Expert Opinion

Prediction of cancer outcome using DNA microarray technology: past, present and future

Olivier Gevaert† & Bart De Moor
Katholieke Universiteit Leuven, Department of Electrical Engineering ESAT-SCD-Sista, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Background: The use of DNA microarray technology to predict cancer outcome already has a history of almost a decade. Although many breakthroughs have been made, the promise of individualized therapy is still not fulfilled. In addition, new technologies are emerging that also show promise in outcome prediction of cancer patients. *Objective*: The impact of DNA microarray and other 'omics' technologies on the outcome prediction of cancer patients was investigated. Whether integration of omics data results in better predictions was also examined. *Methods*: DNA microarray technology was focused on as a starting point because this technology is considered to be the most mature technology from all omics technologies. Next, emerging technologies that may accomplish the same goals but have been less extensively studied are described. *Conclusion*: Besides DNA microarray technology, other omics technologies have shown promise in predicting the cancer outcome or have potential to replace microarray technology in the near future. Moreover, it is shown that integration of multiple omics data can result in better predictions of cancer outcome; but, owing to the lack of comprehensive studies, validation studies are required to verify which omics has the most information and whether a combination of multiple omics data improves predictive performance.

Keywords: cancer, microarray technology, omics, prognosis

Expert Opin. Med. Diagn. (2009) 3(2):157-165

1. Introduction

In the past century the clinical management of cancer patients was based on histopathology or clinical data such as patient history, laboratory analysis or ultrasound parameters. Also, empirical knowledge present in the literature or personal experience of the clinician plays an important role. This implies that the current management of cancer patients potentially suffers from a considerable amount of inter-observer variability. More importantly, not all information that is clinically relevant can be extracted from the data that clinicians have access to at present. The fundamental mechanisms underlying carcinogenesis on a molecular level are not taken into account to make the optimal decision regarding therapy choice, thus influencing the outcome of the patient. More generally put, owing to the absence of any knowledge of the biological processes in the tumor at a molecular level, it is not possible to tailor the therapy to the patient based solely on clinical data.

In the past decade, however, a new technology has been developed that has the potential to change cancer management. DNA microarray technology allows profiling of the expression of thousands of genes at once, possibly representing the whole genome. Usually a microarray consists of a selection of probes representing several genes applied onto a solid surface [1,2]. Reverse transcribed

informa
healthcare

10.1517/17530050802680172 © 2009 Informa UK Ltd ISSN 1753-0059

mRNA extracted from a tumor sample is first labeled and then hybridized with the probes on this surface. After scanning the array, this results in expression levels of thousands of genes for every tumor sample that is hybridized. DNA microarray technology thus captures the molecular signals in the tumor at the transcriptional level and possibly contains information reflecting the outcome of the individual patient.

After the introduction of DNA microarray technology, the data resulting from this technology were used for many applications in cancer, such as class discovery, prediction of diagnosis, prognosis or therapy response [3-9]. In addition, at the time of the first applications of DNA microarray technology on cancer, the first draft of the human genome was published and unlocked a vast resource of data that were previously unavailable [10]. For the first time it was possible to have an idea of all genes in the human genome. This knowledge allowed probes to be designed for all predicted genes, both known and unknown. This accelerated research on a genome-scale level increased the number of probes on a microarray and accelerated the uptake of microarrays in cancer.

One issue that was not foreseen is the high dimensionality of the data resulting from DNA microarray technology. As this technology allows probing of every possible gene, microarray data often contain > 25,000 expression values per patient. Obviously, it is impossible for a clinician to interpret these data directly and identify the genes that contain outcome-related information. This problem sparked considerably interest in the statistical and mathematical community to develop methods to unlock the outcome-related information in microarray data. A further complexity, caused by the relatively high financial cost of microarrays, is that microarray data sets are often small when looking at the number of tumors that is typically profiled, ranging from tens in early studies [4] to at most a few hundred samples in the latest studies [11]. This led statisticians to identify the curse of dimensionality, which entails that the number of data points to represent adequately a space increases super-exponentially with increasing dimensionality [12]. This makes it statistically challenging to identify a subset of genes that contain outcome-related information when few tumors are profiled on a whole genome scale. Anyway, many methods have been proposed to analyze microarray data for different purposes, such as discovering new classes and prediction of outcomes related to diagnosis, prognosis or drug response [13-15].

