Microarray data analysis: strategies, pitfalls and applications in clinical oncology

S.J. Van Laere, P.A. van Dam, L.Y. Dirix, P.B. Vermeulen

Microarray technology has rapidly established a critical site in cancer research. Although the procedures to perform microarray experiments require specialised laboratory equipment and trained personnel, the task of data analysis is often a more tantalising job. Great care must be taken with the interpretation of the results, as there is often no a priori way to detect flaws in the analytical procedures. The power of the microarray technology lies in its great versatility. The technology can be used to identify new, previously undefined subgroups of supposedly homogeneous tumour samples, to identify new (single-gene and multi-gene) biomarkers for disease progression (prognosis) or treatment response (prediction) and to functionally and molecularly unravel poorly characterised conditions. The purpose of this review is to provide the reader with a glimpse of the possible strategies of analysis that exist for microarray data, their advantages and disadvantages, their applicability to cancer research and how they might, reshape the field of clinical oncology in the future.

(Belg J Med Oncol 2011;5:50-6)

Introduction

Since the landmark paper published by Schena et al, describing a novel technique that allows the investigation of the expression of multiple genes in one experiment, a true avalanche of studies reporting on the microarray technology arose.[1] It took only 1 year before the first paper applying microarrays to cancer (i.e. melanoma) was published.[2] To date, approximately 13,000 studies have used a derivate of the microarray technology to study the biology of cancer from different perspectives (biology, diagnosis, prognosis, treatment prediction). *Figure 1* demonstrates the explosive increase in studies associating microarrays and cancer since the inception of the technology back in 1995.

In 2003, microarrays were introduced in clinical trials. DePrimo et al applied the technology to for the identification of predictive biomarkers for response in patients with metastatic colorectal cancer treated with SU5416 in a Phase III trial (*Figure 1*).[3] More recently, 2 clinical randomised trials have been initiated to investigate the possibility of using gene expression assays for the selection of patients with breast cancer able to benefit from chemotherapy. The MINDACT trial aims at prospectively validating the use of the 70-gene poor prognosis signature as a tool for the improvement of the selection of patients with good prognosis who would not benefit from adjuvant chemotherapy.[4-6] The TAILORx-trial aims at the validation of the OncotypeDX recurrence

Authors: S.J. Van Laere MsC PhD, Oncology Centre, St. Augustinus Hospital; P.A. van Dam MD PhD; L.Y. Dirix MD PhD; P.B. Vermeulen MD PhD, Translational Cancer Research Group (Laboratory of Pathology, University of Antwerp, Belgium; Oncology Centre, St. Augustinus Hospital, Wilrijk, Belgium).

Please send all correspondence to: S.J. Van Laere, MsC PhD, Oncology Centre, St. Augustinus Hospital, Oosterveldlaan 24, 2610 Wilrijk, Belgium; E-mail: Steven.VanLaere@GZA.be.

Conflict of interest: the authors have nothing to disclose and indicate no potential conflicts of interest.

Key words: microarray, class discovery, class prediction, class comparison.

score as a tool for discriminating between patients who would benefit and patients who would not benefit from the addition of chemotherapy to hormonal treatment.[7]
The principle of the microarray technology is based on the unique feature of RNA- and DNA-molecules to bind to RNA- or DNA molecules with a complementary sequence (i.e A with T/U and G with C). This process is called hybridisation. When performing a microarray experiment, RNA is extracted from a sample of interest (e.g. a breast biopsy). This RNA sample, often referred to as the target, is subsequently converted to copy DNA (cDNA), which is further processed and ultimately coupled to a fluorescent dye. Once the cDNA is fluorescently labeled, it is hybridised to the microarray. The latter is a matrix of spots and each spot contains cDNA molecules, often called probes, each with a unique sequence designed to specifically bind to the cDNA molecules from only 1 corresponding gene. During hybridisation, the cDNA molecules in the spots will capture their fluorescently labeled complements from the target, and, as such, the spot will acquire a fluorescent signal. The intensity of this fluorescent signal is proportional to the amount of complementary cDNA molecules in the target, and hence is a measure of the expression of the corresponding gene.
The microarray is scanned with an automated confocal laser scanner allowing the researcher to determine the fluorescent intensity of all spots on the microarray. After scanning, the resulting images are processed during the image analysis procedure to extract raw signal intensity values. These raw signal intensity values need to be normalised to reduce technical variation and increase array comparability. The guiding principle behind data normalisation is that the majority of the genes are not differentially expressed and, hence, that global differences in fluorescence intensities between different arrays are not expected. After normalisation of the fluorescence intensities, quality assessment is performed by filtering out spots with low signal-to-noise ratios, to prevent biased results.
The procedures to perform a microarray experiment require trained personnel and specialised laboratory equipment. Nevertheless, although the process of data analysis is logistically more feasible (it only requires powerful computers with specific software), it is without doubt one of the most tantalising tasks due to the need to process vast amounts of data. To unravel the biological data provided by a microarray experiment requires both skills and cautiousness, as frequently there is no a priori way to detect flaws in the analysis procedures. The focus of the current review is to give the reader a glimpse of the pitfalls associated with microarray data analysis.

