

# Predicting node positivity in gastric cancer from gene expression profiles

Michael J. Korenberg · Bryan J. Dicken ·  
Sambasivarao Damaraju · Kathryn Graham ·  
Carol E. Cass

Received: 30 April 2009 / Accepted: 7 May 2009  
© Springer Science+Business Media B.V. 2009

**Abstract** Lymphovascular invasion (LVI) in gastric cancer is readily demonstrated pre-operatively by mucosal biopsy during endoscopy, which can also provide samples for gene expression profiling. We have examined gene expression associated with lymphovascular invasion in a cohort of gastric cancer patients and have developed a 28-gene predictor of tumor aggressiveness for forecasting risk of node positivity (N+), which could potentially be useful pre-operatively. The resulting model's ranking of increasing tumor aggressiveness correlated positively with N+ status, reaching statistical significance, and was three times the correlation of LVI with N+ status.

**Keywords** Gastric cancer · Lymphovascular invasion · Microarray · Tumor-node-metastasis

## Introduction

Gastric cancer is one of the main causes of cancer-related deaths worldwide, causing about 14,000 deaths annually in the United States (Dicken et al. 2005; Karpheh and Brennan 1998). Adjuvant and neoadjuvant therapy and decision-making on surgical margins in gastric and other cancers would benefit if it were possible to forecast risk of nodal involvement soon after diagnosis. Previously, microarray analysis of chromosomal *copy number changes*, and amplifications, gains, and losses at a number of chromosomal regions, revealed various groups that correlated significantly with lymph node status and survival in gastric cancer (Weiss et al. 2003, 2004). Lymphovascular invasion (LVI) can be readily demonstrated pre-operatively by mucosal biopsy during endoscopy and has been shown to correlate with advancing nodal status in a study involving several hundred patients (Dicken et al. 2004). Prognostic value of LVI has been shown for patients with adenocarcinomas of the esophagogastric junction (von Rahden et al. 2005).

Though LVI status does provide a measure of tumor aggressiveness, not all cases with lymphovascular invasion show nodal involvement (N+), and not all LVI- cases are node negative (N0), so it is useful

---

M. J. Korenberg (✉)  
Department of Electrical and Computer Engineering,  
Queen's University, Kingston, Ontario K7L 3N6, Canada  
e-mail: korenber@queensu.ca

B. J. Dicken · S. Damaraju · K. Graham · C. E. Cass  
PolyomX Program, Cross Cancer Institute, 11560  
University Ave, Edmonton, Alberta T6G 1Z2, Canada

B. J. Dicken  
Department of Surgery, University of Alberta, Edmonton,  
Alberta, Canada

S. Damaraju  
Department of Laboratory Medicine and Pathology,  
University of Alberta, Edmonton, Alberta, Canada

K. Graham · C. E. Cass  
Department of Oncology, University of Alberta,  
Edmonton, Alberta, Canada

to develop a refined model of tumor aggressiveness. A recent paper examining gene expression of fresh gastric cancer tissue found differential expression of oligophrenin-1 (OPHN1) and ribophorin-II (RPNII) with respect to LVI (Dicken et al. 2006). In a prospective microarray study, gene expression profiles were obtained from tumors of 20 patients undergoing surgery for gastric cancer (Dicken et al. 2006); 15 tumors had lymphovascular invasion (were LVI+), five were LVI-.

The present paper is an extension in which a data-mining tool was applied to the same gene expression profiles to produce a clinically-relevant predictor in a proof of principle study. In particular, starting with known LVI status, a gene-expression-based model of tumor aggressiveness was developed for prediction of nodal involvement, in which the model was itself created blinded to any nodal information about the test cases. This blind testing of the predicted risk of nodal involvement distinguishes the present work from many other reports of gene-expression-based predictors. One important exception is the classic paper of Khan et al whose artificial neural networks classified flawlessly blinded test gene expression profiles of small round blue-cell tumors (Khan et al. 2001).