In this review, the potential of data generated by DNA microarray technology and how they influence prediction of cancer outcome are discussed. First, two different microarray technologies that exist to unlock the transcriptome are described briefly, and methods to analyze the data. Next, an overview is given of the history of DNA microarray technology in cancer. This overview is not meant to be exhaustive; instead the milestones and problems that arose when using DNA microarray technology to analyze cancer tissue are focused on. Most examples come from breast cancer, because in the early days of DNA microarray technology this was the most extensively studied cancer site. Later, more studies were published investigating the use of DNA microarray technology on other cancer sites such as colon cancer [16], lung cancer [17] and ovarian cancer [18]. Finally, other, less mature, technologies are briefly introduced that profile other levels of the central dogma of molecular biology or technologies that may replace DNA microarray technology altogether. These technologies are often called 'omics' in the computational biology and bioinformatics community.

2. DNA microarray technology and data analysis

In the Introduction, the basic idea of DNA microarray technology was described. However, two quite different technologies were introduced at about the same time. First, in the case of cDNA microarray technology [19,20], the probes are long DNA sequences that are mechanically spotted on a glass slide. Next, different dyes are used for the test and control sample. Then both the test (e.g., breast tumor sample) and the control sample (e.g., normal breast tissue sample) are applied onto the glass slide and competitively hybridize with the probes on the array. Next, the array is scanned by a laser and the fluorescent intensities of both the test and the control dye are measured. If only the test sample bounded to a probe only its corresponding dye will light up, and vice versa for the control sample. The result is a relative expression of each gene of the test sample versus the control sample. As two samples are simultaneously hybridized to the array, this is often called a two-channel array.

The other technology is called oligonucleotide microarray and is commercially available [2]. In this case the probes are very short, namely 25 bases, but multiple probes are present representing the same gene. Also, the probes are not spotted but grown on an array using a lithographic process similar to how computer chips are made [21]. Owing to this process there is no variation in the amount of probe material that is present on the array. As such each sample is separately hybridized to the array and after scanning each gene expression value is represented by a single fluorescence intensity. This array system is therefore called a one-channel array.

After these experimental procedures normalization of the data is required to remove any non-biological noise coming from different labeling, array or other effects. This process is different for cDNA [22] and oligonucleotide arrays [23]. An overview of normalization methods has been given in [24]. Subsequent microarray data analysis methods can be subdivided in unsupervised and supervised methods. Semi-supervised methods, which make use of both labeled and unlabeled data, exist as well but are beyond the scope of this review.

Unsupervised analysis or class discovery is an unbiased analysis of microarray data. No prior class information is used and clustering methods are used to group the samples

based solely on the microarray data. In Kerr *et al.* [25], an overview is given of many clustering algorithms with different properties and limitations.

Second, in the case of supervised analysis, also called class prediction, previous knowledge is taken into account. Often tumor samples for microarray studies come from well-defined groups, for example good and poor prognosis patients. The aim is then to identify genes or develop a model that is able to assign patients to the good or poor prognosis class based on the microarray data of its corresponding tumor. The simplest supervised method is the identification of differentially expressed genes using statistical tests. Parametric or non-parametric univariate statistical tests such as the t-test or the Wilcoxon rank sum test (often also called the Mann-Whitney test) can be used to give each gene a p-value. Then the genes are ordered based on ascending p-values, with the top genes the most important genes necessary to differentiate the classes. More advanced methods do not solely rank the genes but build a model based on microarray data with or without gene selection to predict the classes. A few examples of modeling strategies are decision trees [26], support vector machines [15,27], nearest neighbors [28,29], neural networks [30] and linear discriminant analysis [4], and many others exist [31]. For a more theoretical description of these methods, refer to the book by Mitchell [32] and to a review by Slonim for an overview of both class discovery and class prediction methods [33]. In addition, combined class prediction and dimensionality reduction, for example by selecting genes in some manner, has been shown to improve classification results [15].