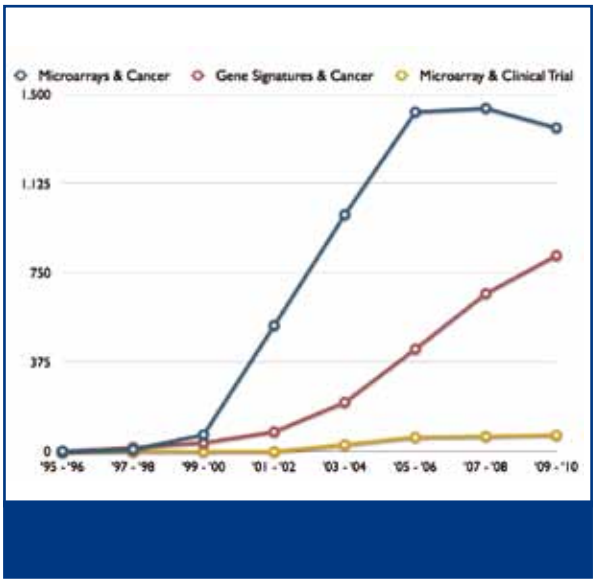

Figure 1. A PubMed-search for "Microarrays and Cancer" (blue), "Gene Signatures and Cancer" (red) and "Microarray and Clinical Trial" (yellow) was performed for each two-year period from 1995 until 2010 (X-axis). Since the inception of the microarray technology in 1995, the number of publications applying the technology to cancer (Y-axis) increased exponentially. Since 2005, a plateau-phase has been reached. The number of studies reporting on gene signatures in cancer is still exponentially increasing at the time of writing. The number of clinical trials involving a microarray study in some way, is lagging behind and is currently less than 5% of the total amount of studies reporting on cancer and involving microarrays.

Data analysis strategies

In general, 2 major analysis strategies exist; the unsupervised and the supervised method. An overview is presented in *Table 1*. The unsupervised methods refer to a group of techniques analysing the data without taking phenotypic features associated with the samples into account. In an unsupervised analysis, one is generally interested in the biological relationship between samples. As such, this strategy is suited to discover new biological subgroups within a set of (tumour) samples and therefore unsu-

Table 1. Overview of microarray data analysis strategies, the associated algorithms and their appropriate references for further reading

STRATEGY	SUBSTRATEGY	ALGORITHMS	REFERENCES
Unsupervised analysis	Class discovery	Hierarchical cluster analysis, K-means clustering, Self-organising maps, principal component analysis	12, 13, 14, 15, 16, 17
Supervised analysis	Class comparison	T-, ANOVA-, Mann-Whitney U- and Wilcoxon Signed Rank-testing, Linear models, Significance analysis of microarrays, Permutation testing	12, 18,19, 20, 21
	Class prediction	Nearest centroid classification, Neural networks, Linear discriminant analysis, K-nearest neighbours, Support vector machines	15, 22, 23, 24, 25
	Functional analysis	Gene Set Enrichment Analysis, Global testing, Global ANCOVA models, Pseudo-comparative genomic hybridisation	26, 27, 28, 29, 30