## Patients and methods

### Patient samples

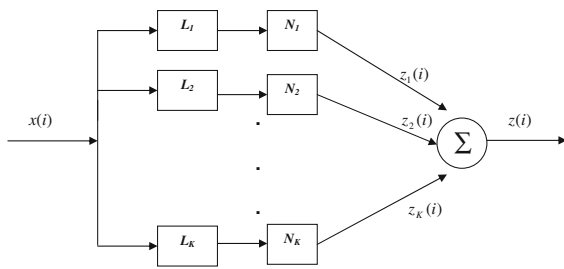
Microarray slides were printed by the Gene Array Facility of Genome British Columbia using the Operon (Alameda, CA) Human 70 mer oligonucleotide set, version 1.1, representing 13,971 genes, which were printed in duplicate on each slide. RNA was obtained from tumors of 20 patients (15 LVI+, 5 LVI-) undergoing surgery for primary gastric cancer, who prospectively provided written informed consent for tissue banking and analysis plans that were approved by the appropriate Research Ethics Board(s). Microarray slides were probed in triplicate with labeled cDNA prepared from 30 µg each of tumor and reference total RNA. Superscript II was used to prepare cDNA, which was then labeled with Cy3 or Cy5 using an indirect amino allyl technique (Botwell and Sambrook 2003). To ensure all relevant gastric-expressed RNAs would be represented, the reference

sample was prepared by mixing RNA from normal gastric mucosa with RNA from ten representative gastric tumor samples. After hybridization, the microarray slides were scanned with an Axon 4000B using GenePix 3.0 software and the data were normalized using lowess normalization as previously described (Listgarten et al. 2003). Further details about the microarray profiles are available (Dicken et al. 2006).

LVI was defined as presence of tumor emboli in either vascular or lymphatic channels (Dicken et al. 2004). In the present study, the LVI status was based upon final pathologic reporting and not obtained preoperatively. TMN staging nodal status was defined by the 1997 American Joint Committee on Cancer staging system as follows. N0 denotes no involved nodes as determined after examination of at least 15 lymph nodes, N1 denotes 1–6 positive nodes, N2 denotes 7–15 positive nodes, and N3 denotes >15 positive nodes (Karpeh et al. 2000). In our study, six tumor samples were N0, 9 were N1, and five were N2, and none were N3. No positive node was sufficiently large to be visualized by CT or endoscopic ultrasound; nodal status was completely unavailable preoperatively.

### Statistical methods

The parallel cascade identification (PCI) approach detailed previously (Korenberg 1991) builds a mimetic model of a nonlinear system, using a parallel array of dynamic linear and static nonlinear elements (Fig. 1), given only the system's input and output. Constructing a PCI model to distinguish between gene expression profiles belonging to different classes was described (Korenberg 2002), briefly reviewed by Kirkpatrick (2002), and is outlined here. First, genes having largest absolute difference in expression level between the training profiles for the two classes are selected. Next, for each exemplar profile, the expression values from the selected genes are appended in the same order to form an input segment representative of the exemplar's class. The input segments from the different classes are concatenated to form a training input  $x(i)$ . The corresponding training output  $y(i)$  is assigned different values over input segments from different classes. For this training input and output, a PCI model (Fig. 1) is identified to approximate the input/output relation (Korenberg 1991). Then an input signal for a novel profile is constructed by appending the expression values of the selected genes in the same order used



**Fig. 1** Parallel cascade model used to predict LVI. Each  $L$  is a dynamic (has memory) linear element and each  $N$  is a polynomial static nonlinearity

above; the resulting model output is used to classify the profile.

To build a PCI predictor of LVI status, only one LVI+ and one LVI− profile were used to construct the training input, so that the test set comprised the remaining 14 LVI+ and 4 LVI− profiles. When selected genes from the first LVI+ (denoted GT4) and first LVI− (denoted GT86) profiles were used to build the training input, the LVI+ segment of the input was significantly smaller at almost every expression value than the LVI− segment. This indicates a colored input, while a white one is advantageous for PCI (Korenberg 2002). Hence one other pair of exemplars, the last LVI+ (UT90) and last LVI− (UT178) profiles, were checked for forming the training input, and a much richer training input resulted, so that these profiles were chosen to find the PCI model. Indeed, the autocovariance curve, for a training input based on GT4 and GT86, indicated a colored input, while that for training profiles UT90, UT178 indicated an almost white input. The training profiles were chosen blinded to nodal involvement and the risk predictor built before any nodal information was revealed. No information could leak to increase classification performance.

