Biomarkers, 2010; 15(8): 693–706

informa
healthcare

ORIGINAL ARTICLE

Multi-gene biomarker panel for reference free prostate cancer diagnosis: determination and independent validation

Miroslava Cuperlovic-Culf[1,2], Nabil Belacel[2], Michelle Davey[1], and Rodney J. Ouellette[1,3]

[1]Atlantic Cancer Research Institute, Moncton, NB, Canada, [2]Institute for Information Technology, National Research Council of Canada, Moncton, NB, Canada, and [3]Dr. Georges Dumont Hospital, Moncton, NB, Canada

Abstract

Identification of biomarkers that can accurately and reliably diagnose prostate cancer is clinically highly desirable. A novel classification method, *K*-closest resemblance was applied to several high-quality transcriptomic datasets of prostate cancer leading to the discovery of a panel of eight gene biomarkers that can detect prostate cancer with over 96% specificity and sensitivity in leave-one-out cross-validation. Independent validation on clinical samples confirmed the discriminatory power of this gene panel, yielding over 95% accuracy of diagnosis based on receiver-operating characteristic curve analyses. Different levels of validation of the proposed biomarker panel have shown that it allows extremely accurate diagnosis of prostate cancer. Application of this panel can possibly add a fast and objective tool to the pathologist's arsenal following further clinical testing.

Keywords: *Prostate cancer; molecular diagnostic; gene expression microarray analysis; data analysis; biomarker discovery*

Introduction

Prostate cancer has evolved as a major health problem in the male population of the Western world. It is currently the most commonly diagnosed malignancy and the second leading cause of cancer death, representing nearly 29% of all male cancer deaths (Li et al. 2006, Karan et al. 2003, Peehl 2005, Berger et al. 2004, Rose et al. 2005, Dhanasekaran et al. 2001, Lapointe et al. 2004, Luo et al. 2001, Singh et al. 2002,Varambally et al. 2002, Welsh et al. 2001, Rhodes et al. 2002, Tanguay 2000, Yu et al. 2004, Tomlins et al. 2007, LaTulippe et al. 2002, Mhawech-Fauceglia et al. 2007, Hessels et al. 2010, Graif et al. 2007, Esserman et al. 2009). The current methods for detecting prostate cancer include the prostate-specific antigen (PSA) blood test, a digital rectal examination (DRE), transrectal ultrasound (TRUS) and biopsy. However each of these tests suffers from significant limitations in sensitivity and specificity. New biomarkers are needed for prostate cancer diagnosis and treatment planning (Li et al. 2006, Graif et al. 2007, Esserman et al. 2009). Newly developed single gene biomarkers, such as initially highly promising prostate cancer antigen 3 (PCA3), are still under investigation. The utility of such single marker tests is often questionable as the expression level of any one gene probably depends on a number of tumour and non-tumour related factors and can, thus, lead to high number of false-positive diagnosis (Hessels et al. 2010). Furthermore, prostate cancer metastasis that may occur *via* lymphatic and haematogenous pathways can target almost any distant organ. When the diagnosis of metastatic prostate cancer is suspected immunohistochemistry is routinely used, usually analysing PSA and prostate-specific acid phosphatase (PAP). Even though the use of these two markers can improve diagnosis, there are many reports showing that the accuracy of this test depends on the tumour grade

Address for Correspondence: MiroslavaCuperlovic-Culf,Institute for Information Technology, National Research Council of Canada, Moncton, 55 Crowley Farm Road, Suite 1100, Moncton, NB E1A 7R1, Canada. Tel: 506-861-0952. Fax: 506-851-3630. E-mail: miroslava.cuperlovic-culf@nrc-cnrc.gc.ca

(Received 04 June 2010; revised 16 July 2010; accepted 25 July 2010)

ISSN 1354-750X print/ISSN 1366-5804 online © 2010 Informa UK, Ltd.
DOI: 10.3109/1354750X.2010.511268

and treatment. Therefore, additional markers would be highly beneficial (Tanguay 2000).

DNA microarrays are able to provide an unbiased screen of expression levels for thousands of genes in a range of samples. The application of DNA microarrays in cancer research over a decade has resulted in the availability of an abundance of literature on gene expression measurements for many different applications. Several recent studies have systematically assessed gene expression profiles in prostate normal and cancer tissues (Dhanasekaran et al. 2001, Lapoine et al. 2004, Luo et al. 2001, Singh et al. 2002, Varambally et al. 2002, Welsh et al. 2001), providing a wealth of relevant and openly accessible information on changes in mRNA levels relating to this disease. Therefore it can be argued that sufficient experimental gene expression information is already available for the extraction of an optimal biomarker set. The challenge lies in applying the best data analysis tools for the extraction of useful information.

Combined analysis of published microarray data and the available information about genes, proteins, metabolites and cellular pathways in the prostate gland provides a significant base for computational determination of clinically relevant biomarkers and their *in silico* validation. Optimization as well as development of methods for supervised and unsupervised analysis of microarray data is expected to bring useful clinical information, including determination of diagnostic and prognostic biomarkers. Detailed analyses of the properties of these biomarkers can then lead to the determination of significant pathways, providing a novel approach for their validation and also possibly leading to non-invasive diagnostics and improved treatment options.

Recent statistical analysis of clinical data suggests increasing the number of biopsies in patients (Graif et al. 2007). Subsequently, fast and accurate diagnosis from biopsy samples will be needed and this requires the identification of optimal sets of diagnostic markers. Several genes and proteins have been individually proposed as diagnostic markers for prostate cancer. However, expression levels of a single gene or a single protein can be altered by non-cancer-related factors leading to false-positive or false-negative diagnoses. Furthermore, reliance on changes in expression levels of a single biomarker for diagnosis requires measurement of the absolute concentration and the determination of scale which is often problematic. Conversely, the use of a biomarker panel reduces the likelihood of false-positive or false-negative results. Furthermore, the inclusion of genes that have both higher and lower expression levels in the malignant tissue provides an internal reference for the diagnostic test. Here, we present the discovery of a highly accurate gene panel for prostate cancer diagnosis as well as an independent *in silico* and experimental validation on a number of independent patient samples.

Materials and methodology

Microarray datasets

All datasets used for the biomarker discovery in this work were previously validated for accuracy in RNA expression measurements (references in Table 1). The primary set used for gene biomarker determination was published by Singh et al. (2002) as this was the largest dataset available. The other listed datasets were used for the initial validation. References to the original datasets as well as the number of tumour and normal samples and microarray platform used are given in Table 1. The descriptions of methods used for sample processing, microarray experimentation, and validation are given in the original publications.

