategy is fixed, subtle changes in the patient data used for determining the biomarker score can lead to remarkably different gene lists and prediction results (Michiels et al., 2005).

Replication in independent patient series is also often lacking (Ioannidis et al., 2009).

The most famous applications of multi-biomarker scores are found in breast cancer where different gene classifiers have been developed to address the same clinical question of whom to treat with adjuvant chemotherapy. However because of imperfect concordance between the tests, this might lead in the future to a situation in which different predictive tests may lead to different treatment decisions for one and the same patient (Koscielny, 2008, 2010). All taken together, this calls for some careful attention to the different steps of microarray analysis.

Many papers have nicely reviewed in detail how to analyze typical microarray data experiments (Allison et al., 2006; Reimers, 2010; Simon et al., 2003), to interpret them (Michiels et al., 2007) and to report the results (Dupuy and Simon, 2007). We will not go into the important transition from the microarray images to data ready to be used for statistical analysis (Owzar et al., 2008).

In this paper, we voluntarily start with playing the devil's advocate: using a provocative point of view to illustrate the multidimensionality of the problem. We will then go over several statistical challenges raised by this multidimensionality. What are the specific analytical tools and solutions that have been proposed and what their problematic is. Can we come up with better ones? We will go over these challenges occurring in each step of microarray analysis on patient data, from the hypothesis and the experimental design to the analysis methods, interpretation of the results and clinical utility.

2. Playing the devil's advocate: the multidimensionality of the problem

For the last few years, the medical literature has been invaded by microarray studies aimed at defining gene profiles to discriminate between good and poor prognosis tumors. The biological postulate underlying prognostic microarray studies is that all tumors acquire a metastasis phenotype through the same unique mechanism, and that gene expression data in tumor tissue obtained at resection of the primary tumor can be used to clearly distinguish tumors that will relapse from those that will not. The results of the pioneering prognostic microarray study concerning breast cancer (van 't Veer et al., 2002) are considered proof of concept and have led to general acceptance of the postulate. However, the performances of microarray studies are poorer than initially thought and published gene signature lists are unstable (Michiels et al., 2005). Some of the multi-biomarker scores do show consistent prognostic value such as is the case in breast cancer, but until the recent advent of large validation studies, microarray studies have not allowed a significantly better prognostic classification than conventional prognostic models (Albain et al., 2010; Buyse et al., 2006). In addition, it has been shown that almost all first-generation gene signatures in breast cancer provide a quantitative read-out of the same biological pathway of proliferation (Haibe-Kains et al., 2008; Wirapati et al., 2008). As of today we are still in need of a precise estimation of the incremental value (EGAPP Working Group, 2009; Koscielny and Michiels, 2010; Marchionni et al., 2008). Moreover, by assuming a unique mechanism for the metastasis phenotype, the microarray postulate is in contradiction with the concept of cancer heterogeneity, and consequently with the need for individualized treatments. This would mean that, without rejecting the potential interest of microarrays, true critical consideration, incorporating, and not opposed to, full clinical evidence is now necessary. For example, early detection and screening are known to be efficient strategies to reduce cancer mortality (CISNET Breast Cancer Collaborators, 2006; Koscielny et al., 2009), which would imply that the prognosis is not a built-in characteristic of a tumor, present in the genes during the evolutionary process of the tumor. Consequently either genomic characteristics of tumors change with time or the meaning of these characteristics depends on tumors' clinical characteristics. In that sense the prognosis is a multidimensional problem, with interlaced complex clinical and biological dimensions.

3. The multidimensionality of the data

A common objective of many -omics studies is to find the genes that are most differentially expressed between two (or more) classes of tumors with different characteristics: for instance, between a group of tumor samples from primary cutaneous melanoma patients who developed a distant metastasis within 4 years after surgery and a group of tumors from patients who did not (Winnepenninckx et al., 2006). A statistic measuring the difference in gene expression between the two types of tumors is selected such as a standard t-statistic, or a logrank statistic for censored survival data, or variants developed especially for microarray data. Genes are then ranked according to this statistic, starting with the most differentially expressed gene. A cutoff is selected leading to a list of genes most differentially expressed. In the cutaneous melanoma example, the top 254 genes were selected to be differentially expressed between the groups of patients with distant metastases versus those without (all genes with an individual p-value lower than 0.001).