3. DNA microarray technology and cancer

One of the first landmark studies using microarray data to analyze tumor samples was done by Golub *et al.* [34]. This study on human acute leukemia showed that it was possible to use microarray data to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without any previous knowledge. For the first time the potential of microarray data was shown by illustrating its use in discovering new classes using the previously introduced class discovery methods (i.e., unsupervised analysis) and, second, by using microarray data to assign tumors to known classes (i.e., supervised analysis). The authors of this first landmark paper using microarray data already acknowledge the potential of microarray data to predict therapy response or survival. Class prediction gives the clinician an unbiased method to predict the outcome of the cancer patient instead of traditional methods based on histopathology or empirical clinical data, which do not always reflect patient outcome.

Perou *et al.* [35] did a similar analysis using hierarchical clustering on breast cancer and found different groups of breast tumors. This was confirmed further and expanded in a follow-up study that resulted in four different breast tumor types: luminal A, luminal B, basal and ERBB2 (also known as HER-2/neu) overexpressing [36]. Moreover, in the same study the authors showed that these subtypes also correlated with clinical outcome in an independent data set. Bhattacharjee *et al.* [37] did a similar analysis on lung cancer patients and also discovered subtypes based on microarray data that correlated with clinical outcome.

Alon *et al.* [38] adopted a two-way clustering method whereby both genes and tumors were clustered. They showed that colon tumors and normal colon tissues were separated based on the microarray data. Also, they showed that co-regulated families of genes also clustered together.

Tibshirani *et al.* [39] built further on the development of supervised microarray data analysis methods by developing the nearest shrunken centroid method, also known as PAM. This technique not only allows predicting of classes, but also tries to limit the number of genes necessary to make the prediction. By limiting the number of genes, it is possible to develop cheaper methods to make a diagnostic test, such as smaller microarrays or quantitative PCR.

After these studies research groups focused more on class prediction because of its potential to influence the clinical management of cancer and improve outcome prediction of cancer patients. Some studies on breast [3,40] and lung cancer [41] were published reporting on models to predict prognosis or prognostic factors of cancer patients.

One of the breast cancer studies, the van't Veer study, was internally validated further in a new set of patients [9] and, additionally, an external retrospective validation was done in five European centers (TRANSBIG consortium) [42], however not without criticism [43-46]. In addition, another issue arose when comparing prognostic models from other labs. The van't Veer prognostic model contained the expression levels of 70 genes constructed based on 99 patients, whereas the Wang prognostic model [47], studying the same breast cancer population and outcome, consisted of 76 genes derived from 286 tumor samples. Both models, however, shared only three genes in common. Moreover, the predictive performance of both models decreased drastically when applied to each other's data [48]. A mathematical analysis, however, showed that many sets of 70 genes can be used to predict survival in the van't Veer data set. The authors concluded that it is relatively easy to create prognostic models, but that gene membership of these models is not necessarily tied to a gene's importance in carcinogenesis. Also, to have a unique prognostic model, unrealistic numbers of patients have to be entered in a microarray study owing to the high dimensionality of the data [49].

The van't Veer study and its follow-up study by van de Vijver have resulted in a prognostic test called MammaPrint, which is FDA approved and commercially available in the US. The test was first verified for feasibility in 16 hospitals in the Netherlands [50], however nonrandomized such that the exact predictive value is not yet known. The MINDACT trial (Microarray in Node Negative Disease May Avoid Chemotherapy), which is a 6000-patient randomized

multi-centric trial, aims to answer this question by comparing treatment decisions based on the MammaPrint prognostic test with current management of breast cancer patients [51]. The results of this trial, however, will only come after many years [52].

Similarly in the US, another test called Oncotype DX based on a 21-gene recurrence score was developed [53] and tested extensively [54,55]. This test is now also the subject of a large trial called TAILORx, which will asses whether genes associated with recurrence are able to classify patients into good and poor prognosis [56,57].

A major criticism of previous microarray studies was voiced by Dupuy and Simon [58]. They analyzed all microarray studies focusing on predicting cancer outcomes published in 2004. Many of them contained at least one of three serious flaws. The most important one related to this review is biased estimation of accuracy. Owing to the curse of dimensionality as explained earlier, it is critical that models based on microarray data are validated on an independent test set with or without cross-validation techniques. An independent test set or left out samples in a cross-validation loop cannot be used for determining pre-processing parameters, selection of differentially expressed genes, model building or model selection. Dupuy and Simon showed that half of the studies, including studies published in high-impact journals, contained flaws.