pervised analysis strategies are often referred to as class discovery methods. For example, in a study by Perou et al, the authors performed an unsupervised analysis on a set of breast tumour samples.[8] This resulted in 2 major subgroups associated with the presence or absence of Estrogen Receptor (ER) protein expression. In addition, new subgroups could be identified in both sets. The authors concluded that breast cancer is composed of different subtypes, some of which are biologically more related (e.g the ER+ subtype Luminal A and B) than others.[8] This example demonstrated that class discovery analysis assists in identifying novel tumour subclasses and thus paves the way for more personalised treatment.

The supervised methods differ from the unsupervised methods in that the associated phenotypic data of the tumour samples are integrated in the analysis. The supervised methods can be subdivided in class comparison- and class prediction-methods. As the name suggests, class comparison-methods essentially compare 2 or more classes or groups of tumour samples, mainly to identify differentially expressed genes. This method is well-suited to identify individual biomarkers, but can also be used as a starting point for downstream functional analysis (vide infra) For example, our research group used microarray analysis to molecularly characterise inflammatory breast cancer (IBC). Therefore, we compared a group of IBC samples to a group of nIBC samples.[9] In order to be able to compare both groups, one needs a priori knowledge of the sample or grouping labels and consequently this is a supervised analysis strategy.

The second supervised analysis strategy, class prediction, refers to a set of techniques that is used to design gene signatures. A gene signature is a biomarker consisting of a collection of genes selected for their ability to discriminate with high specificity and sensitivity between 2 or more groups of tumour samples. In essence, each gene comprising a gene signature is a biomarker on its own. For the gene selection procedure, again, this technique requires prior knowledge of the sample or group labels. Gene signatures are among the holy grails in cancer research due to their clinical potential, as they can be used both as prognostic and therapeutic (i.e. predictive) markers. For example, Van 't Veer et al described a gene signature composed of 70 genes able to discriminate between patients with lymph node negative breast cancer who developed distant metastases within 5 years and those who remained metastases-free for more than 5 years.[10] An example of a predictive study is described by Hess et al, who designed a 30 gene signature capable of predicting pathological complete response to neoadjuvant Paclitaxel and Fluorouracil + Doxorubicin + Cyclophosphamide in patients with breast cancer.[11] At the time this review was written, about 2,300 papers associating gene signatures with cancer have been published (*Figure 1*).

From the perspective of tailored treatment, the development of gene signatures allowing for the identification of samples with certain activated pathways is particularly interesting. For example, Creighton et al designed a gene signature predicting the activation of the IGF1-induced signal transduction pathway in breast cancer. They demonstrated that this gene signature is able to predict sensitivity to anti-IGF1R therapy in cell lines and xenografts. Particularly cell lines and xenografts of triple negative breast tumours were predicted to be highly sensitive to anti-IGF1 therapy

(BMS-754807). Indeed, treatment of xenografts or triple negative breast tumours with BMS-754807 in combination with Docetaxel demonstrated significant tumour regression until no tumour was palpable.[32] In addition, the same authors demonstrated that a subgroup of ER+ breast tumours, particularly the endocrine resistant Luminal B tumours, exhibit the expression of the IGF1-activated gene signature.[33] Combined, these data suggest that patients with triple negative or endocrine resistant breast cancer could benefit from anti-IGF1R therapy. More importantly, these data also demonstrate how prediction of pathway activation through gene expression profiling can become a valuable tool in establishing tailored treatment.

A third analysis strategy that needs mentioning is the functional analysis. This too can be regarded as a supervised analysis strategy, because its starting point is either the class prediction or the class comparison analysis. The functional analysis translates lists of genes into biological processes or signal transduction pathways. The rationale behind the functional analysis is, that most genes can play vital roles in different processes or signal transduction pathways. Therefore, it is difficult to unravel the biology based upon lists of individual genes alone. On the other hand, a biological process or signal transduction pathway is usually composed of a unique collection of genes. If these process- or pathway-specific genes can be identified in a list of differentially expressed genes, the associated biological process or signal transduction pathway will be of importance to the biology of the samples that have been used to define the list of differentially expressed genes. In addition, the functional analysis has the advantage of downscaling the amount of possibly involved parameters to a more feasible number. The gene lists associated with biological processes and signal transduction pathways can be found in publicly available databases like Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Transcription Factor databases (TransFac).[27]

In the remaining parts of this review, the aspects of the above outlined analysis strategies (class discovery, class comparison and class prediction) will be discussed in more detail with special focus on potential pitfalls. For more elaborate background on the items and algorithms discussed in the following sections, the reader is directed towards the references in *Table 1*.