When one applies a statistical test for each gene, the number of tests performed is equal to the number of genes. If 10 000 genes are studied and none are really associated with the characteristics under study, then, taking the usual 5% limit for a significant p-value, one expects 5% of the genes, that is, 500 genes to appear as significantly associated with the characteristics, all being false positives. One can lower the p-value cutoff such as in the melanoma example above (for which only 10 out of 10 000 genes would be expected to be false positives at the 0.001 cutoff), but one must not forget that lowering the p-value cutoff will reduce the number of

false positives but can also increase the number of false negatives.

Another solution to reduce the risk of false positives is to select more stringent rules to define statistical significance. The family-wise error rate criterion aims to control the probability of making at least one false positive among all biomarkers tested. A less stringent criterion aims at the false discovery rate or the expected proportion of false positive genes among those declared as differentially expressed. For instance, Benjamini and Hochberg (Benjamini and Hochberg, 1995) suggest to rank the genes according to the *p*-values, starting with the most significant, and to compare the ith *p*-value pi to 5% × *i*/*n*, where *i* is the rank in the list and *n* is the total number of genes. Under some "soft" hypotheses on possible dependence between genes, this limits to 5% on average the proportion of false positives among the genes declared significant. Because many genes might actually have strong correlation patterns among each other, more computer-intensive permutation methods are available which control these criteria (Ge et al., 2003).

4. Multidimensional analytical tools and multidimensional solutions

4.1. *Experimental design*

Once the primary objective of the array study is chosen, an experimental design has to be defined to control and exclude as many biases as possible and reliably test the study hypothesis (Ransohoff, 2007). Possibilities of bias start with specimen collection biases (Ransohoff and Gourlay, 2010). Guidelines have been proposed for the different phases of development (Buyse et al., 2010; Pepe et al., 2001). Here, we adapt the phases as proposed in the cardiovascular field with some minor modifications (Hlatky et al., 2009), displayed in Table 1.

Possible improvements in study design include randomization of the samples to the laboratory processes, blinding of the laboratory staff to clinical outcome, use of nested case-control designs (Pepe et al., 2008) and also matched retrospective analyses. An expensive way to tackle the multidimensionality problem is to increase the sample size of the array experiments, which will lower the false discovery and false negative rate (Pawitan et al., 2005).

4.2. *Cluster analysis or unsupervised classification*

One way to reduce the multidimensionality of the data is to transform the entire data set into a limited set of clusters without a priori knowledge. For example, breast cancers were among the first cancer sites to have been divided into several subgroups using cluster analysis of microarray data (Perou et al., 2000).

A commonly used hierarchical clustering method starts by defining a distance between two tumors as a function of the difference in gene expression. One then regroups the two closest tumors and proceeds by regrouping tumors to obtain a cluster tree, which can be split into branches by selecting a cutoff distance. There are many algorithms available for performing clusterization, and for a given algorithm there are many ways to define a cutoff distance. Furthermore, even in the case of random noise, the technique produces a cluster tree (Miller et al., 2002). It is thus very difficult to know whether the results observed are a characteristic of the sample considered or whether they would be reproducible in another similar collection of tumors. One way to evaluate whether the obtained clusters are stable or not is to perturb the original data set by resampling and to investigate whether the same clusters are found again (Suzuki and Shimodaira, 2006).

Clustering results have shown to be very dependent on the type of normalization method used for the array data set (Lusa et al., 2007) and it is very hard to project the clusters to independent data sets. Thus, unsupervised analyses pose several problems: classification instability because the inclusion of a new patient may modify it, arbitrariness in the choice of the algorithm used for clustering and in the choice of the number of classes. It seems that cluster analysis methods have been somehow overused in the first burst of enthusiasm for microarrays and methodologists consider that outcome-related problems should be addressed with supervised strategies as those presented in the next section (Allison et al., 2006; Dupuy and Simon, 2007; Michiels et al., 2007).