Data preprocessing

Several standard methods for microarray data normalization as well as combination of measurement replicates are utilized here. The original Affymetrix data from Singh et al. (2002) was normalized using the quantile normalization method. The RMA (robust multi-array average) method proposed by Irizarry et al. (2003) was used for probe summarization. In this method, summarization is performed using the median polish and the mismatch probes' fluorescence values are used only for background adjustment. Finally, $\log_2$-transformed data was scaled to a median value of zero and a standard deviation of 1 over all samples and genes. Data preprocessing was primarily performed using microarray analysis tools provided as part of Partek Genomics Suite (Partek Inc., St Louis, MO, USA) and TMeV (http://www.tm4.org/mev/). Other methods and tools such as Gene Pattern (Reich 2006), Bioconductor (www.bioconductor.org) and Matlab

Table 1. Prostate microarray datasets used in this discovery, analysis and first stage of validation of diagnostic biomarker panel.

Dataset	Microarray platform	Probes	Normal samples	Cancer samples
Singh et al. (2002)	Affymetrix HG_U95Av2	12 626	50	52
Lapointe et al. (2004)	Spotted cDNA	4416043008	932	1349
Yu et al. (2004)	Affymetrix HG_U95Av2	12 626	23	64
Dhanasekaran et al. (2001)	Spotted cDNA	5520 genes; 4464 ESTs	6	59
Welsh et al. (2001)	Affymetrix HG_U95Av2	12626	9	25

(Matworks Inc, Natcik, MA, USA) were also utilized for method optimization.

Feature analysis and selection

The existence of several microarray studies of prostate cancer offers possibilities for biomarker discovery. Combined meta-analysis of different datasets has previously been suggested as an interesting way to increase the sample size (Grutzmann et al. 2005). However, direct combination of datasets is not possible because of the differences in sample processing and experimental procedures as well as probe differences. Although several methods for data transformation and meta-analysis have been proposed, there is still no consensus or clear guidance on the best meta-analysis procedure. Thus, meta-analysis, although leading to increased sample size, can lead to erroneous results. Therefore, we used an alternative approach that still takes advantage of the many different datasets available. The largest dataset provided by Singh et al. (2002) was used for biomarker panel determination. This high-quality dataset has matched normal (50) to cancer (52) samples. The other published datasets were used for first round of independent validation of gene panels.

Determination of the most significantly differentially expressed genes from publically available microarray dataset provided by Singh et al. (2002) was performed in two steps. Initial determination of a large subset of all significantly differently expressed genes was performed using significant analysis of microarrays (SAM) (Tusher et al.2001), Wilcox (Mann–Whitney) rank sum test and ANOVA. The gene subset selected by SAM was chosen for further analysis because all the genes determined as significant using the SAM method were also selected by either ANOVA or Mann–Whitney procedures. The selection of a small, diagnostic biomarker panel was performed using a method developed in our group called *K*closest resemblance (*K-CR*) (Belacel 2004, Belacel 2000). *K-CR* is a type of multiple criteria decision analysis (Belacel 2000) where the determination of a subset of *K* prototypes representing the closest resemblance to an object, or in this case a tissue sample, is based on the scoring function. In *K-CR*, final assignment of the tissue samples to different classes or to different types of tumours is performed using the majority-voting rule as used in the *k* nearest neighbour procedure (Dasarathy 1991). The method generates gene weights that can be used to select a subset of genes that results in the highest classification specificity and sensitivity for the dataset. In order to avoid over-fitting, the performance of the method for each gene subset was evaluated using the 'leave-one-out' technique (reviewed in Cuperlovic-Culf et al.2005). In order to avoid suboptimal selection, results from several consecutive runs of *K-CR* were evaluated based on the accuracy of classification, validation against other datasets and biological knowledge. The method is schematically represented in Figure 1.

Gene analysis

The analysis of properties and characteristics of biomarker genes in terms of their expression, function, sequence properties, interactions, gene ontologies and chromosomal localization as well as gene correlation networks was performed using various methods for gene clustering, sequence and pathway analysis. The methods used for gene and sample clustering included Fuzzy J-Means (Belacel 2004) as well as principal components analysis (PCA) and hierarchical clustering as applied in Partek Genomics Suite (Partek Inc.). Gene properties and expression levels in other microarray experiments were investigated using Oncomine (Rhodes et al.2007). Gene network analysis was performed using Pathway Studio (Ariadne Genomics, Rockville, MD, USA). Pathway Studio develops networks of related genes for a given list of genes based on literature search.

Experimental validation

The panel of genes was experimentally validated using quantitative real-time polymerase chain reaction (qPCR) gene expression measurements in an independent set of 19 tumour, 13 adjacent-to-tumour and 14 normal prostate RNA samples extracted from patient tissue biopsies purchased from a commercial source (Asterandplc, Detroit, MI, USA). Each RNA sample was assessed for 28S and 18S ribosomal degradation using the Experion™ Automated Electrophoresis System (Bio-Rad Laboratories, Hercules, CA, USA). All samples demonstrated 28S/18S ratios of 1.6 or higher. Total RNA (500ng) was synthesized into cDNA using oligo$(dT)_{20}$ primers and SuperScript III reverse transcriptase (Invitrogen, San Diego, CA, USA) according to the manufacturer's suggested protocol. Each sample was reverse transcribed in triplicate. Real-time primers for the eight target genes were designed using Primer3 v0.3.0 (primer3.sourceforge.net) and the appropriate mRNA consensus sequences from the NCBI Entrez nucleotide database (ncbi.nlm.nih.gov). PCR product sizes were limited to a maximum of 150 bases in length. Primers were designed to amplify intron-spanning regions and were supplied by Integrated DNA Technologies (Coralville, IO, USA). qPCR reactions were done on a Mastercyclerep *realplex²* (Eppendorf, Westbury, NY, USA) instrument using B-R SYBR® Green SuperMix for iQ™ (Quanta BioSciences, Gaithersburg, MD, USA) reagent, optimized primer amounts for each primer set and cDNA input amounts of 10ng per 20µl reaction. Each gene measurement for each sample was