Class discovery analysis

Class discovery analysis is an unsupervised analysis directed at analysing the biological relationships between the samples. The biological relationships between samples are defined by how similar the gene expression profiles of the samples are. This implies that the algorithms need a measure of similarity. The most widely used similarity metrics are the Euclidean distance, the correlation coefficient and the Manhattan distance, but many more exist. Class discovery algorithms organise data in such way that samples with similar expression profiles are grouped together and that the gene expression profiles of the samples in different groups are maximally dissimilar. Groups of samples resulting from class discovery algorithms are called clusters and the algorithms themselves are often referred to as cluster algorithms. Examples of these algorithms are hierarchical cluster analysis, k-means cluster analysis and self-organising maps.[12-17]

Another powerful class discovery algorithm that is widely gaining more and more interest is principal component analysis (PCA). Although this algorithm is regarded as an unsupervised analysis method, it should be mentioned separately from the cluster algorithms because no real distance metric is defined from the start. The rationale of PCA is to visualise multidimensional data in 2 or 3 dimensions through data decomposition. An average microarray experiment consists of 100 samples and 20,000 genes. This means that the 100 samples can be visualised in a 20,000-dimensional hyperspace. PCA can be used to visualise these multidimensional datasets by reducing the number of dimensions down to 2 or 3 theoretical dimensions, the principal components. These capture most of the gene expression variation present in the dataset and, as such, preserve the gene expression relationships between the samples.[12]

Although class discovery techniques are extremely powerful, great care must be taken in applying these algorithms. Even though the methods used are objective in the sense that the algorithms are well-defined and reproducible, they are still subjective in that selecting different algorithms or different distance metrics will place different objects into different clusters. Furthermore, clustering unrelated data would still produce clusters, although they might not be biologically meaningful. The challenge is therefore, to select the data and to apply the algo-

rithms appropriately so that the classification is sensible.[12] To obtain a useful analysis while at the same time preserving the expression relationships in the array dataset, it is standard practice to use those genes with a high coefficient of variation and mean expression level. In addition, each clustering result should be analysed for cluster stability or robustness to prevent reporting results that are not supported by the data.

Class comparison analysis

In a class comparison analysis, one is interested in finding lists of genes that are differentially expressed between the conditions during the study. These gene lists can be used to unravel the molecular biology of previously uncharacterised conditions or they can be used to define potential biomarkers to follow-up on treatment or disease progression. Different methods exist to define lists of differentially expressed genes, all derived from the field of traditional statistics, for example T-testing, ANOVA-testing, Mann-Whitney U-testing and Wilcoxon Signed Rank-testing. Also linear models (i.e. linear regression analysis) can be used for this purpose. For example, when studying differential gene expression between ER+ and ER- tumours, the power of each gene in the microarray experiment to correctly predict ER+ from ER- breast tumours is evaluated using linear regression analysis. In this, the predictive power is proportional to the level of differential expression.