4. Other data providers

DNA microarray technology is often considered to be the most mature technology, whereas other technologies, probing different layers of the central dogma of molecular biology, have different properties and may also contain outcome-related information. It is hypothesized that microarray data are a good approximation of the final amount of protein product, but processes such as alternative splicing and post-translational modifications can influence this significantly. Therefore, studying other layers could add important complementary data to predict the outcome of cancer patients. A few other technologies are discussed here that potentially provide data on a genome-scale level for outcome prediction of cancer patients: array comparative genomic hybridization (arrayCGH) [59], mass spectrometry-based proteomics [60-62], protein/antibody microarrays [63], microRNA [64] and DNA methylation [65]. Finally, potential new technologies are discussed that may replace DNA microarray technology in the near future and make possible the study of many of the above-mentioned biological processes using one platform.

ArrayCGH allows study of the genome by identifying copy number variants with microarray technology. This technology is very similar to DNA microarray technology but probes the genome instead of the transcriptome. ArrayCGH measures variations in DNA copy number within the entire genome of a disease sample compared with a normal sample. ArrayCGH was first applied to breast cancer in 1998 and offers a much higher resolution than traditional CGH [66]. Moreover, arrayCGH can couple the copy number alterations directly to the genomic sequence. This makes this technology a candidate to search for genetic alterations relevant for cancer outcome on a genome-wide scale.

In addition, it only recently became clear that the copy number of genes could differ greatly between individuals [67]. It is still unclear what the 'normal' human genome looks like but the number of genes affected by this process is bigger than expected. This raises the question of how these copy number variations are related to disease. Recently, many studies have shown that copy number variations are related to cancer outcomes in many cancer sites [68-72]. Many defects in human development leading to cancer are due to gains and losses of chromosomes and chromosomal segments. These aberrations, defined as regions of aberrantly increased or decreased DNA copy number, can be detected using arrayCGH, making this a possible alternative for DNA microarray technology.

The second technology that is emerging as an important provider of data for studying cancer outcome is mass spectrometry-based proteomics. This technology has already been around a long time but only recently moved to high-throughput analysis of complex proteins samples [73,74]. Proteomics technology based on mass spectrometry was first used in cancer for diagnostic purposes by Petricoin *et al.*, based on surface-enhanced laser desorption and ionization (SELDI) technology [75]. This study raised many controversies [76-80], among which reproducibility may be the most important one. Others have taken this into account; for example, Taguchi *et al.* [81] used matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) to profile pretreatment serum of patients with non-small cell lung cancer (NSCLC). Their model, based on eight distinct m/z values, could predict response to treatment with an EGFR inhibitor. The multi-centric nature of this study is a strength; however, this study's weakness is the lack of identification of the eight m/z values. In the near future, however, mass spectrometry has the potential to have a significant impact on outcome prediction of cancer patients, especially when taking into account its ability to detect post-translational modifications [82-84].

Another approach to studying the proteome is by using protein microarray (also often called antibody microarrays) [85]. Whereas mass spectrometry-based proteomics is considered to be an unbiased approach, protein/antibody microarrays are a focused approach to study the proteome quantitatively [63]. A protein microarray consists of several affinity reagents attached to a solid surface. Many different techniques assist in accomplishing this; see Spisak *et al.* for an overview of different technologies [86]. An important evolution is that currently up to a few thousand proteins can be assayed at the same time, which increases its potential in studying cancer outcome significantly. A recent application of this technology showed potential in the diagnosis of pancreatic cancer using serum samples [87]. Similarly, the same group also showed

that serum samples applied to antibody arrays can be used to distinguish metastatic breast cancer from healthy controls [88]. Finally, this technology can also be used to distinguish differential phosphorylation in non-small cell lung cancer [89].

