University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 179

Prognosis of Stage II Colon Cancer by Non-Neoplastic Mucosa Gene Expresssion Profiling

Alain Barrier* Sandrine Dudoit†

et al.‡

*Dept of Digestive Surgery, Hopital Tenon, Paris; INSERM U444, Faculte de Medecine Saint-Antoine, Universite Pierre et Marie Curie, Paris, alain.barrier@tnn.ap-hop-paris.fr

†Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

‡

http://biostats.bepress.com/ucbbiostat/paper179

Prognosis of Stage II Colon Cancer by Non-Neoplastic Mucosa Gene Expresssion Profiling

Alain Barrier, Sandrine Dudoit, and et al.

Abstract

Aims. This study assessed the possibility to build a prognosis predictor, based on non-neoplastic mucosa microarray gene expression measures, in stage II colon cancer patients. Materials and Methods. Non-neoplastic colonic mucosa mRNA samples from 24 patients (10 with a metachronous metastasis, 14 with no recurrence) were profiled using the Affymetrix HGU133A GeneChip. The k-nearest neighbor method was used for prognosis prediction using microarray gene expression measures. Leave-one-out cross-validation was used to select the number of neighbors and number of informative genes to include in the predictor. Based on this information, a prognosis predictor was proposed and its accuracy estimated by double cross-validation. Results. In leave-one-out cross-validation, the lowest number of informative genes giving the lowest number of false predictions (3 out of 24) was 65. A 65-gene prognosis predictor was then built, with an estimated accuracy of 79%. Genes included in this predictor suggested branching signal transduction pathways with possible extensive networks between individual pathways. It also included genes coding for proteins involved in immune surveillance. Conclusion. This study suggests that one can build an accurate prognosis predictor for stage II colon cancer patients, based on non-neoplastic mucosa microarray gene expression measures.

Introduction

Despite numerous clinical trials, the benefit of adjuvant chemotherapy in the treatment of stage II colon cancer patients has never been proved in a randomized study. In most meta-analyses, there is a trend towards a benefit of adjuvant chemotherapy, but statistical significance is not reached [1]. Thus, this benefit seems to exist but, as it is slight, studies are not powerful enough to demonstrate it. This ambiguous situation is perfectly summarized by the conclusion of the 2004 recommendations of the American Society of Clinical Oncology [2]: "Direct evidence from randomized controlled trials does not support the routine use of adjuvant chemotherapy for patients with stage II colon cancer. Patients and oncologists … are justified in considering the use of adjuvant chemotherapy, particularly for those patients with high-risk stage II disease. … Patients with stage II disease should be encouraged to participate in randomized trials".

Including all stage II colon cancer patients in a randomized trial is debatable. Even if a properly-designed study, comprising thousands of patients, demonstrated a statistically significant benefit of adjuvant chemotherapy, it may not be logical to conclude that this treatment should be given to all stage II patients. Such a conclusion would not take into account that three fourths of the patients are cured by surgery alone and would lead to administering to all patients a treatment that would be useful for only a few. Another more rational approach would be to identify a subgroup of patients at high risk of recurrence, thus more likely to benefit from adjuvant chemotherapy, and to include only these selected patients in a randomized trial. This presupposes finding accurate prognosis predictors for stage II colon cancer patients.

Microarray gene expression profiling has been reported to accurately predict the prognosis of several malignant tumors (breast carcinomas [3,4], lung carcinomas [5,6],

lymphomas [7,8]). Thus, by analogy with these tumors, it may be postulated that gene expression profiling represents a valuable tool in predicting the prognosis of stage II colon cancer patients and thereby in identifying a subgroup of patients at high risk of recurrence. To date, this hypothesis has only been addressed in the study of Wang et al. [9], with good results (overall prediction accuracy of 78%).

The present study aimed to assess the possibility to build a microarray-based prognosis predictor for stage II colon cancer patients using non-neoplastic mucosa gene expression profiles. The rationale for studying the non-neoplastic mucosa, in contrast to tumor tissue as in Wang et al [9], may be summarized as follows. There is an increasing evidence that interactions between stromal and cancer cells are a prerequisite for metastases to occur [10]. However, it remains unclear whether this metastatic potential originates in cancer cells and/or in stromal compartments. Metastatic potential may be present from the start of the tumor [11,12]. Accepting this theory, non-neoplastic mucosa on which the tumor has arisen may contain some helpful information. Non-neoplastic mucosa mRNA samples from 24 patients, with homogeneous disease (stage II) and postoperative treatment (no adjuvant chemotherapy), but different outcomes (10 with metastatic recurrence, 14 with no recurrence), were profiled using the *Affymetrix HGU133A* GeneChip.