4.3. *Development of a prediction rule or supervised classification*

A popular way to tackle the multidimensionality problem in an outcome-related array study is to develop a clinically relevant multi-biomarker score or prediction rule from the data. We encourage the pre-specification of the different components – choice of gene selection method, prediction rule and

Table 1 – Different phases of development and evaluation of a clinically useful prognostic multi-biomarker score or "biomarker" using array technologies.

N°	Phase	Elaboration
1	Proof of concept	Do biomarker levels differ between subjects with and without outcome?
2	External validation	Does the biomarker predict development of future outcomes in a cohort or nested case-cohort/case-cohort study?
3	Incremental value	Does the biomarker add information to established, standard risk markers?
4	Clinical utility	Does the biomarker change predicted risk sufficiently to change recommended therapy?
5	Clinical outcome	Does use of the biomarker improve clinical outcome, especially when tested in a randomized clinical trial?
6	Cost-effectiveness	Does use of the biomarker improve clinical outcome sufficiently to justify the additional costs of testing and treatment?

if necessary cutoffs – in a translational research protocol. A metric needs also to be chosen in order to evaluate performance of the prediction rule, such as the misclassification rate (error rate), sensitivity and specificity, or measures of predictive accuracy such as area under the Receiving Operating Characteristics curve (sensitivity versus 1 – specificity) and explained variation (Gerds et al., 2008).

Let's have a closer look at one of the breast cancer multi-biomarker scores that is one of the closest to implementation in clinical practice: van't Veer's 70-gene signature for predicting the occurrence of distant metastasis in breast cancer patients, patented under the name MammaPrint. This multi-biomarker score used a nearest centroid prediction rule, relatively simple but still top of the cream of the published gene signatures (Koscielny, 2010). For each new patient, the expression of the 70 genes is measured and a distance is calculated to the average values of those 70 genes among the subset of patients who experienced a distant relapse in the retrospective series used to train the prediction rule (van 't Veer et al., 2002). If the distance is low enough (a correlation value above 0.4) the patient will be classified as poor prognosis.

Many more complicated prediction rules have been suggested in the microarray literature. The results have been adequately described as a statistical tower of Babel (Allison et al., 2006). Some microarray analysis packages present systematically the results of several classification methods for a single data set. It is then very tempting to publish only the best-looking result, leading to a biased evaluation of the performance of the prediction rule (Ioannidis, 2005). We wrote in 2005 that, in principle, there is no biological or mathematical reason why one particular classification method should be better than another for the prediction of the outcome of cancer patients based on microarray data since there are many possible solutions in the multidimensional gene expression space (Michiels et al., 2005). But what is the reason that relatively "simple" methods work well in the data cursed by the multidimensionality, beyond an Occam's razor interpretation? There is not much empirical evidence available that more complex models would outperform simpler ones and no one method is widely accepted as superior or optimal. All methods, simple or complicated ones, are actually susceptible to overfitting to a certain degree. The large MicroArray Quality Consortium II, led by the FDA has recently shown that variations on univariate gene selection methods and prediction rules have only a modest impact on performance (Shi et al., 2010) and several statistically equally good predictors can be developed for any given classification problem (Popovici et al., 2010).

Still the same question remains: how should the huge amount of biomarker information be summarized into one single multi-biomarker equation? The main types of approaches have been coined: the 'top-down' approach and the hypothesis-driven or 'bottom-up' approach. The top-down approach derives from a prognostic model simply by looking for gene expression patterns associated with clinical outcome without any a priori biological assumption, whereas the bottom-up approach first identifies gene expression profiles linked with a specific biological phenotype and subsequently correlates these findings to an appropriately defined clinical outcome (Liu, 2005; Sotiriou and Piccart, 2007).