Next, the epigenome is receiving increased attention. More specifically, DNA methylation is being considered as an important process influencing cancer outcome. DNA methylation involves the binding of a methyl group to CpG islands in the genome [90]. CpG islands are often found in the regulatory regions of genes and are often associated with transcriptional inactivation. The first attempt to characterize the methylome in cancer was done by Costello *et al.* [91]. They characterized the methylation status of 1184 CpG islands in 98 primary tumors. Before this study, DNA methylation in cancer was limited to 15 of the estimated 45,000 CpG islands in the human genome. Costello and colleagues showed that DNA methylation status was different between different cancer sites and they hypothesized that different tumor subtypes exist based on the DNA methylation profile. Most studies focus on using DNA methylation for early detection of cancer [92], however its use in predicting prognosis has also been shown [93-97].

Next, the discovery of microRNAs, a group of small RNAs, has shown promise in predicting cancer diagnosis and prognosis [98,99]. A microRNA is typically between 20 and 22 nucleotides and is thought to function as transcription or post-transcriptional regulators by binding with their target mRNAs [100]. One of the first studies investigating expression of all known microRNAs in cancer was done by Lu *et al.* [101]. They studied 217 microRNAs in 334 samples representing different tissues and tumors. The results showed that although the number of profiled microRNAs was limited, different expression was seen in tumors and normal samples. Moreover, microRNA expression profiles had better classification performance compared with mRNA on the same samples. MicroRNA profiles have also shown promise in predicting prognosis of chronic lymphocytic leukemia [102,103], breast cancer [104,105], lung cancer [106] and colon cancer [107].

Finally, the rise of second-generation sequencing technology may subsume DNA microarray technology in the near future, especially the less expensive platforms with short reads such as Solexa [108], SOLiD [109] or HeliScope [110], which are at present not suited for *de novo* assembly owing to short read length, but are ideal for counting applications such as studying gene expression or microRNA expression [111]. As soon as these technologies become cheaper, DNA microarray technology may be replaced by next-generation sequencing sooner than expected [112,113]. In addition, these new sequencing platforms offer a wide range of applications, such as studying gene expression, miRNA expression and copy number variation, all previously measured using separate microarray technologies, and also add new information such as alternative splicing [114].

5. Conclusion

Based on the numerous publications investigating the use of DNA microarray technology to predict outcome in different cancer sites, this technology seems to be the most mature technology from all the omics. Moreover, judging by the two randomized trials that have been set up, DNA microarray technology is expected to be the first technology used in the clinical management of cancer and thus have an influence on the outcome of cancer patients. On the other hand, other omics are arising that may be better predictors of cancer outcome than DNA microarray technology. Several studies have been described that have shown that arrayCGH, mass spectrometry-based proteomics, DNA methylation or microRNAs are also able to predict cancer outcome. It is not known at present which omics has the most outcome-related information. Finally, second-generation sequencing may replace general microarray technology in the near future [115,116].

6. Expert opinion

Few studies have been done evaluating different omics technologies on the same patients to assess which omics layer has the most outcome-related information. Furthermore, there are few groups investigating how information from many omics technologies together with the clinical data of the patient can be integrated for outcome prediction. Microarray data are high dimensional, characterized by many variables and few observations. Moreover, this technique suffers from a low signal-to-noise ratio, which causes instability in gene signatures. Integration of other sources of information could be important to counter randomly generated differences in expression levels. Nevertheless, the focus in most studies is on the microarray analysis, whereas, for example, clinical data are not used in the same manner. Clinical data include, for example, patient history, laboratory analysis or ultra-sound parameters. These data are still the basis of research and fully guide the clinical management of cancer. Model development where both clinical variables and gene expression are combined should give an answer to the question of whether gene expression measurements improve outcome prediction independently of clinical factors.

The authors have developed methods using both Bayesian networks and kernel methods that are able to integrate clinical and microarray data using previously discussed publicly available data sets [3,9]. In Gevaert *et al.* [117], it was shown that partial integration of clinical and microarray data improved prognosis prediction of breast cancer patients. Also, Daemen *et al.* integrated clinical and microarray data based on kernel methods [118] and saw similar results. Similar conclusions have been reached by others [119]. For example, a microarray data analysis from a large lung cancer study showed that models including clinical variables performed in most cases better than those without clinical variables [11].