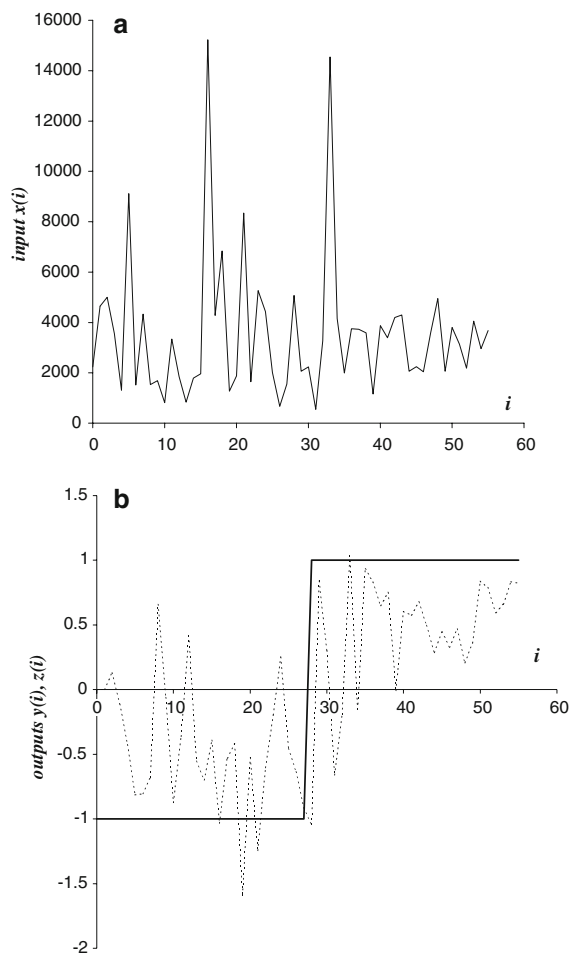
Some parameters chiefly related to architecture had to be set (Korenberg 2002): *memory length* of a dynamic linear element  $L$ , *polynomial degree* of a static nonlinear element  $N$ , *maximum number* of cascade paths, *threshold* for accepting a candidate cascade into the model, and *number* of genes to select. These parameters' values were set by comparing the performance of PCI models, identified for trial values of the parameters, over an evaluation set that never included the profile to be classified.

Since there were relatively few profiles in total, the protocol used to develop and test the PCI classifiers was a variant of leave-one-out testing. Three sets were used: (1) UT90, UT178 to construct training inputs; (2) 17 profiles in an evaluation set; (3) a held-out profile to be classified. Several PCI models were trained once and for all from set (1) for trial values of number (200, 26–30, and 22) of genes, and the various architectural values listed above. The trial architectural parameter settings and number of genes covered a narrow range around those employed earlier (Korenberg 2002, 2003), where an effective two-cascade model with memory length 4 and polynomial degree 5 was identified using threshold 6. The performance of the different PCI models was compared; the best one, with highest percentage correct over LVI+ and LVI− classes, was selected. In the event of a tie, the PCI model having minimum MSE over the training data (in set 1) was selected, because PCI models with lower MSE over the same training data tend to be better classifiers.

Sets (2) and (3) were changed until each of the 18 profiles (all but UT90, UT178) had been classified for LVI. Since the training profiles never change, it is possible that the same model may turn out to be the best over the evaluation set each time, in which case the same model could be shown to be effective over the 18 test profiles. Here, the same PCI model using 28 gene expression values emerged as best over every evaluation set. The corresponding training input for this model is shown in Fig. 2a. Model memory length was three, with 7th degree polynomial static nonlinearities and two cascade paths in total, identified using a threshold of six—settings similar or identical to values used previously (Korenberg 2003). On the training data, model MSE was 46.56%, relative to the training output variance.