Materials and Methods

Patients and samples

Twenty-four patients operated on for a stage II colonic adenocarcinoma in the Department of Digestive Surgery of the Hospital Tenon between 1997 and 1999 were included in this study. None of these 24 patients had any adjuvant chemotherapy. Patients were evaluated at 3-month intervals for the first postoperative year and at 6-month intervals thereafter. Metastatic reccurrences were identified by clinical examination, completed by chest X-ray and liver ultrasound (or CT scan). Ten among the 24 patients developped a liver metastasis in the follow-up, while the other 14 patients remained disease-free for at least 60 months.
For each patient, adjacent non-neoplastic colon mucosa (distance greater than 5 cm from the gross tumor limit) was collected at the time of surgery, with patients' informed consent, and was stored in liquid nitrogen within 0.5 hour after the resection. Samples were reviewed by a pathologist to check the absence of tumor cells.

Total RNA was extracted using Trizol reagent. mRNA target samples were hybridized to *Affymetrix HGU133A* GeneChips, containing a total of 22,283 probe-sets (Affymetrix, Santa Clara, CA), as described in the Affymetrix GeneChip Expression Analysis Manual (Affymetrix, Wooburn Green, UK). Briefly, 5 µg (100 ng/µl) of total RNA was used to synthesize double-stranded cDNA with SuperScript II reverse transcriptase (Invitrogen, Cergy Pontoise, France) and a T7-(dT)24 primer (Proligo Biochemie GmbH, Hamburg, Germany). Then, biotinylated cRNA was synthesized from the double-stranded cDNA using the RNA Transcript Labeling kit (Enzo Life Sciences, Farmingdale, NY) and was purified and fragmented. The fragmented cRNA was hybridized to the oligonucleotide microarray, which

was washed and stained with streptavidin-phycoerythrin. Scanning was performed with a GeneArray Scanner Update (Affymetrix, Wooburn Green, UK).

Data analysis

Data pre-processing

Starting from the 24 CEL files, gene expression measures were computed using the Robust Multichip Average (RMA) method described in Irizarry et al. [13] and implemented in the Bioconductor R package *affy*. This method includes the following successive steps : 1) Background correction ; 2) Probe-level quantile normalization ; 3) Calculation of expression measures using median polish.

Prognosis prediction

The prognosis prediction method consists of the following two steps.

a) Selection of informative genes. Genes that are differentially expressed between patients who experienced a tumor relapse and patients who remained disease-free are identified based on two-sample t-statistics with equal variance. The m genes with the largest absolute t-statistics are retained to build a prognosis predictor.

b) Prognosis prediction. The k-nearest neighbor method, based on the Euclidean distance between the expression profiles for the m informative genes of step a), is applied to predict prognosis. Specifically, the prognosis of a given patient is predicted as the most common prognosis among its k nearest neighbors, i.e., the k patients with the closest expression profiles.

Selection of prognosis predictor parameters

Leave-one-out cross-validation was used to select the two prognosis predictor parameters, namely the number of informative genes m and the number of nearest neighbors k. A total of 150 prognosis predictors were considered, corresponding to the following parameter values: k = 1, 3, and 5, and m = 5, 10, …, 250. The performance of a given prognosis predictor, indexed by the pair (m,k), was assessed as follows. Each of the 24 samples was used in turn as the *validation set*; the prognosis predictor was built using the *training set* formed by the remaining 23 samples and used to assign a prognosis (recurrence or no recurrence) to the validation sample; the predicted prognosis was then compared to the actual recurrence status; the numbers of false predictions (discordance between the predicted and actual evolutions) and true predictions were recorded for each of the 24 samples. Thus, for each of the 150 prognosis predictors, i.e., each (m,k) pair, a prediction error rate (out of 24) was obtained.

Proposition of a prognosis predictor

Because of ties in the error rates from leave-one-out cross-validation, the number of informative genes of this predictor was set to be the lowest number of informative genes, giving the lowest number of false predictions. Selection of informative genes was based on the 24 samples.

Estimation of the generalization error of the prognosis predictor

A double cross-validation scheme was used to assess the performance of the proposed prognosis predictor. For the "outer level" of cross-validation, the 24 samples were divided into 6 sets of 4 samples each (6-fold cross-validation). Each of these 6 sets was used in turn as the validation set, the other 5 sets (20 patients) being used as the training set. For each of the 6 steps in the cross-validation, a prognosis predictor was built based on the training set using

the method previously described: i) determination of the lowest number of genes and the lowest number of nearest neighbors giving the lowest number of false predictions (out of 20), using leave-one-out cross-validation ("inner level" of cross-validation); ii) selection of the m informative genes based on the 20 patients. The predictor was used to assign a prognosis to the 4 "outer level" validation set samples. The predicted prognoses were then compared to the actual recurrence status, giving a false prediction rate (out of 4). The 6 false prediction rates (one for each of the 6 steps of the outer level cross-validation) were averaged to provide an estimate of the generalization error.