The greatest statistical challenge is based on the conclusions concerning the expression of a great number of genes on the basis of a small number of samples. This problem is also referred to as the multiple testing or multiple comparisons problem. The major culprit for this phenomenon is the significance level or alpha-level that is the preset probability that a statistical test will be a false positive. Another way to consider the significance level is that it is the percentage of statistical tests that will be false positive. In microarray analysis, which usually deals with on average 20,000 genes, this means that 1,000 false positives are expected for a significance level of 0.05. For this kind of experiment, the False Discovery Rate (FDR), which is the percentage of false positives in the total amount of significant results, can easily be about 50%. The problem is that one cannot determine which statistical results are true/false positives.[12,18-21]. A number of algorithms have been developed to tackle this problem. The Bonferroni correction is very stringent and aims at reducing the probability of having only one false positive in the full set of comparisons. This method is widely used when dealing with the identification of individual biomarkers. The FDR-correction proposed by Benjamini and Hochberg aims at reducing the expected proportion of false positives.[20] This method is less stringent and is more appropriate when aiming at the molecular characterization of a condition or the identification of gene signatures. FDR-levels of up to 10% are considered appropriate in microarray literature.[12, 18-21] Of note, biologically interesting genes with more elevated FDR-levels (i.e. greater than 10%) can still be relevant. For example, genes with an FDR of 20% still have 80% chance of being true positives. It is standard practice to evaluate such genes with alternative expression profiling techniques (e.g. qRT-PCR), preferentially on a different and larger set of samples.

Class comparison analysis

Perhaps the most promising application of microarrays for expression profiling, is class prediction. In this setting it is not necessary to understand the underlying molecular biology of the condition. Rather, it is a purely statistical exercise in linking a certain pattern of expression to a certain diagnosis or prognosis.[15,22-25] The procedure of classification is a well-established field in statistics, from which a wealth of methods can be drawn. Each modelling method consists of 3 stages: feature selection, model building and model assessment.[15,22-25]

Feature selection involves the process of selecting genes that will be used to construct the classifier. Feature selection should favour informative genes without being too restrictive in their selection criteria. Usually, simple fold change statistics, T-test or ANOVA-statistics (for multiple groups) are used. Also, genes can be selected based on their standard deviation. Genes that ideally compose a gene signature are those genes demonstrating a huge difference in mean or median gene expression between groups of tumour samples (e.g. high fold-change or high T-test or ANOVA-statistics) but show little variation within each group of tumour samples under study (e.g. low within-group standard deviation). After the feature

Key messages for clinical practice

1. Class discovery analysis can be used to identify novel tumour subgroups and, as such, can assist in paving the way for tailored therapy.
2. Class comparison analysis can be used to study the biology of poorly characterised tumour specimens and to identify potential molecular targets for targeted therapy.
3. Class prediction analysis can be used to identify multi-gene biomarkers to assist in therapeutic decision-making and patient prognosis.
4. Gene expression profiling can provide guidance in the establishment of tailored therapy through prediction of signal transduction pathway activation in clinical tumour samples.

selection process, the model should be built. Many model-building procedures exist, including nearest centroid classification, neural networks, linear discriminant analysis, K-nearest neighbours and support vector machines. A popular choice amongst these is the nearest centroid classification, which compares each sample under study to a set of centroids, one for each class. A centroid is the average expression profile of a certain class and is based on the genes selected during the feature selection process. The final step in model building is to define a prediction rule, for example a threshold on the Euclidean distance between the samples and the centroids. When this Euclidean distance exceeds a given value, the sample is not classified in the group represented by the centroid and vice versa. Once the prediction model is constructed, the next step is model assessment and validation, essentially performed on a series of samples not used for feature selection or model-building. In the model assessment stage, metrics such as sensitivity (% of correctly classified positive samples), specificity (% of correctly classified negative samples), accuracy (% correct classifications) and prediction error (% misclassifications) can be examined to determine how well the prediction model is able to correctly classify a series of unknown samples. For example, we designed a gene signature able to predict IBC samples from nIBC samples. Testing this gene signature onto an independent series of IBC/nIBC samples proved that the gene signature correctly classified all IBC samples (sensitivity=100%) and only 1 nIBC sample was misclassified (specificity=91%). The total percentage of misclassifications was 5.8% (=prediction error) and the accuracy was 94.2%.[31] The rationale behind model assessment is, that a classifier can perform well on the samples used for feature selection/model-building, but may perform poor on other samples subject to the same classification procedure. This phenomenon is called overfitting. The best way to perform model assessment is to use an independent dataset, composed of samples not used in the feature selection/model-building phase. This can be done by splitting a group of samples in a training set, used for feature selection/model-building and a test set used for model assessment. If the initial set of samples is too small for splitting, other model assessment procedures exist such as cross-validation or bootstrap analysis.[15,22-25]