Table 1 shows that, of the selected genes, eukaryotic translation elongation factor alpha 1 (EEF1A1) was chosen twice. This is because the gene was represented on the array by two different oligonucleotide segments since, when the array was designed, the segments were believed to correspond to different genes. Strictly speaking, the model used 28 oligonucleotides corresponding to 27 genes; however for simplicity, it will be referred to as a 28-gene model.

When the training input from Fig. 2a was fed through the model, the resulting output (Fig. 2b) was predominately negative over the LVI+ portion, and



**Fig. 2** **a** Training input  $x(i)$  formed by splicing together the raw expression levels of selected genes from LVI+ profile UT90 and LVI- profile UT178. The genes used (Table 1) were those having greatest difference in expression levels between the two profiles based on 28 different oligonucleotides. **b** Training output  $y(i)$  (solid line) defined as  $-1$  over the LVI+ portion of the training input and  $1$  over the LVI- portion. The training input and output were used to identify a parallel cascade model of the form in Fig. 1. The dashed line represents calculated output  $z(i)$  when the identified model is stimulated by training input  $x(i)$ . Note that  $z(i)$  is predominately negative (average value:  $-0.4799$ ) over the LVI+ portion, and positive (average value:  $0.4799$ ) over the LVI- portion, of the training input

positive over the LVI- portion, of the training input. To predict LVI status of a novel profile, a 28-point input signal prepared from the expression values of the selected genes was fed through the model. The resulting model output was averaged following the

**Table 1** Genes used to predict lymphovascular invasion of gastric cancer

Gene description	Gene symbol	Accession number
Collagen type III, alpha 1	COL3A1	NM_000090
Persephin	PSPN	AF040962*
Oligophrenin 1	OPHN1	NM_002547*
Lysozyme (renal amyloidosis)	LYZ	NM_000329*
Neuromedin U	NMU	NM_019515
Immunoglobulin kappa variable	IGCK	NM-00834
Tripartate motif containing 3	TRIM3	NM_033278
Ribosomal protein L18a	RPL18A	X80822*
Unknown	N/A	AK001590
Ribosomal protein S24	MRPS24	NM_032014
Eukaryotic translation elongation factor alpha 1	EEF1A1	NM_001402
H3 histone family 3A	H3F3A	M11353*
Adenosine deaminase tRNA-specific 1	ADAT1	NM_012091
Actin, alpha 2, smooth muscle, aorta	ACTA2	NM_001613
Zinc finger protein 160	ZNF160	NM_198893
Telomerase-associated protein 1	TEP1	NM_007110
CDC14 cell division cycle 14 homolog A	CDC14A	NM_033312***
Ribosomal protein L12 pseudogene	N/A	U85977*
Myosin-binding protein	MYBP3	Y10129*
Eukaryotic translation elongation factor alpha 1	EEF1A1	NM_001402
Proline dehydrogenase	PRODH2	AK001359
Matrix Gla protein	MGP	AK000309**
PRO1900 protein	N/A	NM_016344
Unknown	N/A	NM_005579*
Ribosomal protein S14	MRPS14	NM_022100*
Unknown	N/A	AK001329
Ribosomal protein large P0	RPLP0	NM_053275
Hydroxysteroid (17-beta) dehydrogenase 1	HSD17B1	NM_000413

Asterisks denote various degrees of up-regulation in LVI+ based on training profiles UT90, UT178, with CDC14 (cell division cycle 14 homolog A) especially active. Note that the EEF1A1 gene, corresponding to two different oligonucleotides (originally designated as unigene nos. 181165 and 274466) on the microarray, was selected twice with different expression values. The 28 gene-expression values were appended using the above order in constructing the input signal. Gene symbols that are not available in public databases are indicated (N/A)

settling period, which depended on the memory length: here, for memory length of 3, the averaging started on the 3rd output point. If average model output was negative the tumor sample was classified as LVI+, and otherwise as LVI-.