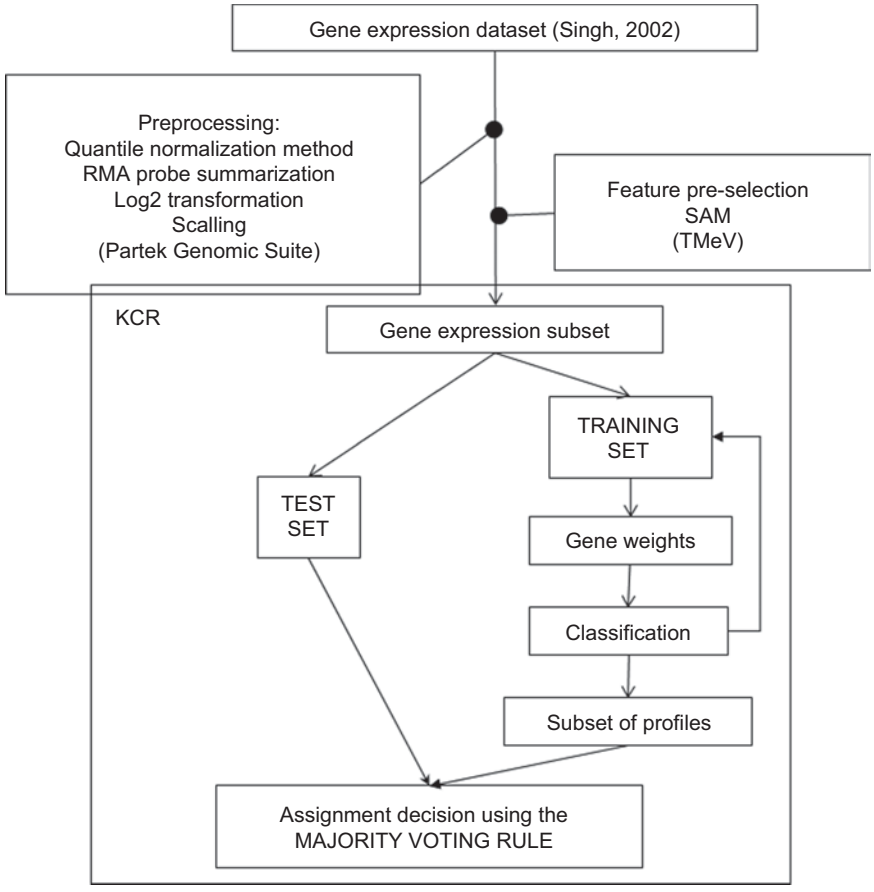

Figure 1. Schematic description of the data analysis procedure including steps in data preprocessing, feature preselection and final feature selection using *K*-closest resemblance procedure (*K-CR*).

performed in triplicate.Melting curves were checked to ensure specificity.

qPCR efficiencies (E) for each gene were calculated using the formula $E=10^{[-1/\text{slope}]}$. The relative expression ratio (R) for each target gene was calculated based on E and the quantification cycle (Cq) deviation of each test sample versus a control sample (normal prostate RNA pool) using the following equation:

$$R=(E_{target})^{\Delta Cq\ target\ (control-sample)}\ (\text{Pfaffl, 2001}) \quad (1)$$

Results and discussion

Feature analysis:biomarker discovery

The microarray dataset made publically available by Singh et al. (2002) was used for biomarker panel determination. This high-quality dataset includes expression data for over 12,000 genes measured in matched normal (50) and prostate cancer (52) biopsy samples. Several statistical methods were tested for their ability to exclude from the dataset genes which show no significant expression level change between cancer and normal samples. These statistical methods include ANOVA, Wilcox rank sum test and SAM. The subsets of significantly differentially expressed genes obtained using these three statistical methods differed and is shown as a Venn diagram in Figure 2.

However, all the genes determined as significant using the SAM method were also selected by either ANOVA or Mann–Whitney procedures, with in fact a large majority of these genes selected by all three methods. Following this observation, further analysis was performed on the SAM-selected gene subset (557 genes). This set was used in the second step of the feature selection. In this step we have used the *K-CR* method developed in our group (Graif et al. 2007). Rather than selecting significant genes from the set, the *K-CR* method deletes the genes which are insignificant in the classification. This type of analysis allows the cross-interaction of significant genes to contribute to the accuracy of classification. As *K-CR* is a heuristic method, several computational analyses were performed and several possible biomarker subsets were determined. The most optimal biomarker set was chosen based on the highest sensitivity, validation

against other published microarray datasets and literature analysis of genes in these panels. The selected biomarker panel (Table 2) showed 96% sensitivity and 98% specificity of classification in the test set using the 'leave-one-out' cross-validation technique.

The selected diagnostic biomarker panel consists of eight genes - three genes that are overexpressed and five that are underexpressed in prostate tumours relative to normal prostate tissue. Presented in the Table are names, synonyms and NCBI gene IDs. Also included is the chromosome location for each gene. The possible significance of the gene chromosomal location is outlined below. The expression levels of the biomarker gene panel based on the microarray measurements (Singhet al. 2002), following normalization and scaling as described in Methods, are shown in Figure 3 as a heat map.

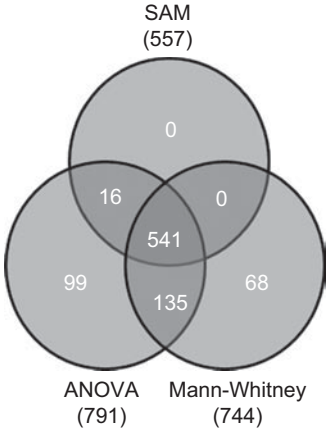

Figure 2. Venn diagram comparing gene subsets selected using ANOVA, Mann–Whitney (Wilcox rank sum) and significant analysis of microarrays (SAM) tests. The ANOVA and Mann–Whitney gene subset is selected to have $p<0.0005$; SAM subset is selected with $D=1.287$ and 0 false-discovery rate.

For optimal sample diagnosis, values for all eight genes should be utilized and the final sample classification should be performed using the majority voting method or the ratio method described below. From the expression level heat map presented in Figure 3 it can be seen that when majority voting from the results of all eight genes is used for diagnosis only one normal and one tumour sample are misclassified, resulting in 96% sensitivity and 98% specificity in detecting the presence of prostate cancer in tissue samples.

Independent microarray set validation

The independent analysis of the classification power of the selected biomarker panel was performed using the microarray dataset published by Lapointe et al. (2004) as well as on several other datasets available in the Oncomine database (Rhodes et al. 2007) (Tables 1 and 3). The dataset provided by Lapointe and co-workers consists of gene expressions measured using spotted cDNA microarrays rather than the Affymetrix short oligonucleotide chips used by Singh and co-workers. Therefore, exploration of gene expressions based on the Lapointe dataset (Lapointe et al. 2004) could demonstrate whether highly significant and diagnostic gene expression changes observed on Singh et al. dataset (Singh et al. 2002) come from phenotypical changes rather than technological biases of the Affymetrix platform. The microarray platforms used by Lapointe and colleagues included only five out of the eight marker genes in our list (*GSTM4*, *HPN*, *ITSN1*, *LTBP4* and *XBP1*). However, the hierarchical clustering (Figure 4) and the PCA (Figure 5) of the expression level values for these five biomarkers still showed good separation between tumour and normal samples. Figure 4 presents the comparison of the result of hierarchical clustering

Table 2. Biomarker gene set determination using *K-CR* analysis on published prostate cancer datasets (Belacel 2007).