Software

The statiscal analysis was performed with the open-source software R, Version 2.0.1. (http://cran.r-project/org), and Bioconductor packages (www.bioconductor.org). The following R packages were used : *affy* Version 1.5.8. (Irizarry RA, Gautier L, Bolstad BM, Miller C), *multtest* Version 1.5.2. (Pollard KS, Ge Y, Dudoit S), *class* Version 7.2.11. (Venables T, Ripley B, Hornik K, Gebhardt A), *hgu133a* Version 1.6.5. (Zhang J), and *annaffy* Version 1.0.11. (Smith CA).

Results

Selection of prognosis predictor parameters

A total of 150 prognosis predictors (50 possible values for the number m of informative genes, 3 possible values for the number k of nearest neighbors) were considered and their performance assessed using leave-one-out cross-validation. The distribution of the numbers of false predictions obtained with each of these 150 predictors is given in Figure 1. No pair of parameters (*m*, *k*) allowed a perfect concordance between the predicted and the observed evolutions. The numbers of false predictions ranged between 3 and 7. Three false predictions (out of 24, accuracy = 88%) represented the best and most frequent result (96 out of 150) . Figure 2 shows the numbers of false predictions obtained with respect to the values of both parameters, *m* and *k*. Predictors built with 30 or fewer informative genes yielded the highest numbers of false predictions (5 to 7). Predictors built with more than 60 informative genes yielded stable results and low numbers of false predictions. For a given number of informative genes, the results were quite similar for different numbers of nearest neighbors. The lowest number of informative genes giving the lowest number of false predictions (=3) was 65.

Proposition of a prognosis predictor

Based on the results of the leave-one-out cross-validation, 65 informative genes were selected using all 24 patients, by taking the 65 top-ranked genes (i.e., the 65 genes with the highest absolute t-statistics). Of these genes, 44 were over-expressed in patients who developed a recurrence while the other 21 were over-expressed in patients who remained disease-free for at least 5 years. Both lists of genes are given in Tables 1 and 2, respectively. Informative genes can be divided into 3 categories: 1) plasma membrane receptors with members of

different signaling pathways and transcription factors, 2) proteins involved in cell growth and/or maintenance such as glucose metabolism, protein biosynthesis, transport and degradation, and 3) proteins involved in the immune response. The following membrane receptors were over-expressed in the mucosa of patients who recurred: *solute carrier family 18*; *translocation protein 1*; *annexin 2*; *exostoses 2*; *ribophorin II*; *transmembrane protein 4*; two *G protein-coupled receptors* involved in positive regulation of I-kappaB kinase/NFkappaB cascade; *KDEL endoplasmic reticulum protein retention receptor 3* that can modulate MAP kinase signalling; *immediate early response 3 interacting protein 1*; and *integral membrane protein 2A*. Membrane receptors that were over-expressed in the mucosa of patients who remained recurrence-free belong to different families, except for transmembrane 4 superfamily member 2. There were : *CD24 antigen*, a protein involved in the humoral immune response that is also a membrane receptor over-expressed on epithelial cancer cells; *signal transducer and activator of transcription 2* that induces the JAK-STAT cascade; *SPPL2b*; *potassium voltage channel shaker-related family beta member 1*; *basigin*; and *major histocompatibility complex class I C*. As most of the cell surface receptors are linked to signal transduction, an over-expression of some signal transducers and factors of transcription was also observed : *WD40 protein ciao 1* that can interact with tumor suppressor proteins, and *ADP-ribosylation factor-like 1* in patients who recurred ; *cyclin-dependent kinase (CDC2-like) 10*, *ankyring repeat and SOCS-box containing 13* in patients who did not recur. Among genes involved in immunity, two transcripts, *CD24* and the *major histocompatibility complex class I, C* were overexpressed in the mucosa of patients who did not recur. Two members of the forkhead-box transcription factors, *forhead boxO1A* and *forkhead box J3*, were overexpressed in patients who recurred and in those who did not recur, respectively.

Estimation of the generalization error of the prognosis predictor

The results obtained at each of the 6 steps of the "outer level" cross-validation are summarized in Table 3. For each step, 20 samples were used as the training set, while the other 4 were used as the validation set. The second column indicates the distribution of the numbers of false predictions obtained with each of the 150 predictors in the "inner level" cross-validation based on the 20 patients of the training set. The third column gives the lowest numbers of informative genes and nearest neighbors that yielded the lowest number of false predictions for "inner level" leave-one-out cross-validation. These parameter values were used to build the prognosis predictor based on the training set of size 20. This predictor was applied to assign a prognosis to each of the 4 patients of the validation set. The false prediction rates, obtained for each of the 6 steps, are given in the fourth column. The average of these 6 false prediction rates (21%) provides an estimate of the accuracy of our proposed prognosis predictor (79%).