## Results

### LVI prediction

Thirteen of 14 LVI+ (~93%) and 3 of 4 LVI- (75%) were correctly classified ( $P < 0.019$  on Fisher's exact test, Matthews' correlation coefficient  $\phi = 0.68$ ); the only errors were (LVI+) GT4 and (LVI-) GT86 profiles.

### Predicting risk of nodal involvement

When nodal status was finally unmasked, it emerged that six tumor samples were N0, while 14 samples were N+. LVI has been shown to correlate with advancing nodal status (Dicken et al. 2004, 2006) when a large number of tumor samples were studied. However, over our 20 tumor samples, the correlation of LVI with node positivity was only 0.13 and did not reach statistical significance.

Recall that the PCI mean outputs, with negative values indicating tumor aggressiveness, had been calculated for all 20 profiles blinded to nodal involvement; their relative ranking is shown in Table 2a together with actual N0 versus N+ status. They are listed in descending order with the profile having the largest positive mean output at the top. While PCI mean output is one way of ranking the profiles from least aggressive to more aggressive cases, there is no claim that the mean outputs form a linear (equal interval) scale. Because of the way the training output had been defined (-1 for LVI+, +1 for LVI-), progressively negative mean output, corresponding to increasing rank number in Table 2a, should indicate increasing tumor aggressiveness and increased risk of node positivity. Thus, N0 profiles should tend to be nearer to the top of Table 2a. Moreover, increasing rank number should correlate positively with N+ status; a negative correlation, no matter how large the magnitude, would not be significant and instead would indicate that the predictor had failed. This means that the relevant

issue is the probability of the same or higher positive correlation occurring by chance. Indeed, increasing rank number correlated with node positivity ( $r = +0.4$ ), with probability  $P < 0.0412$  of obtaining as high or higher positive correlation by chance.

The exact Mann-Whitney U-test (Lowry 2007) confirms that PCI ranked the profiles according to increasing risk of nodal involvement, with least aggressive cases at the top, in Table 2a. Here the statistic U measured the number of times an N+ profile had a smaller rank number than an N0 profile in Table 2a; for six N0 and 14 N+ the critical value is  $U = 21$  at or below which a ranking would be significant. If the PCI model could not predict increased risk of node positivity, then the relative ordering of N0 and N+ profiles should be random in Table 2a, with expected value of  $U = 42$ . However as predicted, N0 cases clearly tend to occur closer to the top in Table 2a, and the ranking reached the critical value  $U = 21$  for significance on the exact Mann-Whitney test ( $P < 0.0457$ ).

The PCI ranking is certainly not perfect, but its correlation with node positivity was three times the correlation of LVI with node positivity. One practical way of using the PCI ranking is by checking whether a profile would rank within the top or the bottom half of Table 2a, of predicted lower-risk and higher-risk cases respectively. This two-class form effectively distinguishes among different stages of nodal status (here, N0, N1, or N2). All our N+ cases were N1, except N2 cases GT4, RT8, RT233, UT90, RT29. Comparing the top half of Table 2a predicted to be at lower risk with the bottom half predicted at higher risk showed that higher risk correlated positively with increased nodal status:  $r = +0.47$ , with probability  $P < 0.018$  of obtaining as high or higher positive correlation by chance. However, this median split dichotomizes a quantitative variable and, though it is not an uncommon practice, typically it is not recommended since it discards useful information (MacCallum et al. 2002). It is presented as one way to make practical use of the PCI ranking but, to establish that the ranking is predictive of node positivity, reliance is instead placed on other analysis such as the exact Mann-Whitney U-test discussed above. Although PCI predicted risk correlated with actual nodal status, the biggest errors were cases GT4, predicted lower-risk yet actually N2, and MT385, predicted higher-risk yet actually N0. All