Name	Synonyms	UniGene	RefSeq	Chromosome	Expression T/N
Intersectin 1 (SH3 domain protein)	*ITSN1,ITSN, SH3D1A, SH3P17, MGC134948, MGC134949*	Hs.160324	NM_003024	21q22.1-q22.2	Down
Glutathione S-transferase M4	*GSTM4, GTM4, GSTM4-4, MGC9247, MGC131945*	Hs.348387	NM_000850	1p13.3	Down
Latent transforming growth factor beta binding protein 4	*LTBP4*	Hs.466766	NM_001042544	19q13.1-q13.2	Down
D component of complement (adipsin)	*DF, CDF, ADN, PFD*	Hs.155597	NM_001928	19p13.3	Down
Nel-like 2	*NELL2, NRP2*	Hs.505326	NM_006159	12q13.11-q13.12	Down
X-box binding protein 1	*XBP1, XBP2, TREB5*	Hs.437638	NM_005080	22q12.1|22q12	Up
Prostate-specific membrane antigen	*PSM, FGCP, FOLH, GCP2, PSMA, mGCP, GCPII, NAALAD1, NAALAdase*	Hs.654487	NM_004476	11p11.2	Up
Hepsin (transmembrane protease, serin 1)	*HPN, TMPRSS1*	Hs.182385	NM_182983	19q11-q13.2	Up

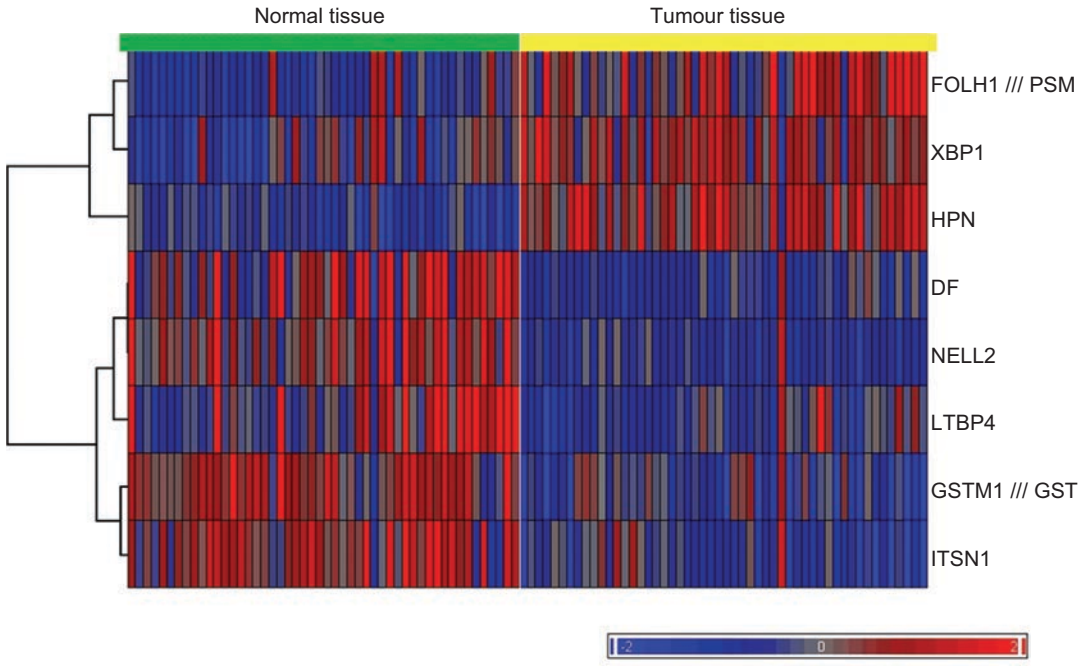

Figure 3. Heat map showing the gene expression values for the biomarker gene panel set as determined in the work by Singh et al. (2002).

Table 3. *p*-Values for difference between expression levels for marker genes between cancer and normal prostate samples in several different studies available in the Oncomine database (Rhodes et al. 2007).

Gene	Rhodes 2002	Lapointe 2004	Luo 2001	Dhanasekaran 2001	Tomlins 2007	Welsh 2001	LaTilippe 2001	Vanaja 2003	Yu 2004	Holzbeierlein 2004
ITSN1	3.5E-10	9.7E-9								
GSTM4	3.7E-19			1.7E-7	4.6E-6					
LTBP4	3.7E-5					6.4E-6				
DF	6.4E-10									
NELL2	1.2E-12	4.1E-6	8.8E-5							
XBP1[a]		7.4E-5		1E-4	9.5E-6					
FOLH1							1.6E-8	1.6E-6		7.2E-6
HPN	6.1E-25		1.8E-7	2.5E-9		2.1E-8		3.8E-8	5.2E-19	

[a]Only comparison metastatic prostate cancer/prostate carcinoma is available in Oncomine; *XBP1* overexpressed in prostate carcinomas.

of samples based on the five gene set (Figure 4A), all genes in the filtered set (Figure 4B), and a subset of five randomly selected genes (Figure 4C). From the results of hierarchical clustering it can be seen that when using the five proposed diagnostic genes, only 7 out of 41 normal samples and 3 out of 62 cancer samples were misclassified resulting in sensitivity and specificity of 95% and 81%, respectively, in this completely independent dataset. On the other hand, sample clustering based on the complete gene dataset as well as the randomly selected five gene subset did not lead to visible grouping of samples by their type.

A similar observation was made from PCA analysis (Figure 5). Once again, good separation of samples based on PCA using the proposed five-gene diagnostic set was obtained (Figure 5A), while poor separation was observed using five randomly selected genes (Figure 5B). As in hierarchical clustering, classification based on the data for the five proposed marker genes appears to be wrong for only three to four cancer and seven to eight normal samples.

The lower sensitivity and specificity obtained in this independent set relative to the original dataset can be easily understood considering that here it was only possible to use five genes from the panel and that spotted cDNA technology introduces larger noise in measurements. However, it should still be noted that the subset of five genes from our biomarker panel allows clear distinction between lymph node metastasis, tumours and, even more clearly, normal samples. Thus it can be inferred that in the metastatic cells expression levels of the marker genes change even further from their normal cell levels. This observation will be further investigated in the future.