5. Challenges in the transfer to the clinic

5.1. *External validation*

Showing that a multi-biomarker score beats chance in a single data set was actually the easy part. The next crucial step in the translation of gene signatures in a clinical setting is external validation, which can be defined as providing evidence that a prediction rule works satisfactorily on patients other than those used to define the biomarker score (Altman and Royston, 2000). This is needed because many different factors can explain the separation between the distinct groups of patients: chance finding, biases, etc… (Ransohoff, 2004, 2005). External validation requires an independent study to be prospectively designed to confirm the results of a previous study, in order to reduce the play of chance and the potential for biases. The same methodological guidelines apply as for the validation of classical tumor markers (REMARK NCI-EORTC Guidelines (McShane et al., 2005)) but there have been many easy-to-avoid mistakes in multi-biomarker score validation studies (Michiels et al., 2007). Ideally, validation of an experimental gene signature should be performed in an independent patient population in a similar clinical setting, by an independent research team.

Some difficulties in replication using publicly available data sets can arise due to cross-platform differences but evidence-based guidelines have been started to be developed for performing solid meta-analyses of array data (Ramasamy et al., 2008).

5.2. *Incremental value*

When a multi-biomarker score has been shown to classify better than chance and this prognostic value has been externally validated, what is its true place in clinical practice compared to clinical prediction methods that are already widely established? This is one of the most important hurdles facing the translation of gene expression signatures into the clinic.

One must therefore study whether these signatures add prognostic information to the clinical decision rules in use. It is not sufficient to perform a multivariate regression analysis, for instance a Cox's regression model, comparing the effects of the clinical prognostic factors and of the signature, and to show that the gene signature is 'more significant' than the clinical factors in this model. It has even been shown that a marker with an odds ratio of 3 is in fact a very poor classification tool and that an odds ratio of 30 or more is desirable (Pepe et al., 2004). A biomarker is of interest only if it provides additional prognostic value, over and above that of all easily measured clinical and pathological characteristics of the patients. The gain in predictive accuracy by the classifier as compared to established clinical prognostic factors should therefore be quantified (Dunkler et al., 2007; Kattan, 2003), by comparing the predictive accuracy of the two multivariate models with and without the gene signature. The updated prognostic model will need to be well calibrated as well by comparing the agreement between the predicted with the observed outcomes (Steyerberg et al., 2010). A multi-biomarker

score could also be of interest if it provides a more reproducible, cheap and precise assay of an already existing tumor measurement that has proven clinical utility so that the clinical prediction rule could be updated.

5.3. *Implementation in clinical practice: impact studies*

The two signatures in breast cancer that have already been used to design clinical trials are the Mammaprint score mentioned above and a 21-gene score called the OncotypeDX assay. In these two randomized trials, chemotherapy is compared to no chemotherapy in the population of patients classified according to the results of a genomic test.

In theory, a trial is actually not needed to validate the prognostic value of a multi-biomarker score since this can be done through multiple independent retrospective validation studies. But randomized trials can be used to validate a gene signature prospectively in the absence of high quality retrospective material (and thus avoid all biases that may affect retrospective validation), investigate the potential of the gene signature as an effect modifier (e.g. associated with the magnitude of the benefit of treatment) and confirm the gene signature's clinical utility or impact on treatment decisions. An overview of some of the recent biomarker-based clinical trial designs can be found in (Buyse and Michiels, 2010).

6. Conclusion

The holy grail in microarray studies seems to have been the search for prognostic multi-biomarker scores based on retrospective convenience samples. But due to the multidimensionality of the data, many different biomarker scores can be constructed that all have a similar amount of prognostic information. No single unique prognostic biomarker score will probably exist for a given disease type. Although gene lists seem at first sight very distinct, it could well be that they are all a read-out from the same biological characteristic of a tumor (such as proliferation in the early breast cancer example).

In our view, a priori specification of the rules and methods used to divide the data and determine the multi-biomarker score -i.e. the selection of the genes, the prediction rule, and the cutoffs- is an optimal approach for hypothesis-driven research and, in this case, leads to the generalizability of prediction accuracy estimates for future patient series. In contrast, trying several alternative methods and choosing the most optimal one among them is still a good way to generate false positive results or to support unduly optimistic views.

A great deal of the concerns raised from multidimensionality of arrays boils down to some of the basics in statistics: solid experimental design. Array technologies do provide fascinating discoveries but we encourage the use of more informative designs for translational research studies, such as designs with repeated time points measuring change in gene expression before, during and after treatment in order to identify targets of new and not so new treatment regimens.