Discussion

The results of the present study clearly suggest the possibility to build a prognosis predictor based on non-neoplastic mucosa gene expression profiles for stage II colon cancer patients. To our knowledge, this is the first time that such a conclusion is reported. Wang et al. [9] have proposed an accurate prognosis predictor for stage II colon cancer patients, but based on tumor gene expression profiles. Non-neoplastic colonic mucosa profiles have already been studied, but only to compare them to tumor profiles [14-17].

Studies aiming to propose a predictor, for either diagnosis or prognosis purposes, are usually designed as follows: samples are split into a training set and a validation set; informative genes are selected based on the training set, using some arbitratory rule; the resulting predictor is assessed on the validation set. The design of the present study, which includes two distinct rounds of cross-validation with different aims, needs to be explained. The first part concerns the selection of a predictor using cross-validation, while the second aims to estimate the generalization error of the selected predictor. The k-nearest neighbor classifier was chosen because it has been shown to be competitive with more complex approaches, such as aggregated classification trees and support vector machines [18,19]. The main parameters of this classifier, namely the numbers of informative genes m and nearest neighbors k, were not chosen a-priori but using cross-validation in the first part of the study.

Specifically, 150 different pairs of parameters were considered and the performance of each was assessed using leave-one out cross-validation. Even this first part mainly aimed to select the predictor parameters, it also allowed to draw some informations about the stability of non-neoplastic mucosa-based prognosis predictors, i.e., the sensitivity of prediction error to the parameters (m, k). With a few informative genes (50 and less), predictor performance was

inversely proportional to the number of genes. With more informative genes, the prediction error rate seemed to stabilise.

Based on results of the first part, a 65-gene prognosis predictor was built on the whole set of patients. When proposing a predictor, it is important to provide an estimate of its accuracy. As a second set of independent samples was not available, a double cross-validation design was used with an "inner level" leave-one-out cross-validation, for parameter selection, and an "outer level" 6-fold cross-validation, for performance assessment of the selected predictor. In order to obtain an honnest estimate of generalization error, it is crucial that all aspects of predictor selection be included in the cross-validation process. Thus, for each of the 6 steps of the "outer level" cross-validation, we reproduced exactly what had been done in the first part of the study with an "inner level" cross-validation: i) selection of the parameters (m, k) yielding the best results by leave-one-out cross-validation, ii) use of this information to build a predictor based on the 20 patients. Note that the estimate of the generalization error, obtained by averaging the estimates of the "outer level" cross-validation, should be conservative, since it is computed based on sets of 20 patients (instead of 24). Thus, one may be confident that the accuracy of the proposed predictor is not over-estimated.

Wang et al. [9] reported a 78% accuracy in predicting the prognosis of stage II colon cancer patients with a predictor based on tumor gene expression profiles, while our predictor, based on non-neoplastic mucosa gene expression profiles, yielded a similar estimated accuracy (79%). The question of whether one should build a prognosis predictor based on tumor or non-neoplastic mucosa gene expression profiles immediately arises. In the present study, the paired tumor samples were not profiled since the aim was not to compare both predictors but to assess non-neoplastic mucosa-based predictors. However, in future studies, it would be of interest to compare the performances of both kinds of predictors. From a practical

point of view, the non-neoplastic mucosa represents an homogeneous pathological sample, while the tumor includes both tumoral and non-tumoral cells. The use of non-neoplastic may though avoid the need of laser-capture microdissection.

Despite the major difference in tissue material, the present study and that of Wang et al. [9] share an important conclusion: gene expression profiling is able to predict, with a great accuracy, the long-term postoperative outcome of stage II colon cancer patients. Thus, by identifying a subgroup of patients at high risk of recurrence, gene expression profiling may be used for postoperative therapeutical indications. To date, there is not enough evidence to claim that adjuvant chemotherapy should be given or not, based on gene expression profiles. But, initially, these profiles may be helpful for clinical studies assessing chemotherapy in stage II colon cancer patients: instead of including all these patients, these studies may be designed to include only patients identified as having a high risk of recurrence, thus more likely to benefit from adjuvant chemotherapy.

Interestingly, genes included in the proposed predictor are not cancer genes or genes encoding elements of the adhesion system, migration or proteolysis, but rather suggest branching signal transduction pathways with possible extensive networks between individual pathways and between cells themselves. For example, in the non-neoplastic mucosa of patients who recurred, we have observed an over-expression of two membrane receptors, *annexin 2,* and *transmembrane protein 4*, previously shown to be involved in tumor invasion [20-22]. Conversely, in the mucosa of patients who did not recur after a follow-up of 5 years, we have observed an increased expression of some genes already reported to induce tumor cell invasion: *basigin*, known to stimulate production of matrix metalloproteinases by fibroblasts [23,24]; a member of *transmembrane protein 4* [21,22]; and *CD24* [25]. CD24 is a mucin-like cell surface molecule on human neutrophils, pre-lymphocytes, and many epithelial