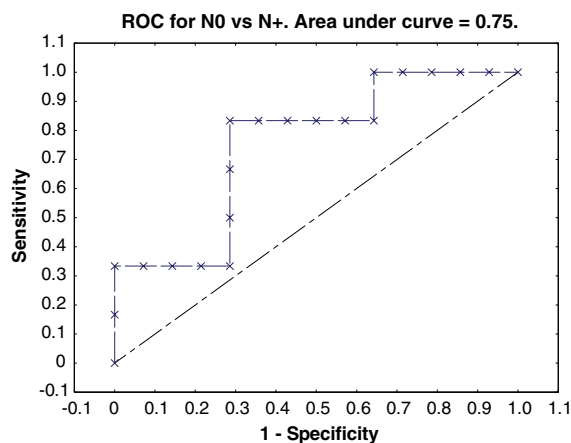
**Table 2** PCI Ranking correlated with N0 versus N+ status (Node Positivity)

	a			b	
	Rank	Node positivity	Profile number	Node positivity	Profile number
	1	N0	UT178	N0	RT191
	2	N0	RT191	N0	UT178
	3	N+	RT258	N+	RT258
	4	N+	RT136	N0	UT260
	5	N+	GT4	N+	RT187
	6	N+	GT64	N0	UT172
	7	N0	UT260	N+	RT8
	8	N0	UT172	N+	RT136
	9	N0	RT46	N+	GT4
	10	N+	RT187	N0	RT46
	11	N+	RT8	N+	GT64
	12	N+	RT233	N+	RT60
	13	N+	UT90	N+	RT233
	14	N+	RT60	N+	UT90
	15	N0	MT385	N+	RT239
	16	N+	RT239	N+	UT278
	17	N+	UT278	N0	MT385
	18	N+	RT29	N+	RT29
	19	N+	GT86	N+	RT182
	20	N+	RT182	N+	GT86

Increasing rank number tends to predict increasing tumor aggressiveness. The exact Mann–Whitney U-test showed that PCI tended to assign smaller rank numbers to N0 (node negative) profiles than to N+ (node positive) profiles. Point biserial correlation of increasing rank number with N+ status produced significant positive coefficient  $r$ . In A, the PCI values for ranking the profiles were obtained blinded to nodal involvement:  $U = 21$ ,  $P < 0.0457$  (Mann–Whitney);  $r = +0.4$ ,  $P < 0.0412$  (point biserial). In B, the analysis was repeated starting with different training profiles, and was carried out after nodal status had been unmasked:  $U = 19$ ,  $P < 0.0314$ ;  $r = +0.44$ ,  $P < 0.028$

other N0 cases fell in the predicted lower-risk half, all other N2 cases in the predicted higher-risk half. These results could be further improved by using LVI information, see below.

Figure 3 shows the receiver operating characteristic (ROC) curve for plot of true positive fraction (*sensitivity*) against false positive fraction (equal to  $1 - \textit{specificity}$ ), for each possible cut-off value. The curve was plotted for predicting node negativity: *sensitivity* refers to the fraction of N0 cases that would be correctly classified for each possible cut-off value, and  $1 - \textit{specificity}$  is the fraction of N+ cases that would be misclassified using the same cut-off. For the 20 profiles, there are 19 intermediate cutoffs corresponding to 19 measured values that exactly trace out the ROC curve shown [in addition to points (0, 0) and (1, 1)]. On the resulting ROC curve, the point corresponding to *sensitivity* of 0.83 has *specificity* of 0.71 ( $1 - \textit{specificity} = 0.29$ ). Area under the curve (AUC) reflects the predictive power, with  $0 < \text{AUC} < 1$ . The null hypothesis of no predictive power corresponds to an  $\text{AUC} = 0.5$ , the closer to



**Fig. 3** Receiver operating characteristic (ROC) curve for N0 versus N+, X denotes a measured value. Area under the curve (AUC) is 0.75, and the standard error is 0.13

1 the better the predictor. Here calculated  $\text{AUC} = 0.75$ , with probability that the AUC is actually not higher than 0.5 only  $P < 0.0275$ , indicating that the PCI ranking has predictive power. The obtained AUC is very similar to that of Wang et al. (2004), where

the AUC equaled 0.741 for their predictor of relapse in Duke's B colon cancer.