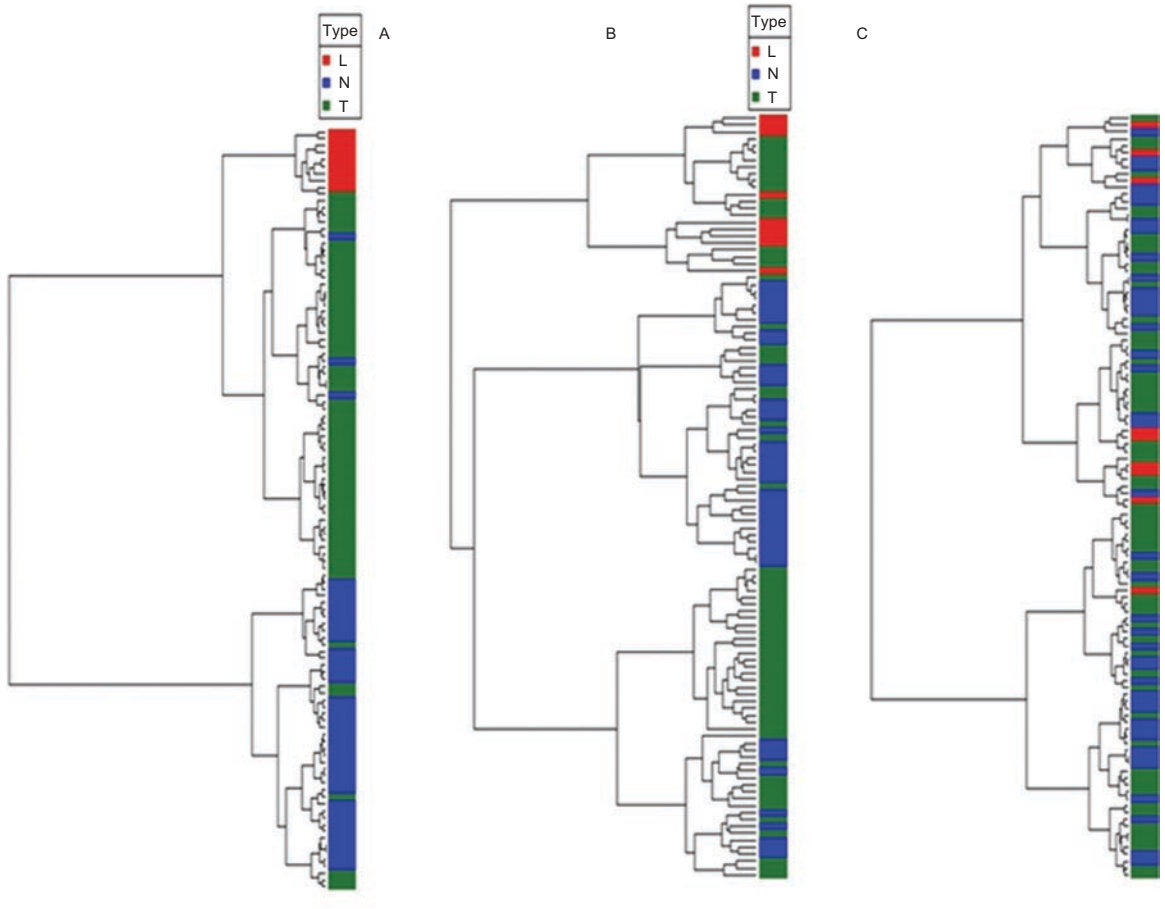

Figure 4. The result of hierarchical agglomerative clustering with Ward's linkage method for Lapointeet al.data(Lapointeet al.2004).Ward's linkage method combines the two clusters which minimizethe increase in total error sum of squares (ESS). The ESS ofa cluster is the sum of squares of the deviations from the meanvalue. (A) Sample clustering using five of eight biomarker genes. Genes available in Lapointe dataset after filtering were: *GSTM4*, *HPN*, *ITSN1*, *LTBP4* and *XBP1*. (B) Sample clustering using all genes in the filtered Lapointe dataset. (C) Sample clustering based on data for five randomly selected genes using the filtered Lapointe dataset.

The Oncomine analysis of the biomarker genes showed consistent expression levels difference between cancer and normal tissues across many different prostate tissue datasets for all eight genes in the panel (Table 3). The ten prostate cancer studies included in Oncomine differed highly in the number of samples, the experimental protocol and the platform. Thus, the observed large variation in the actual *p*-values for biomarker genes was to be expected (Table 3). However, for all eight marker genes in all available datasets *p*-values showed a significant difference between gene expression in tumours and normal samples.

Experimental validation and application method

To validate further the diagnostic utility of the eight-gene panel derived from microarray dataset analyses, the expression levels of these biomarkers were tested using an independent measurement technique, real-time PCR and an independent set of commercially obtained patient prostate samples. The relative expression levels for each of the eight genes were determined for each of 19 tumour, 14 normal and 13 normal/adjacent-to-tumour samples (the measurements are available from authors upon request).

In addition to independent, experimental, validation the goal of real-time PCR analysis was to devise a diagnostic method that would be robust to individual gene changes and that could take full advantage of the presence of both overexpressed and underexpressed genes in the proposed diagnostic panel. Using the average value of the relative expression results for *XBP1*, *PSMA* and *HPN* (overexpressed genes) divided by the average value of the relative expression results for *GSTM4*, *LTBP4*, *ADIPSIN*, *NELL2* and *ITSN1* (underexpressed genes) allowed for a direct and simple diagnostic index calculation. Furthermore, such application of the panel allowed determination of the index without utilization of reference genes, i.e.'house-keeping' genes, which are necessary for absolute calculation of gene expression.

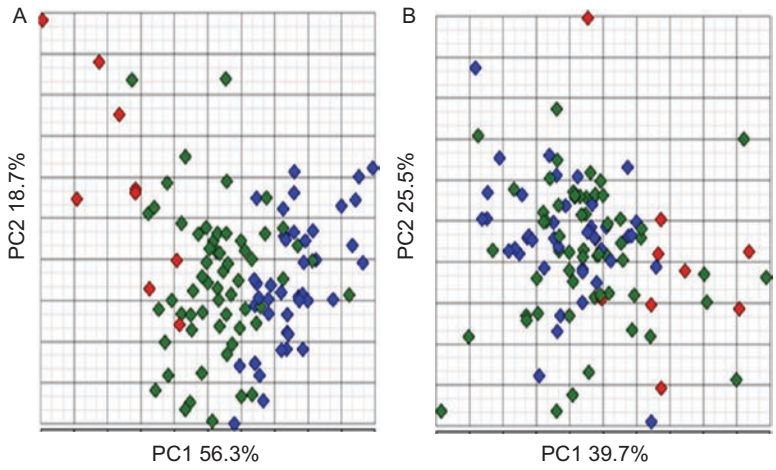

Figure 5. Principle component analysis of (A)five marker genes available in Lapointe experiments (Lapointeet al.2004) and (B) randomly selected five genes from the same set. Red symbols, lymph node metastasis; green, tumour; blue, normal prostate tissue.