Conclusion

Since the introduction of the microarray technology in 1996, the technique has acquired a permanent status in cancer research due to its versatile applicability. The technique can be used to discover new tumour subtypes, to identify single-gene or multiple-gene biomarkers for response to treatment or disease progression and to unravel the molecular biology of previously uncharacterised or poorly characterised conditions. Nevertheless, despite the wide range of possibilities, great care must be taken when analysing data, as many pitfalls exist that can easily trap a researcher into reporting results unsupported by the data. In addition, as

microarray technology is gradually finding its way into clinical trials, knowledge of the possible analysis strategies and existing methods, with their advantages and disadvantages, is important to the clinician in order to be able to interpret the results correctly.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467-70.

2. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457-60.

3. DePrimo SE, Wong LM, Khatry DB, Nicholas SL, Manning WC, Smolich BD, et al. Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. BMC Cancer 2003;3:3.

4. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. J Clin Oncol 2008;26:729-35.

5. Cardoso F, Piccart-Gebhart M, Van't Veer L, Rutgers E. The MINDACT trial: the first prospective clinical validation of a genomic tool. Mol Oncol 2007;1:246-51.

6. Mook S, Van't Veer LJ, Rutgers EJ, Piccart-Gebhart MJ, Cardoso F. Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genomics Proteomics 2007;4:147-55.

7. Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 2007;7:545-53.

8. Perou CM, Sørlie T, Eisen MB, Van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747-52.

9. Van Laere S, Van der Auwera I, Van den Eynden G, Van Hummelen P, Van Dam P, Van Marck E, et al. Distinct molecular phenotype of inflammatory breast cancer compared to non-inflammatory breast cancer using Affymetrix-based genome-wide gene-expression analysis. Br J Cancer 2007;97:1165-74.

10. Van 't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.

11. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol 2006;24:4236-44.

12. Quackenbush. Computational analysis of microarray data. Nat Rev Genet 2001;2:418-27.

13. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.

14. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8:37-49.

15. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform 2005;6:331-43.

16. Eisen MB, Spellman PT, Brown PO, Botstein. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863-8.

17. Azuaje F. Clustering-based approaches to discovering and visualising microarray data patterns. Brief Bioinform 2003;4:31-42.

18. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368-75.

19. Dudoit S. Statistical methods for identifying differentially expressed genes in replicated cDNA micorarray experiments. Stat Sinica 2002;12:111-39.

20. Benjamini Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 1995;57:289-300.

21. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116-21.

22. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23:7332-41.

23. Huang J, Bhandarkar SM. A comparison of physical mapping algorithms based on the maximum likelihood model. Bioinformatics 2003;19:1303-10.

24. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 2008;9:S13.

25. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 2002;99:6562-6.

26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.

27. Stoeckert CJ Jr, Causton HC, Ball CA. Microarray databases: standards and ontologies. Nat Genet 2002;32 Suppl:469-73.

28. Furge KA, Dykema KJ, Ho C, Chen X. Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma. BMC Genomics 2005;6:67.

29. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007;23:980-7.

30. Goeman JJ, Van de Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004;20:93-9.

31. Van Laere SJ, Beissbarth T, Van der Auwera I, Van den Eynden GG, Trinh XB, Elst H, et al. Relapse-free survival in breast cancer patients is associated with a gene expression signature characteristic for inflammatory breast cancer. Clin Cancer Res 2008;14:7452-60.

32. Litzenburger BC, Creighton CJ, Tsimelzon A, Chan BT, Hilsenbeck SG, Wang T, et al. High IGF-IR activity in triple-negative breast cancer cell lines and tumorgrafts correlates with sensitivity to anti-IGF-IR therapy. Clin Cancer Res December 2010 [Epub ahead of print]

33. Creighton CJ, Casa A, Lazard Z, Huang S, Tsimelzon A, Hilsenbeck SG, et al. Insulin-like growth factor-I activates gene transcription programs strongly associated with poor breast cancer prognosis. J Clin Oncol 2008;26:4078-85.