tumors. Its over-expression in some epithelial tumors, frequently associated with a high tumor grade, has suggested its prognostic value as a routine marker [25]. However, an absence of expression of CD24 mRNA has also been observed in invasive mammary carcinoma derived cells compared to non invasive cells [26]. The role of CD24 in the dissemination of tumor cells could be due to different pathophysiological processes according to its cellular localization. Indeed, it has been described to facilitate the interactions with P-selectin in platelets or endothelial cells [27,28] and to regulate T-cell proliferation in lymphopenic host [29]. Recently, the role of CD24 signaling in the mitochondrial regulation of apoptosis has also been shown [30].

The over-expression of CD24 in the mucosa of patients who remained disease-free was associated with over-expressions of the *major histocompatibility complex* and a member of the *forkhead box*, a family of transcription factors recently shown to play a crucial role in the immune system [31]. Emerging evidences suggest that epigenetic events associated with tumor development and progression, such as deregulated methylation of CpG dinucleotides and aberrant histone acetylation, may impair the immunogenic potential of cancer cells. A central question in cancer immunology remains how the additional genetic alterations, both in primary tumor and in the stromal cells, and the inherent proinflammatory processes can activate tumoral immunity and thus induce immune tolerance. Thus, although the role of the abnormal expression of these genes was not clearly defined as pro- or anti-invasive, our results emphasize that the immune response, to promote the survival or the death of malignant cells, is not restricted to immune cells that infiltrate tumors. The recruitment of cells distant from the primary tumor could constitute a possible mechanism for the presence of lymph-node metastases in some solid tumors such as colorectal cancers.

The over-expression of genes coding for membrane receptors coupled with signal transduction, such as G protein-coupled-receptor and protein kinases, transcription factors and

members of cellular proteolysis systems, such as the Cop9 signalosome and 26S proteasome, suggests an important cross talk between cells, probably connecting the initial events, *e.g.* activation of receptors, to the activation of gene expression in the nucleus. The activation of signalling pathways has been already shown to play a central function in invasion–related cellular activities determining the cells response to extrinsic or intrinsic modulators [32].

In conclusion, the present study clearly suggests the possibility to build a prognosis predictor, based on non-neoplastic mucosa gene expression profiles for stage II colon cancer patients. It also raises questions regarding the role of the so-called "normal mucosa" surrounding the tumor. Genomic alterations in epithelial cells which lead to primary tumors may disturb the molecular cross-talk between cancer cells and the underlying stroma. This conversation may be relayed by other host cells distant from the primary tumors, these cells presenting normal phenotype and thus allowing an adaptated signalling. Several questions remain to be elucidated. One of these is to determine whether normal cells distant from the tumors are contacted to stop and repair or to help the cancer cell invasion.

References

1. Figueredo A, Charette ML, Maroun J, Brouwers MC, Zuraw L. Adjuvant therapy for stage II colon cancer. A systematic review from the Cancer Care Ontario Program in evidence-bades care's gastrointestinal cancer disease site group. J Clin Oncol 2004;22:3395-407.

2. Benson AB 3rd, Schrag D, Somerfield MR, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. J Clin Oncol 2004;22:3408-19.

3. Van't Veer LJ, Dai H, Van De Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.

4. Van de Vijver MJ, Yudong DH, 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.

5. Beer DG, Kardia SLR, Huang C, et al. Gene expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816-24.

6. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung caecinoma by mRNA expression profiling reveals distinct acenocarcinoma subclasses. Proc Natl Acad Sci 2001;98:13790-5.

7. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8:68-74.

8. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. N Engl J Med 2002;346:1937-47.

9. Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 2004;27:1564-71.

10. Mueller MM, Fusenig NE. Friends or foes - bipolar effects of the tumour stroma in cancer. Nat Rev Cancer 2004;4:839-49.

11. Bernards R, Weinberg RA. A progression puzzle. Nature 2002;418:823-4.

12. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33:49-54.

13. Irizarry RA, Gautier L, Cope L. An R Package for analyses of Affymetrix oligonucleotide arrays. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, editors. The analysis of gene expression data: methods and software. New York: Springer; 2003. p.102-119.

14. Zou TT, Selaru FM, Xu Y, et al. Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. Oncogene 2002;21:4855-62.

15. Kitahara O, Furukawa Y, Tanaka T, et al. Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Res 2001;61:3544-9.

16. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma and normal tissue examined by oligonucleotide arrays. Cancer Res 2001;61:3124-30.

17. 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 1999;96:6745-50.

18. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science 2003;18:71-103.

19. Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed T, editor. Statistical analysis of gene expression microarray data. Boca Raton: Chapman & Hall; 2003. p.93-158.