Finally, averaging over several genes in the panel has the potential to reduce errors caused by individual variations in gene expression due to factors other than the cancer.

Therefore, we would like to propose that real-time PCR expression results for the eight-gene panel are employed for prostate cancer diagnosis using the following index:

$$d = \frac{\text{RHPN, RFOLH1, RXBP1}}{\text{RITSN1, RGTM4, RLTBP4, RNELL2}}$$

where R is the relative expression ratio calculated using Equation 1.

From the obtained results it is possible to define an index for diagnosis as:

$$d > 1 - \text{cancer tissue}; d < 1 - \text{normal tissue}$$

The boxplots of values obtained for index d, as determined using the equation above, for tumour, adjacent-to-tumour and normal sample groups within the independent sample set are presented in Figure 6.

From these measurements it can be observed that in tumours the ratio was always significantly over 1 with an average value for all tumour samples of about 5. For normal samples all ratios were below 1 with average value of approximately 0.5. For adjacent-to-tumour samples ratios were all over 1 with an average value of around 2. The receiver-operator curve (ROC) analyses for classification accuracy from ratios for normal versus tumour (Figure 7A), adjacent-to-tumour versustumour (Figure 7B) and normal versus adjacent-to-tumour (Figure 7C) are presented in Figure 7.

ROC analysis showed that the sensitivity and specificity for classification of tumour from normal was over 90% in these independent samples. The area under the curve (AUC) for the classification of tumour versus normal was 0.955. The ROC curves for classification of normal versus adjacent-to-tumour and adjacent-to-tumour vs. tumour demonstrated that although adjacent-to-tumour samples could still be distinguished from tumour samples with an AUC of 0.80, these samples still displayed very different gene expression from non-cancer normal samples.

The validity of this index will be further established in the ongoing preclinical trials (Dr Georges Dumont Hospital, Moncton, Canada); however its diagnostic power is quite indicative in these 46 samples tested thus far. The possibility for application of novel ROC analysis methods (Li & Fine 2008, Ogdie et al. 2010) including ROC surface analysis will be explored in the future once a larger number of samples becomes available.

Properties of genes in the biomarker panel

The network analysis of marker genes presented in Figure 8 shows a very general overview of all proteins/genes connected to the marker set either as regulators or as being regulated by the genes in the panel obtained from a detailed literature search. It is clear from this analysis that the marker genes have only a small number of co-regulators. Therefore, the expression levels of the proposed marker genes are unlikely to be affected by the same processes. Although biological variations in the samples can affect individual gene expression levels, they will not have a common effect on all marker genes. Therefore, when the selected group of genes is used as a panel for diagnosis with the method described above, it should be highly resistant to variations in individual genes.

The majority of the determined genes have already been independently analysed for their significance in cancer development and for their diagnostic power in

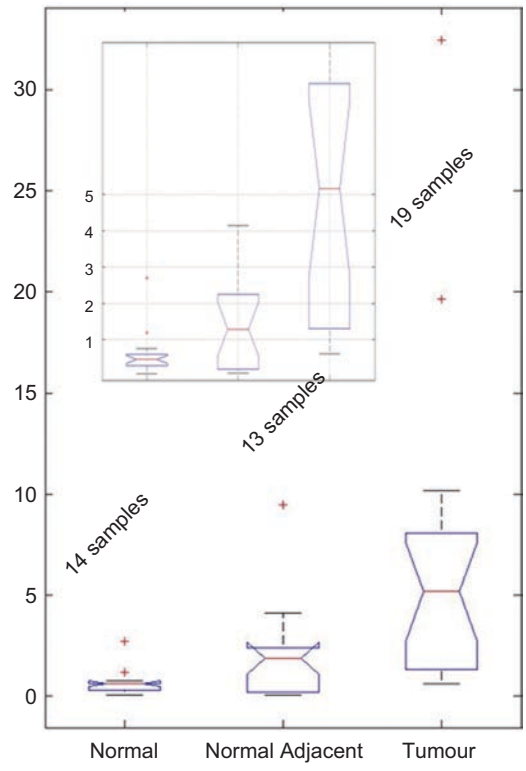

Figure 6. Boxplot of the real-time polymerase chain reaction measurements for normal, adjacent-to-tumour and tumour samples. The measurements are presented using the *d* index – ratio of the average value of all overexpressed relative to average value of all underexpressed genes ratio measurements.

prostate cancers. This provides further validation of the selected marker set. All of the genes in the panel that were found to be overexpressed in tumours are known to be highly diagnostic for prostate (hepsin, prostate-specific membrane antigen and *XBP1*) as well as breast cancers (*XBP1*). General analysis of literature search (Pathway Studio 5.0) looking for published connections between genes in the panel with cancer related processes is shown in Figure 9.

Hepsin (HPN) is a transmembrane protease which plays an essential role in cell growth and maintenance of cell morphology and is proposed as a stimulant of prostate cancer cell proliferation. There is overwhelming evidence in the literature of overexpression of hepsin in prostate cancer. Hepsin has been identified as a major significantly differentially expressed gene in all microarray studies (Nelson 2004) and this result has been independently observed in several hepsin gene analysis publications (Pal et al.2006, Magee et al.2001, Landers et al.2005, Chenet al.2003, Stephen et al.2004). In addition, HPN protein was determined by Northern blot as well as tissue microarray analysis to be a significant marker of prostate cancer, with overexpression in about 90% of the cancer samples studied (Dhanasekaran et al.2001). Furthermore, the analyses of single-nucleotide polymorphisms (SNPs) have shown the association of prostate cancer susceptibility with the 19q locus, with several SNPs at the location of *HPN* (Pal et al.2006).