Role of the funding source

None.

Conflict of interest

None.

REFERENCES

Albain, K.S., Barlow, W.E., Shak, S., Hortobagyi, G.N., Livingston, R.B., Yeh, I.T., Ravdin, P., Bugarini, R., Baehner, F.L., Davidson, N.E., Sledge, G.W., Winer, E.P., Hudis, C., Ingle, J.N., Perez, E.A., Pritchard, K.I., Shepherd, L., Gralow, J.R., Yoshizawa, C., Allred, D.C., Osborne, C.K., Hayes, D.F., 2010. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. The Lancet Oncology 11, 55–65.

Allison, D.B., Cui, X., Page, G.P., Sabripour, M., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics 7, 55–65.

Altman, D.G., Royston, P., 2000. What do we mean by validating a prognostic model? Statistics in Medicine 19, 453–473.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289–300.

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A.M., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., Piccart, M.J., 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. Journal of the National Cancer Institute 98, 1183–1192.

Buyse, M., Michiels, S., 2010. Biomarkers and surrogate endpoints in clinical trials. In: Kelly, Kevin, Halabi, Susan (Eds.), Oncology Clinical Trials: Successful Design, Conduct and Analysis. Demos Medical Publishing.

Buyse, M., Sargent, D.J., Grothey, A., Matheson, A., de Gramont, A., 2010. Biomarkers and surrogate end points–the challenge of statistical validation. Nature Reviews 7, 309–317.

CISNET Breast Cancer Collaborators, 2006. The impact of mammography and adjuvant therapy on U.S. breast cancer mortality (1975–2000): collective results from the Cancer Intervention and Surveillance Modeling Network. Journal of National Cancer Institute Monographs 36, 1–126.

Coombes, K.R., Wang, J., Baggerly, K.A., 2007. Microarrays: retracing steps. Nature Medicine 13, 1276–1277, author reply pp. 1277–1278.

Dunkler, D., Michiels, S., Schemper, M., 2007. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? European Journal of Cancer 43, 745–751.

Dupuy, A., Simon, R.M., 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute 99, 147–157.

EGAPP Working Group, 2009. Recommendations from the EGAPP Working Group: can tumor gene expression profiling improve outcomes in patients with breast cancer? Genetics in Medicine 11, 66–73.

Ge, Y.C., Dudoit, S., Speed, T.P., 2003. Resampling-based multiple testing for microarray data analysis. Test 12, 1–77.