20. Tanaka T, Akatsuka S, Ozeki M, Shirase T, Hiai H, Toyokuni S. Redox regulation of annexin 2 and its implications for oxidative stress-induced renal carcinogenesis and metastasis. Oncogene 2004;23:3980-9.

21. Hashida H, Takabayashi A, Tokuhara T, et al. Clinical significance of transmembrane 4 superfamily in colon cancer. Br J Cancer 2003;89:158-67.

22. Mazzocca A, Carloni V, Sciammetta S, et al. Expression of transmembrane 4 superfamily (TM4SF) proteins and their role in hepatic stellate cell motility and wound healing migration. J Hepatol 2002;37:322-30.

23. Kanekura T, Chen X, Kanzaki T. Basigin (CD147) is expressed on melanoma cells and induces tumor cell invasion by stimulating production of matrix metalloproteinases by fibroblasts. Int J Cancer 2002;99:520-8.

24. Li R, Huang L, Guo H, Toole BP. Basigin (murine EMMPRIN) stimulates matrix metalloproteinase production by fibroblasts. J Cell Physiol 2001;186:371-9.

25. Kristiansen G, Sammar M, Altevogt P. Tumour biological aspects of CD24, a mucin-like adhesion molecule. J Mol Histol 2004;35:255-62.

26. Schindelmann S, Windish J, Grundmann R, Kreienberg R, Zeillinger R, Deissler H. Expression profiling of mammary carcinoma cell lines: correlation of in vitro invasiveness with expression of CD24. Tumour Biol 2002;23:139-45.

27. Aigner S, Ramos CL, Hafezi-Moghadam A, et al. CD24 mediates rolling of breast carcinoma cells on P-selectin. FASEB J 1998;12:1241-51.

28. Aigner S, Sthoeger ZM, Fogel M, et al. CD24, a mucin-type glycoprotein, is a ligand for P-selectin on human tumor cells. Blood 1997;89:3385-95.

29. Li O, Zheng P, Liu Y. CD24 expression on T cells is required for optimal T cell proliferation in lymphopenic host. J Exp Med 2004;200:1083-9.

30. Taguchi T, Kiyokawa N, Mimori K, et al. Pre-B cell antigen receptor-mediated signal inhibits CD24-induced apoptosis in human pre-B cells. J Immunol 2003;170:252-60.

31. Coffer PJ, Burgering BM. Forkhead-box transcription factors and their role in the immune system. Nat Rev Immunol 2004;4:889-99.

32. Barbier M, Attoub S, Calvez R, et al. Tumour biology. Weakening link to colorectal cancer? Nature 2001;413:796.

Table 1. Over-expressed genes in patients who developped a recurrence

Affy probeID	Gene Name	GenBank Accession Number
207074_s_at	Solute carrier family 18 (vesicular monoamine), member 1	NM_003053
213800_at	complement factor H	X04697
208942_s_at	translocation protein 1	BE866511
202141_s_at	COP9 constitutive photomorphogenic homolog subunit 8 (Arabidopsis)	BC003090
206884_s_at	Sciellin	NM_003843
201606_s_at	nuclear phosphoprotein similar to S. cerevisiae PWP1	BE796924
213503_x_at	annexin A2	BE908217
203536_s_at	WD40 protein Ciao1	NM_004804
202341_s_at	tripartite motif-containing 2	AA149745
211023_at	pyruvate dehydrogenase (lipoamide) beta	AL117618
202013_s_at	exostoses (multiple) 2	NM_000401
210427_x_at	annexin A2	BC001388
212836_at	polymerase (DNA-directed), delta 3, accessory subunit	D26018
208093_s_at	nudE nuclear distribution gene E homolog like 1 (A. nidulans)	NM_030808
201067_at	proteasome (prosome, macropain) 26S subunit, ATPase, 2	BF215487
201590_x_at	annexin A2	NM_004039
213399_x_at	ribophorin II	AI560720
218976_at	DnaJ (Hsp40) homolog, subfamily C, member 12	NM_021800
201543_s_at		NM_020150
202857_at	transmembrane protein 4	NM_014255
202723_s_at	forkhead box O1A (rhabdomyosarcoma)	AW117498
209045_at	X-prolyl aminopeptidase (aminopeptidase P) 1, soluble	AF195530
220841_s_at	Abelson helper integration site	NM_017651
218135_at	PTX1 protein	NM_016570
214307_at	homogentisate 1,2-dioxygenase (homogentisate oxidase)	AI478172
207651_at	G protein-coupled receptor 171	NM_013308
204017_at	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3	NM_006855
211406_at	immediate early response 3 interacting protein 1	AF119875
218257_s_at	UDP-glucose ceramide glucosyltransferase-like 1	NM_020120
219553_at	non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)	NM_013330
212342_at	hypothetical protein MGC21416	BG500611
222122_s_at	THO complex 2	BG403671
221766_s_at	family with sequence similarity 46, member A	AW246673
201658_at	ADP-ribosylation factor-like 1	AU151560
217868_s_at	DORA reverse strand protein 1	NM_016025
201077_s_at	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)	AF155235
205141_at	angiogenin, ribonuclease, RNase A family, 5	NM_001145
205342_s_at	sulfotransferase family, cytosolic, 1C, member 1	AF026303
216228_s_at	WD repeat and HMG-box DNA binding protein 1	AK001538
222140_s_at	G protein-coupled receptor 89	AK021758
208095_s_at	signal recognition particle 72kDa	NM_001222
213491_x_at	ribophorin II	AL514285
202747_s_at	integral membrane protein 2A	NM_004867
201822_at	translocase of inner mitochondrial membrane 17 homolog A (yeast)	NM_006335