Prostate-specific membrane antigen (PSMA, PSM or FOLH1), both gene and protein, has also been indicated as prostate cancer marker (Mhawech-Fauceglia et al.2007, Landers et al.2005). PSMA is a membrane-bound glycoprotein that functions as a folate hydrolase with high expression in benign and malignant prostate tissues. The *PSMA* gene is mapped on the short arm of chromosome 11. Recently published results from the PROGRESS study (Johanneson et al.2007) demonstrated a linkage between hereditary prostate cancers and chromosome 11p11.2-q12.2, the location which includes *PSMA*. This 750-amino acid protein is expressed in both normal and neoplastic prostate cells, however during the progression of cancer from androgen-sensitivity to androgen-independence, the overall expression of PSMA increases with its appearance in the plasma membrane. In addition, PSMA expression is known to be inversely correlated with the degree of prostatic cancer differentiation. A recent tissue microarray analysis on normal tissues and 3161 benign and malignant tumours (Tanguay 2000) has shown that sensitivity and specificity of PSMA in distinguishing prostate adenocarcinoma from any other type of malignancy is 65.9% and 94.5%, respectively. Furthermore, the sensitivity and specificity of PSMA in differentiating prostate cancers from urothelial cancer is 65.9% and 82.9%, respectively.

The final gene determined as a highly overexpressed marker in prostate tumours is X-box binding protein 1 (*XBP1*). XBP1 is a basic leucine zipper (bZIP)-containing transcription factor capable of specific binding to the endoplasmic reticulum stress response element I. XBP-1 is known to be associated to the so-called 'unfolded protein response' (UPR) (www.ihop-net.org/UniPub/iHOP). Changes in the endoplasmic reticulum and UPR can result in changes to secretory and extracellular proteins as well as steroids and lipid production and the upregulation of chaperones. mRNA levels of XBP1 were highly elevated in estrogen receptor (ER)-α-positive breast tumours (Lacroix & Leclerc 2004). Furthermore, Oncomine investigation (Rhodes et al.2007) of the *XBP1* sequence resulted in several corresponding sequences including known oncogenes Jun, ATF6, Fos, JunB, JunD and Maf. The *XBP1* gene is located on chromosome 22q12.1, close to the location of D22S689 that has been indicated as a potentially significant genetic marker for hereditary prostate cancers together with many other regions of chromosome 22q12.2 and 22q12.3 (Camp et al.2007). Furthermore, the *XBP1* gene and corresponding protein have been shown to have an increased expression in organ-confined prostate cancers relative to normal prostate

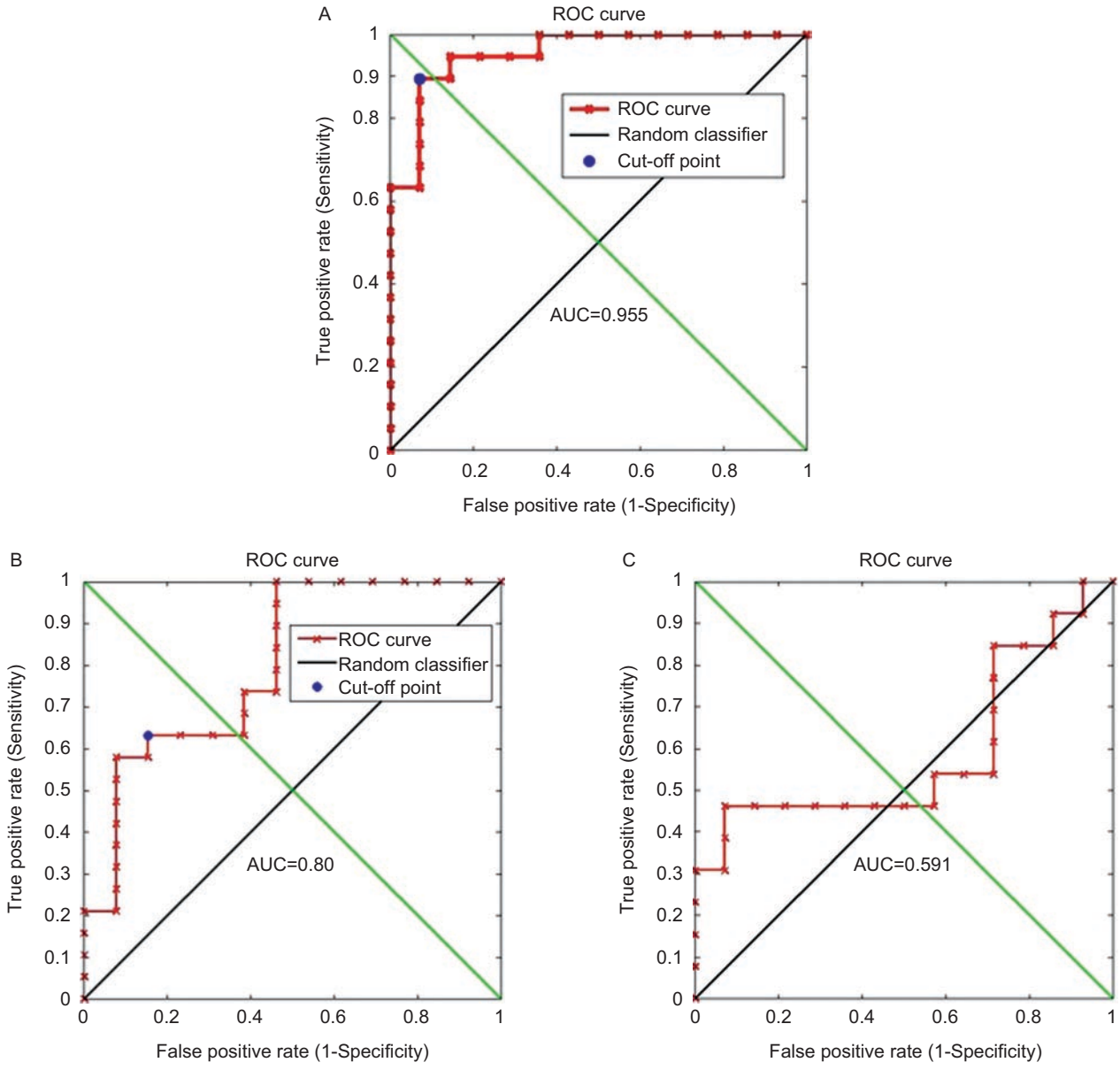

Figure 7. Receiver-operating characteristic (ROC) plots for the classification accuracy calculation for experimental samples. ROCs were calculated using the *d* index values. (A) Classification between normal and tumour samples. The cut-off point represents the classification sensitivity and specificity of 90%. (B) Classification between adjacent-to-tumour and tumour samples. The cut-off point represents the classification sensitivity and specificity of 60%. This result shows that adjacent-to-tumour samples are more closely related to normal than tumour samples. (C) Classification between adjacent-to-tumour and normal samples. Adjacent-to-tumour samples, although altered from completely normal tissues, cannot be used for diagnosing prostate tumours.

tissues; however XBP1 expression in refractory cancers has been shown to drop relative to both normal and cancer prostate tissues (Takahashi et al.2002). Work of Takahashi et al. (2002) concluded that the expression of XBP1 has an intimate connection with the differentiation of prostate adenocarcinomas.