Gerds, T.A., Cai, T., Schumacher, M., 2008. The performance of risk prediction models. Biometrical Journal 50, 457–479.
Haibe-Kains, B., Desmedt, C., Sotiriou, C., Bontempi, G., 2008. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics (Oxford, England) 24, 2200–2208.
Hlatky, M.A., Greenland, P., Arnett, D.K., Ballantyne, C.M., Criqui, M.H., Elkind, M.S., Go, A.S., Harrell Jr., F.E., Hong, Y., Howard, B.V., Howard, V.J., Hsue, P.Y., Kramer, C.M., McConnell, J.P., Normand, S.L., O'Donnell, C.J., Smith Jr., S.C., Wilson, P.W., 2009. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 119, 2408–2416.
Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Medicine 2, e124.
Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G.P., Petretto, E., van Noort, V., 2009. Repeatability of published microarray gene expression analyses. Nature Genetics 41, 149–155.
Kattan, M.W., 2003. Judging new markers by their ability to improve predictive accuracy. Journal of the National Cancer Institute 95, 634–635.
Koscielny, S., 2008. Critical review of microarray-based prognostic tests and trials in breast cancer. Current Opinion in Obstetrics & Gynecology 20, 47–50.
Koscielny, S., 2010. Why most gene expression signatures of tumors have not been useful in the clinic. Science Translational Medicine 2, 14ps2.
Koscielny, S., Arriagada, R., Adolfsson, J., Fornander, T., Bergh, J., 2009. Impact of tumour size on axillary involvement and distant dissemination in breast cancer. British Journal of Cancer 101, 902–907.
Koscielny, S., Michiels, S., 2010. Clinical usefulness of microarrays for cancer prognosis in 2010–letter. Clinical Cancer Research 16, 6180.
Liu, E.T., 2005. New technologies for high-throughput analysis. Pharmacogenomics 6, 469–471.
Liu, E.T., Karuturi, K.R., 2004. Microarrays and clinical investigations. The New England Journal of Medicine 350, 1595–1597.
Lusa, L., McShane, L.M., Reid, J.F., De Cecco, L., Ambrogi, F., Biganzoli, E., Gariboldi, M., Pierotti, M.A., 2007. Challenges in projecting clustering results across gene expression-profiling datasets. Journal of the National Cancer Institute 99, 1715–1723.
Marchionni, L., Wilson, R.F., Wolff, A.C., Marinopoulos, S., Parmigiani, G., Bass, E.B., Goodman, S.N., 2008. Systematic review: gene expression profiling assays in early-stage breast cancer. Annals of Internal Medicine 148, 358–369.
McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., Clark, G.M., 2005. Reporting recommendations for tumor marker prognostic studies (REMARK). Journal of the National Cancer Institute 97, 1180–1184.
Michiels, S., Koscielny, S., Hill, C., 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365, 488–492.
Michiels, S., Koscielny, S., Hill, C., 2007. Interpretation of microarray data in cancer. British Journal of Cancer 96, 1155–1158.
Miller, L.D., Long, P.M., Wong, L., Mukherjee, S., McShane, L.M., Liu, E.T., 2002. Optimal gene expression analysis by microarrays. Cancer Cell 2, 353–361.
Owzar, K., Barry, W.T., Jung, S.H., Sohn, I., George, S.L., 2008. Statistical challenges in preprocessing in microarray experiments in cancer. Clinical Cancer Research 14, 5959–5966.
Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., Ploner, A., 2005. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics (Oxford, England) 21, 3017–3024.
Pepe, M.S., Etzioni, R., Feng, Z., Potter, J.D., Thompson, M.L., Thornquist, M., Winget, M., Yasui, Y., 2001. Phases of biomarker development for early detection of cancer. Journal of the National Cancer Institute 93, 1054–1061.
Pepe, M.S., Feng, Z., Janes, H., Bossuyt, P.M., Potter, J.D., 2008. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. Journal of the National Cancer Institute 100, 1432–1438.
Pepe, M.S., Janes, H., Longton, G., Leisenring, W., Newcomb, P., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology 159, 882–890.
Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O., Botstein, D., 2000. Molecular portraits of human breast tumours. Nature 406, 747–752.
Popovici, V., Chen, W., Gallas, B.G., Hatzis, C., Shi, W., Samuelson, F.W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., Hess, K.R., Valero, V., Booser, D., Delorenzi, M., Hortobagyi, G.N., Shi, L., Symmans, W.F., Pusztai, L., 2010. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Research 12, R5.
Ramasamy, A., Mondry, A., Holmes, C.C., Altman, D.G., 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Medicine 5, e184.
Ransohoff, D.F., 2004. Rules of evidence for cancer molecular-marker discovery and validation. Nature Reviews Cancer 4, 309–314.
Ransohoff, D.F., 2005. Bias as a threat to the validity of cancer molecular-marker research. Nature Reviews Cancer 5, 142–149.
Ransohoff, D.F., 2007. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. Journal of Clinical Epidemiology 60, 1205–1219.
Ransohoff, D.F., Gourlay, M.L., 2010. Sources of bias in specimens for research about molecular markers for cancer. Journal of Clinical Oncology 28, 698–704.
Reimers, M., 2010. Making informed choices about microarray data analysis. PLoS Computational Biology 6 e1000786.
Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.M., Goodsaid, F.M., Pusztai, L., Shaughnessy Jr., J.D., Oberthuer, A., Thomas, R.S., Paules, R.S., Fielden, M., Barlogie, B., Chen, W., Du, P., Fischer, M., Furlanello, C., Gallas, B.D., Ge, X., Megherbi, D.B., Symmans, W.F., Wang, M.D., Zhang, J., Bitter, H., Brors, B., Bushel, P.R., Bylesjo, M., Chen, M., Cheng, J., Cheng, J., Chou, J., Davison, T.S., Delorenzi, M., Deng, Y., Devanarayan, V., Dix, D.J., Dopazo, J., Dorff, K.C., Elloumi, F., Fan, J., Fan, S., Fan, X., Fang, H., Gonzaludo, N., Hess, K.R., Hong, H., Huan, J., Irizarry, R.A., Judson, R., Juraeva, D., Lababidi, S., Lambert, C.G., Li, L., Li, Y., Li, Z., Lin, S.M., Liu, G., Lobenhofer, E.K., Luo, J., Luo, W., McCall, M.N., Nikolsky, Y., Pennello, G.A., Perkins, R.G., Philip, R., Popovici, V., Price, N.D., Qian, F., Scherer, A., Shi, T., Shi, W., Sung, J., Thierry-Mieg, D., Thierry-Mieg, J., Thodima, V., Trygg, J., Vishnuvajjala, L., Wang, S.J., Wu, J., Wu, Y., Xie, Q., Yousef, W.A., Zhang, L., Zhang, X., Zhong, S., Zhou, Y., Zhu, S., Arasappan, D., Bao, W., Lucas, A.B., Berthold, F., Brennan, R.J., Buness, A., Catalano, J.G., Chang, C., Chen, R., Cheng, Y., Cui, J., Czika, W., Demichelis, F., Deng, X., Dosymbekov, D., Eils, R., Feng, Y., Fostel, J., Fulmer-Smentek, S., Fuscoe, J.C., Gatto, L., Ge, W., Goldstein, D.R., Guo, L., Halbert, D.N., Han, J., Harris, S.C., Hatzis, C., Herman, D., Huang, J., Jensen, R.V., Jiang, R., Johnson, C.D., Jurman, G., Kahlert, Y., Khuder, S.A., Kohl, M.,