Table 2. Over-expressed genes in patients who remained disease-free

Affy probeID	Gene Name	GenBank Accession Number
209771_x_at	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	AA761181
205170_at	signal transducer and activator of transcription 2, 113kDa	NM_005419
215833_s_at	SPPL2b	AC004410
216379_x_at	KIAA1919	AK000168
210622_x_at	cyclin-dependent kinase (CDC2-like) 10	AF153430
208651_x_at	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	M58664
266_s_at	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	L33930
207980_s_at	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	NM_006079
202242_at	transmembrane 4 superfamily member 2	NM_004615
200661_at	protective protein for beta-galactosidase (galactosialidosis)	NM_000308
213827_at	sorting nexin 26	AL137579
208156_x_at	epiplakin 1	NM_031308
210079_x_at	potassium voltage-gated channel, shaker-related subfamily, beta member 1	U16953
211065_x_at	phosphofructokinase, liver	BC006422
209357_at	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	AF109161
216103_at	thioesterase, adipose associated	AB014607
218862_at	ankyrin repeat and SOCS box-containing 13	NM_024701
208677_s_at	basigin (OK blood group)	AL550657
211799_x_at	major histocompatibility complex, class I, C	U62824
200646_s_at	nucleobindin 1	NM_006184
217310_s_at	forkhead box J3	AK027075

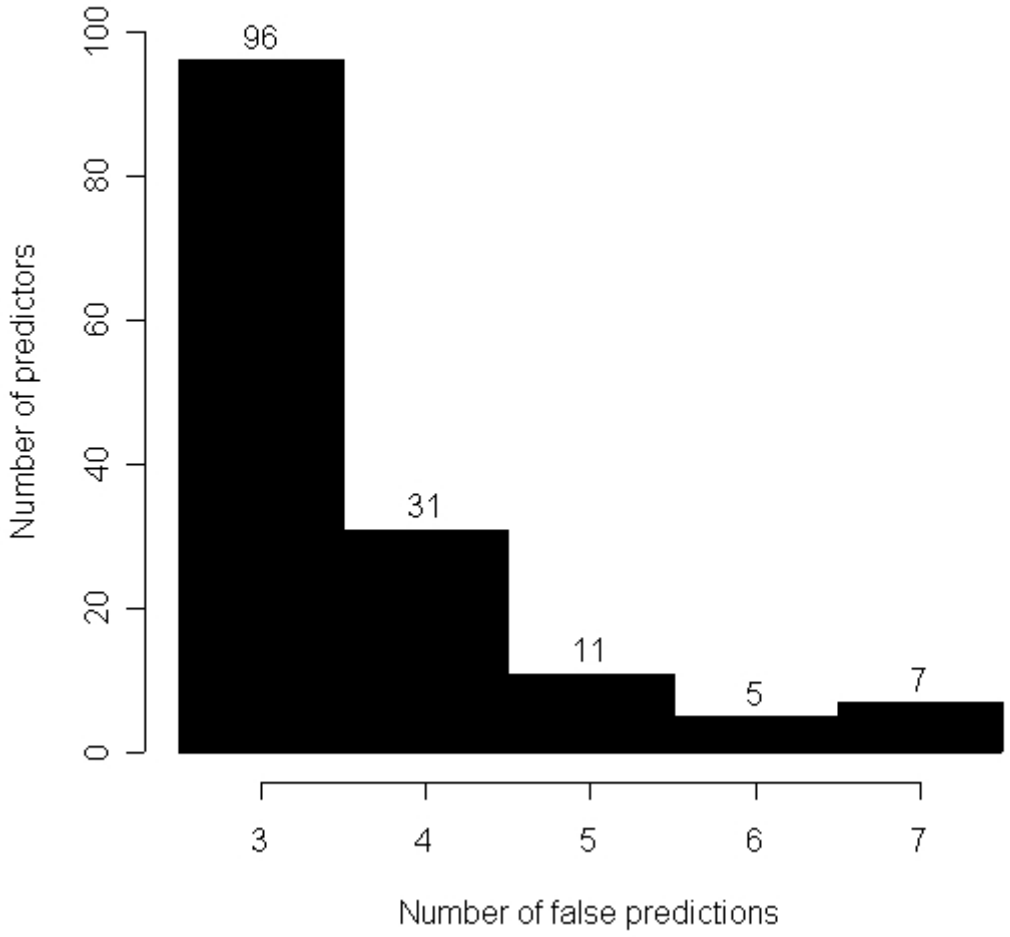
Table 3. Results of the double cross-validation.