Similarly, data from new technologies introduced earlier could contain complementary information that is not present in the clinical or microarray data and thus improve predictive performance. A proof-of-principle approach has been developed integrating microarray and proteomics data gathered at two important time points during therapy of rectal cancer patients using kernel integration methods [120]. This study showed that integrating more than one data source performed better than any data source alone when predicting therapy response. Also, this study showed that integrating genome scale information at different time points during therapy also has a positive effect.

Finally, it has been shown that other sources of data such as literature abstracts can be integrated using text mining techniques and improve predictive performance [121]. Moreover, this model is sufficiently general such that other sources of information such as pathways and known protein–protein interactions can be integrated using the same framework [122]. In addition, other groups have followed the same approach [123].

At present, it is unknown which technology and thus which level of molecular biology is the most relevant for outcome prediction. To be able to assess the omics platform with the most outcome-related information, it is required that studies be designed where multiple technologies are applied on the same tumor samples in sufficiently large number, or new technologies, such as second-generation sequencing, providing information on multiple omics layers. The authors acknowledge that this will increase costs substantially, however this is a necessary investment to move away from costly empirical models to patient-tailored therapy [52,124,125]. It is not unthinkable that the best model to predict cancer outcome could be a combination of two downregulated genes, a methylated gene, a deleted region on a certain chromosome and the presence of a phosphorylated protein.

Declaration of interest

The authors state no conflict of interest and have received no payment in preparation of this manuscript.

Bibliography

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. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675-80
3. Veer V, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6
4. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7
5. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790-5
6. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behaviour. Cancer Cell 2002;1:203-9
7. Spentzos D, Levine DA, Ramoni MF, et al. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. J Clin Oncol 2004;22:4700-10
8. Spentzos D, Levine DA, Kolia S, et al. Unique gene expression profile based on pathologic response in epithelial ovarian cancer. J Clin Oncol 2005;23:7911-8
9. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999-2009
10. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. Science 2001;291:1304-51
11. Shedden K, Taylor J, Enkemann S, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008;14(8):822-7
12. Clarke R, Ressom H, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8:37-49
13. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116-21
14. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567-72
15. Pochet N, De Smet F, Suykens J, De Moor B. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004;20:3185-95
16. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745-50
17. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790-5
18. Schwartz DR, Kardia SL, Shedden KA, et al. Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. Cancer Res 2002;62:4722-9
19. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467-70
20. DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457-60
21. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. BioTechniques 1995;19:442-7
22. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of Methods for Image Analysis on cDNA Microarray Data. J Comput Graphical Stat 2002;11(1):108-36

23. Irizarry R, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat 2003;4:249-64
24. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002;32(Suppl):496-501
25. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med 2008;38:283-93
26. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007;356:11-20
27. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001;98:15149-54
28. Berrar D, Bradbury I, Dubitzky W. Instance-based concept learning from multiclass DNA microarray data. BMC Bioinformatics 2006;7:73
29. Prasad NB, Somervell H, Tufano RP, et al. Identification of genes differentially expressed in benign versus malignant thyroid tumors. Clinical cancer research : an official journal of the American Association for Cancer Research 2008;14:3327-37
30. Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-9
31. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567-72
32. Mitchell T. Machine Learning. McGraw-Hill Science/Engineering/Math, 1997
33. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 2002;32(Suppl):502-8
34. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7
35. Perou CM, SÃ¸rlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747-52
36. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003;100:8418-23
37. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790-5
38. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745-50
39. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567-72
40. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98:11462-7
41. Beer D, Kardia S, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816-24
42. Buyse M, Loi S, Van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183-92
43. Ransohoff DF. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715-7
44. Kunkler IH. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715-7
45. Kopans DB. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715-7
46. Helmbold P, Haerting J, Kölbl H. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715-7
47. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671-9
48. Ein-Dor L, Kela I, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171-8
49. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006;103:5923-8
50. Bueno-De-Mesquita JM, van Harten WH, Retel VP, et al. Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). Lancet Oncol 2007;8:1079-87
51. Cardoso F, Van't Veer L, Rutgers E, et al. Clinical application of the 70-gene profile: the MINDACT trial. Journal of clinical oncology: official journal of the American Society of Clinical Oncology 2008;26:729-35
52. Sotiriou C, Piccart M. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 2007;7:545-53
53. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817-26
54. Esteva FJ, Sahin AA, Cristofanilli M, et al. Prognostic role of a multigene reverse transcriptase-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. Clin Cancer Res 2005;11:3315-9
55. Habel L, Shak S, Jacobs M, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res 2006;8:R25
56. Sparano JA. TAILORx: trial assigning individualized options for treatment (Rx). Clin Breast Cancer 2006;7:347-50
57. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. Journal of clinical oncology: official journal of the American Society of Clinical Oncology 2008;26:721-8
58. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147-57
59. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37(Suppl):S11-17
60. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198-207
61. Dalton WS, Friend SH. Cancer biomarkers–an invitation to the table. Science 2006;312:1165-8
62. Koomen JM, Haura EB, Bepler G, et al. Proteomic contributions to personalized cancer care. Mol Cell Proteomics 2008;7(10):1780-94