The GT4 (LVI+) profile had been narrowly misclassified as LVI-, the GT86 (LVI-) profile had been misclassified as LVI+. Although not done in the present study, actual LVI status could be obtained pre-operatively via endoscopy (before nodal status is known), then LVI status used to improve PCI risk prediction by removing the two misclassified profiles. The remaining profiles, in same order as in Table 2a but with rank numbers 1–18, had these numbers of involved nodes respectively: 0,0,1,3,1,0,0,0,3,9,7,7,1,0,1,3,8,3 (ranks 1–9 had 8 involved nodes, ranks 10–18 had 39 such nodes). PCI ranking of the 18 profiles correlated with ranking by number of involved nodes on the Spearman correlation test ( $r_s = 0.4837$ ), with probability  $P < 0.0211$  of obtaining as high or higher positive correlation by chance. Thus, statistical analysis of the blind test clearly indicated that the PCI model is a good predictor of nodal involvement.

The above results represent a one-shot blind test to predict risk of node positivity. However, once nodal status was revealed, the analysis was repeated to show what would happen if the training had started with different profiles. First, profiles RT29 (LVI+) and RT191 (LVI-) were considered but, for the selected 28 gene expression values from each of these profiles, the resulting training input was strongly colored. Next, RT8 (LVI+) and RT191 were considered as a training pair and this time, for the selected gene expression values, the resulting 56-point training input was nearly white. For this training input, one particular PCI model emerged as best over every evaluation set. Over the training data, the model had an MSE of 44.44%. In predicting LVI status, the model correctly classified 2 of 4 LVI- test profiles, and all 14 LVI+ test profiles (Matthews' correlation coefficient  $\phi = 0.66$ , Fisher's exact test  $P < 0.0393$ ). When all 20 profiles were ranked by mean PCI output (Table 2b) with largest positive mean output at the top, the N+ profiles tended to have larger rank numbers, as expected ( $P < 0.0314$ , on exact Mann-Whitney U-test). Corresponding point biserial correlation of increasing rank number with node positivity was +0.44, with probability  $P < 0.028$  of obtaining as high or higher positive correlation by chance. Thus, very similar results were obtained as in the blind test.

## PCI Application to other small data sets

PCI has shown similar predictive accuracy on two small datasets from the literature (Golub et al. 1999, MacDonald et al. 2001). On the first dataset (7 successful, 8 failed treatments), PCI predicted treatment response of acute myeloid leukemia patients, employing a very similar training and test regime as used here (Korenberg 2002). On the second dataset, PCI predicted metastatic potential of primary medulloblastomas (9 metastatic, 14 non-metastatic), where the first 3 profiles from each class were used to develop the model, and the remainder served as the test set (Korenberg 2004). For both small datasets, the PCI predictors reached statistical significance, performing similarly as herein.

## Discussion

This work demonstrates in a blind test that PCI could be used to predict node positivity in gastric cancer. In a clinical setting, the mucosal biopsy obtained pre-operatively during endoscopy could be used both to detect LVI and for gene expression profiling. Then the PCI mean output could be employed to place the patient within a ranking for which the other cases have known nodal status, and thus predict both the risk of node positivity and, more generally, tumor aggressiveness. Even if the LVI status is not determined pre-operatively, the PCI predictor would still be useful so long as the required 28 gene expression values are measured from the biopsy, since the PCI ranking's correlation with node positivity over all cases was three times that for LVI. In addition, as shown above, the PCI 28-gene model could be combined with demonstrated LVI status to better predict risk of nodal involvement prior to surgery. This could provide useful information for clinical practice, including identification of patients who would benefit from neoadjuvant therapy. As noted above, the number of involved nodes in the bottom half of the PCI ranking is then 39, almost five times the number (8) in the top half. Hence new cases whose LVI status is correctly predicted and whose PCI output ranks them with the bottom half of the patients might be candidates for more aggressive or neoadjuvant treatment.

Several genes used by the predictor have known cancer roles; cell division cycle 14 homolog A was especially up-regulated in LVI+. Immunoglobulin kappa variable (IGCK) was the most down-regulated in LVI+, followed by proline dehydrogenase (PRODH2).