	Leave-one-out cross-validation Number of predictors with							Parameters (m, k)	FP Rate (validation set)
	1 FP *	2 FP *	3 FP *	4 FP *	5 FP *	6 FP *	7 FP *		
Step 1		119	18	9	3	1		m=15 , k=3	2/4 (50%)
Step 2	1	6	62	70	10	1		m=10 , k=1	1/4 (25%)
Step 3		13	20	79	28	1	9	m=70 , k=1	0/4 (0%)
Step 4		9	34	67	35	5		m=105 , k=1	1/4 (25%)
Step 5		1	8	36	27	30	48	m=20 , k=3	0/4 (0%)
Step 6		50	47	29	17	7		m=40 , k=1	1/4 (25%)

FP : False Prediction

* : out of 20 samples

Figure 1. Distribution of the numbers of false predictions.



A total of 150 prognosis predictors - 50 possible values for the number m of informative genes (m = 5, 10, …, 250), 3 possible values for the number k of nearest neighbors (k = 1, 3, and 5) - were considered and their performance assessed using leave-one-out cross-validation. Figure 1 shows the distribution of the numbers of false predictions (out of 24) obtained with each of these 150 predictors.

Figure 2. Number of false predictions.

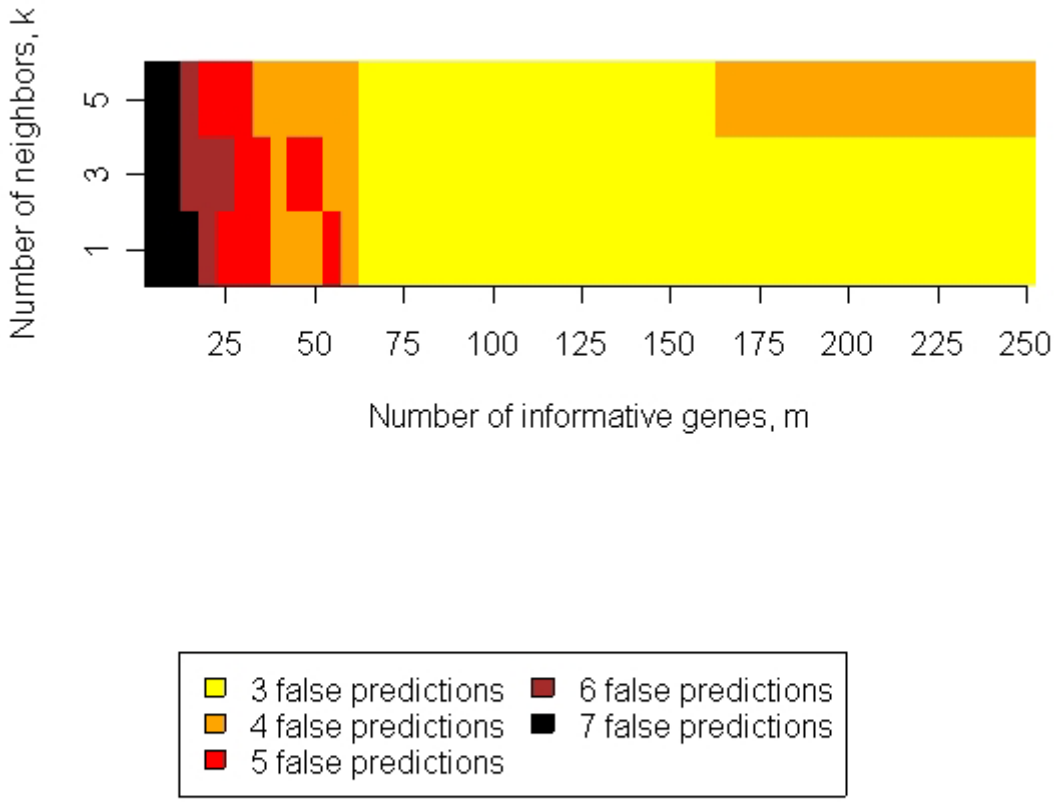


Figure 2 shows the number of false predictions as a function of the number *m* of informative genes (x-axis) and the number *k* of nearest neighbors (y-axis). In these pseudo-color images, colored rectangles indicate the number of false predictions, with yellow (black) corresponding to the lowest (highest) numbers of errors.