63. MacBeath G. Protein microarrays and proteomics. Nat Genet 2002;32:526-32
64. Calin GA, Liu CG, Sevignani C, et al. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. Proc Natl Acad Sci USA 2004;101:11755-60
65. Esteller M. Epigenetics in cancer. N Engl J Med 2008;358:1148-59
66. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37(Suppl):S11-17
67. Redon R, Ishikawa S, Fitch K, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54
68. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37(Suppl)
69. Fridlyand J, Snijders AM, Ylstra B, et al. Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 2006;6
70. Snijders A, Schmidt B, Fridlyand J, et al. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. Oncogene 2005;24:4232-42
71. Garnis C, Lockwood W, Vucic E, et al. High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. Int J Cancer 2006;118:1556-64
72. Chin K, DeVries S, Fridlyand J, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006;10:529-41
73. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198-207
74. Cox J, Mann M. Is proteomics the new genomics? Cell 2007;130:395-8
75. Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572-7
76. Rockhill B. Proteomic patterns in serum and identification of ovarian cancer. Lancet 2002;360
77. Elwood M. Proteomic patterns in serum and identification of ovarian cancer. Lancet 2002;360
78. Pearl DC. Proteomic patterns in serum and identification of ovarian cancer. Lancet 2002;360
79. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20:777-85
80. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4
81. Taguchi F, Solomon B, Gregorc V, et al. Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. J Natl Cancer Inst 2007;99:838-46
82. Dalton WS, Friend SH. Cancer biomarkers–an invitation to the table. Science 2006;312:1165-8
83. Koomen JM, Haura EB, Bepler G, et al. Proteomic contributions to personalized cancer care. Mol Cell Proteomics 2008
84. Hanash S, Pitteri S, Faca V. Mining the plasma proteome for cancer biomarkers. Nature 2008;452:571-9
85. Wingren C, Borrebaeck CAK. Antibody microarrays: Current status and key technological advances. OMICS 2006;10:411-27
86. Spisak S, Tulassay Z, Molnar B, Guttman A. Protein microchips in biomedicine and biomarker discovery. Electrophoresis 2007;28:4261-73
87. Ingvarsson J, Wingren C, Carlsson A, et al. Detection of pancreatic cancer using antibody microarray-based serum protein profiling. Proteomics 2008;8:2211-9
88. Carlsson A, Wingren C, Ingvarsson J, et al. Serum proteome profiling of metastatic breast cancer using recombinant antibody microarrays. Eur J Cancer 2008;44:472-80
89. VanMeter AJ, Rodriguez AS, Bowman ED, et al. Laser Capture Microdissection and Protein Microarray Analysis of Human Non-small Cell Lung Cancer: differential epidermal growth factor receptor (EGPR) phosphorylation events associated with mutated EGFR compared with wild type. Mol Cell Proteomics 2008;7:1902-24
90. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer 2003;3:253-66
91. Costello JF, Frühwald MC, Smiraglia DJ, et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet 2000;24:132-8
92. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer 2003;3:253-66
93. Esteller M. Epigenetics in cancer. N Engl J Med 2008;358:1148-59
94. Kawakami K, Brabender J, Lord RV, et al. Hypermethylated APC DNA in plasma and prognosis of patients with esophageal adenocarcinoma. J Natl Cancer Inst 2000;92:1805-11
95. Alaminos M, Davalos V, Cheung NK, et al. Clustering of gene hypermethylation associated with clinical risk groups in neuroblastoma. J Natl Cancer Inst 2004;96:1208-19
96. Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. Ann Rev Pharmacol Toxicol 2005;45:629-56
97. Wei SH, Balch C, Paik HH, et al. Prognostic DNA methylation biomarkers in ovarian cancer. Clinical cancer research: an official journal of the American Association for Cancer Research 2006;12:2788-94
98. Cho WCS. OncomiRs: the discovery and progress of microRNAs in cancers. Mol Cancer 2007;6:60
99. Blenkiron C, Miska EA. MiRNAs in cancer: approaches, aetiology, diagnostics and therapy. Hum Mol Genet 2007;16:R106-R113
100. Calin G, Croce C. MicroRNA-Cancer Connection: The Beginning of a New Tale. Cancer Res 2006;66:7390-4
101. Lu J, Getz G, Miska E, et al. MicroRNA expression profiles classify human cancers. Nature 2005;435:834-8
102. Calin GA, Liu CG, Sevignani C, et al. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. Proc Natl Acad Sci USA 2004;101:11755-60
103. Calin GA, Trapasso F, Shimizu M, et al. Familial cancer associated with a polymorphism in ARLTS1. N Engl J Med 2005;352:1667-76
104. Iorio MV, Ferracin M, Liu CG, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res 2005;65:7065-70
105. Murakami Y, Yasuda T, Saigo K, et al. Comprehensive analysis of microRNA