The most important limitation in this study was sample size. An encouraging consideration is that the predictor was validated in a blind test. The methods presented here offer researchers alternative methods of data analysis that may have advantages over conventional statistical approaches. We made an attempt to use the PCI method in addition to statistical approaches to illustrate the advantages and limitations of each of these techniques. Genomics studies are likely to produce more complex data sets and therefore should be amenable to newer methods of analysis. The use of machine learning techniques in the context of genomic data is now becoming increasingly popular. However, in most biological studies, the sample size available to the researchers is often the limiting factor, despite the high throughput capabilities of the analytical techniques. In this study we demonstrated the PCI method as an alternative under conditions of limited sample size. The conclusions drawn here should be further validated in an independent study.

**Acknowledgments** CEC is Canada Research Chair in Oncology. Funding support to the PolyomX Program was provided by the Alberta Cancer Foundation and the Alberta Cancer Board and through the Clinical Investigator Program of the Royal College of Physicians and Surgeons of Canada and the University of Alberta Hospital Foundation to BJD.

## References

- Botwell D, Sambrook J (eds) (2003) DNA microarrays: a cloning manual. Cold Spring Harbour Laboratory Pr, Cold Spring Harbour, NY
- Dicken BJ, Saunders LD, Jhangri GS et al (2004) Gastric cancer: establishing predictors of biologic behavior with use of population-based data. *Ann Surg Oncol* 11(6):629–635
- Dicken BJ, Bigam DL, Cass C et al (2005) Gastric adenocarcinoma: review and considerations for future directions. *Ann Surg* 241(1):27–39
- Dicken BJ, Graham K, Hamilton SM et al (2006) Lymphovascular invasion is associated with poor survival in gastric cancer: an application of gene-expression and tissue array techniques. *Ann Surg* 243:64–73
- Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537
- Karpeh MS, Brennan MF (1998) Gastric carcinoma. *Ann Surg Oncol* 5:650–656
- Karpeh MS, Leon L, Klimstra D et al (2000) Lymph node staging in gastric cancer: is location more important than number? An analysis of 1, 038 patients. *Ann Surg* 232(3):362–371
- Khan J, Wei JS, Ringnér M et al (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 7(6):673–679
- Kirkpatrick P (2002) Look into the future. *Nat Rev Drug Discov* 1(5):334
- Korenberg MJ (1991) Parallel cascade identification and kernel estimation for nonlinear systems. *Ann Biomed Eng* 19:429–455
- Korenberg MJ (2002) Prediction of treatment response using gene expression profiles. *J Proteome Res* 1:55–61
- Korenberg MJ (2003) Gene expression monitoring accurately predicts medulloblastoma positive and negative clinical outcomes. *FEBS Lett* 533:110–114
- Korenberg MJ (2004) On predicting medulloblastoma metastasis by gene expression profiling. *J Proteome Res* 3:91–96
- Listgarten J, Graham K, Damaraju S et al (2003) Clinically validated benchmarking of normalisation. *Appl Bioinformatics* 2(4):219–228
- Lowry R (2007) *Concepts and Applications of Inferential Statistics*. Poughkeepsie, NY: Vassar College. <http://faculty.vassar.edu/lowry/webtext.html>
- MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. *Psychol Methods* 7:19–40
- MacDonald TJ, Brown KM, LaFleur B et al. (2001) Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nature Genet* 29:143–152. Datasets: [http://pepr.cnmcresearch.org/browse.do?action=list\\_prj\\_exp&projectId=63](http://pepr.cnmcresearch.org/browse.do?action=list_prj_exp&projectId=63)
- von Rahden BH, Stein HJ, Feith M et al (2005) Lymphatic vessel invasion as a prognostic factor in patients with primary resected adenocarcinomas of the esophagogastric junction. *J Clin Oncol* 23(4):874–879
- Wang Y, Jatkoa T, Zhang Y et al (2004) Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol* 22(9):1564–1571
- Weiss MM, Kuipers EJ, Postma C et al (2003) Genomic profiling of gastric cancer predicts lymph node status and survival. *Oncogene* 22(12):1872–1879
- Weiss MM, Kuipers EJ, Postma C et al (2004) Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival. *Cell Oncol* 26(5–6):307–317