Li, J., Li, L., Li, M., Li, Q.Z., Li, S., Li, Z., Liu, J., Liu, Y., Liu, Z., Meng, L., Madera, M., Martinez-Murillo, F., Medina, I., Meehan, J., Miclaus, K., Moffitt, R.A., Montaner, D., Mukherjee, P., Mulligan, G.J., Neville, P., Nikolskaya, T., Ning, B., Page, G.P., Parker, J., Parry, R.M., Peng, X., Peterson, R.L., Phan, J.H., Quanz, B., Ren, Y., Riccadonna, S., Roter, A.H., Samuelson, F.W., Schumacher, M.M., Shambaugh, J.D., Shi, Q., Shippy, R., Si, S., Smalter, A., Sotiriou, C., Soukup, M., Staedtler, F., Steiner, G., Stokes, T.H., Sun, Q., Tan, P.Y., Tang, R., Tezak, Z., Thorn, B., Tsyganova, M., Turpaz, Y., Vega, S.C., Visintainer, R., von Frese, J., Wang, C., Wang, E., Wang, J., Wang, W., Westermann, F., Willey, J.C., Woods, M., Wu, S., Xiao, N., Xu, J., Xu, L., Yang, L., Zeng, X., Zhang, J., Zhang, L., Zhang, M., Zhao, C., Puri, R.K., Scherf, U., Tong, W., Wolfinger, R.D., 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology 28, 827–838.

Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M., 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95, 14–18.

Sotiriou, C., Piccart, M.J., 2007. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7, 545–553.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass.) 21, 128–138.

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics (Oxford, England) 22, 1540–1542.

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., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S.R., Avril, M.F., Ortiz Romero, P.L., Robert, T., Balacescu, O., Eggermont, A.M., Lenoir, G., Sarasin, A., Tursz, T., van den Oord, J.J., Spatz, A., 2006. Gene expression profiling of primary cutaneous melanoma and clinical outcome. Journal of the National Cancer Institute 98, 472–482.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., Goldstein, D.R., Piccart, M., Delorenzi, M., 2008. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research 10, R65.