expression patterns in hepatocellular carcinoma and non-tumorous tissues. Oncogene 2005

106. Yanaihara N, Caplen N, Bowman E, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 2006;9:189-98
107. Schetter AJ, Leung SY, Sohn JJ, et al. MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. JAMA 2008;299:425-36
108. Bentley D. Whole-genome re-sequencing. Curr Opin Genet Dev 2006;16:545-52
109. Shendure J, Porreca G, Reppas N, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science 2005;309:1728-32
110. Harris T, Buzby P, Babcock H, et al. Single-Molecule DNA Sequencing of a Viral Genome. Science 2008;320:106-9
111. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotech 2008;26:1135-45
112. Ledford H. The death of microarrays? Nature 2008;455:847
113. Shendure J. The beginning of the end for microarrays? Nat Meth 2008;5:585-7
114. Kahvejian A, Quackenbush J, Thompson J. What would you do if you could sequence everything? Nat Biotech 2008;26:1125-33
115. Shendure J. The beginning of the end for microarrays? Nat Meth 2008;5:585-7
116. Ledford H. The death of microarrays? Nature 2008;455:847
117. Gevaert O, Smet FD, Timmerman D, et al. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22
118. Daemen A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference 2007; 2007:5411-5
119. Stephenson AJ, Smith A, Kattan MW, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer 2005;104:290-8
120. Daemen A, Gevaert O, De Bie T, et al. Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2008;166-77
121. Gevaert O, Van Vooren S, De Moor B. Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2008;279-90
122. Gevaert O, Van Vooren S, De Moor B. A Framework for Elucidating Regulatory Networks Based on Prior Information and Expression Data. Ann NY Acad Sci 2007;1115:240-8
123. Djebbari A, Quackenbush J. Seeded Bayesian Networks: constructing genetic networks from microarray data. BMC Syst Biol 2008;2
124. Nevins JR, Huang ES, Dressman H, et al. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum Mol Genet 2003;12 Spec No 2
125. Pittman J, Huang E, Dressman H, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. PNAS 2004;101:8431-6

Affiliation

Olivier Gevaert[†] & Bart De Moor
[†]Author for correspondence
Katholieke Universiteit Leuven,
Department of Electrical Engineering
ESAT-SCD-Sista,
Kasteelpark Arenberg 10,
3001 Leuven, Belgium
Tel: +32 16 328646; Fax: +32 16 32;
E-mail: olivier.gevaert@esat.kuleuven.be