In the proposed diagnostic panel, most of the genes that are underexpressed in prostate tumours have also been observed as part of larger groups of significantly underexpressed genes in other studies. LTBP4 is one of the isoforms of latent transforming growth factor-beta binding protein (TGF-β). LTBP proteins are believed to be structural components of connective tissue microfibrils and local regulators of TGF-β tissue deposition and signalling. In mouse experiments it has been shown that disrupted expression of LTBP4 results in severe pulmonary emphysema, cardiomyopathy and colorectal cancer. These highly tissue-specific abnormalities are associated

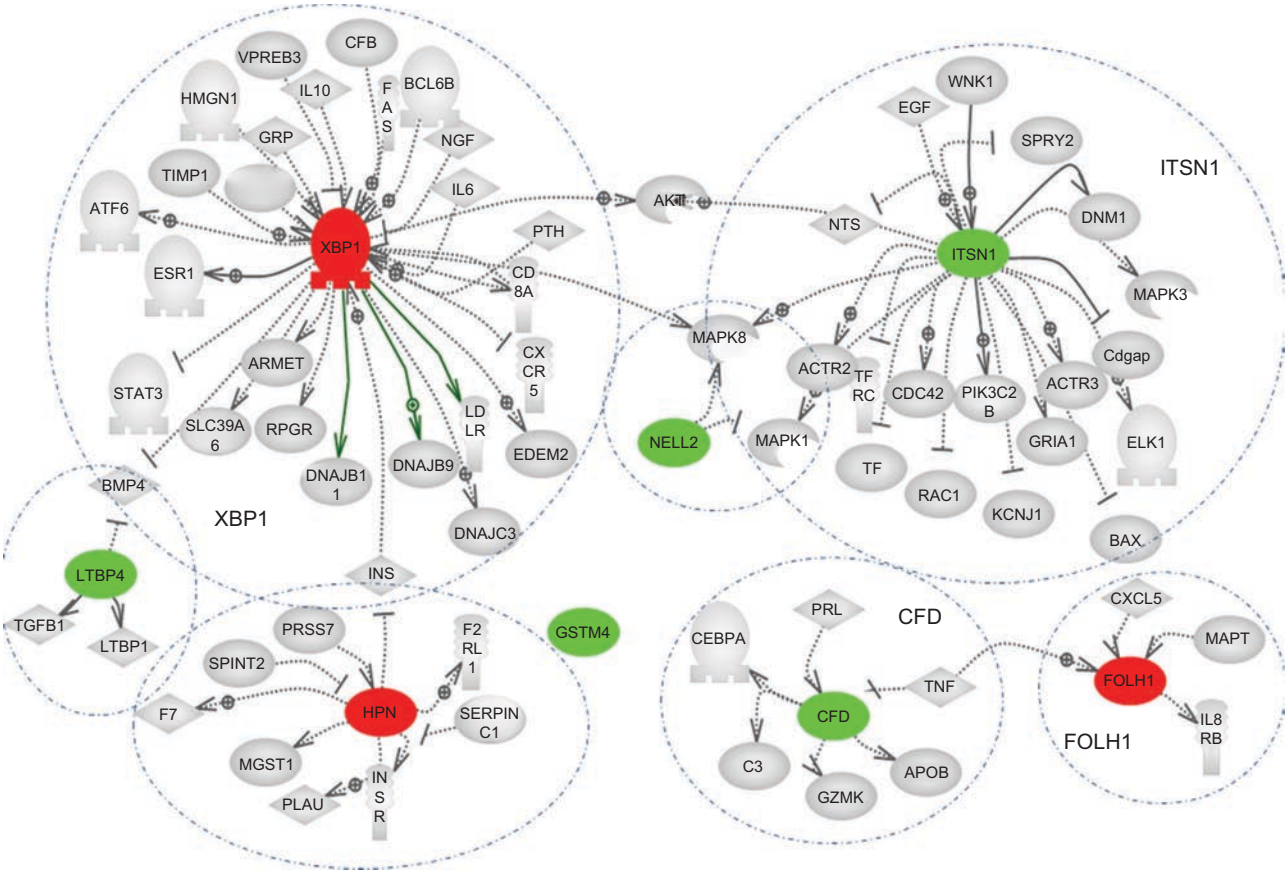

Figure 8. Regulators of the analysed genes determined from Pathway Studio literature search (Pathway Studio 7.0). Biomarkers overexpressed in cancer are shown in red and biomarkers underexpressed in cancer are shown in green. Individual regulatory partners are encircled.

with profound defects in the elastic fibre structure and with a reduced deposition of TGF-β in the extracellular space. As a consequence, epithelial cells have reduced levels of phosphorylated Smad2 proteins, overexpress c-myc, and undergo uncontrolled proliferation. This phenotype supports the predicted dual role of LTBP4 as a structural component of the extracellular matrix and as a local regulator of TGF-β tissue deposition and signalling (Sterner-Kock et al.2002), suggesting its possible role in prostate cancer.

Although there are limited data about the function and possible significance of adipsin, *GSTM4*, *NELL2* and *ITSN1* genes in cancers, once again they are consistently and significantly underexpressed in several different microarray experiments comparing prostate tumours and normal tissues (Rhodes et al.2002). Similarly to the *HPN* gene, both *LTBP4* and adipsin are located on chromosome 19, with *LTBP4* being in the region of known SNPs important for prostate cancer progression (Pal et al.2006).

Neural epidermal growth factor-like 2 gene (*NELL2*) encodes a cytoplasmic protein that contains epidermal growth factor (EGF)-like repeats. The encoded heterotrimeric protein may be involved in cell growth regulation and differentiation. Although no further analysis has been performed in terms of significance in prostate cancer development, the gene expression data for *NELL2* consistently shows underexpression in prostate cancers (Rhodes et al.2002).

Finally, glutathione *S*-transferase M4 is an enzyme involved in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione (www.ihop-net.org/UniPub/iHOP/). The genes encoding the GSTM4 class of gluthathione*S*-transferase are organized in a gene cluster on chromosome 1p13.3 and are known to be highly polymorphic. These genetic variations can change an individual's susceptibility to carcinogens and toxins as well as affect the toxicity and efficacy of certain drugs. Diversification of these genes has occurred in regions encoding substrate-binding domains, as well as in tissue expression patterns, to accommodate an increasing number of foreign compounds. Although very little work has been published on the significance of GSTM4 in cancers its function in detoxification might be indicative of its importance in carcinogenesis. We hope that the discovery of *GSTM4* as one of the members of the diagnostic panel will initiate further analysis of its involvement in cancer.

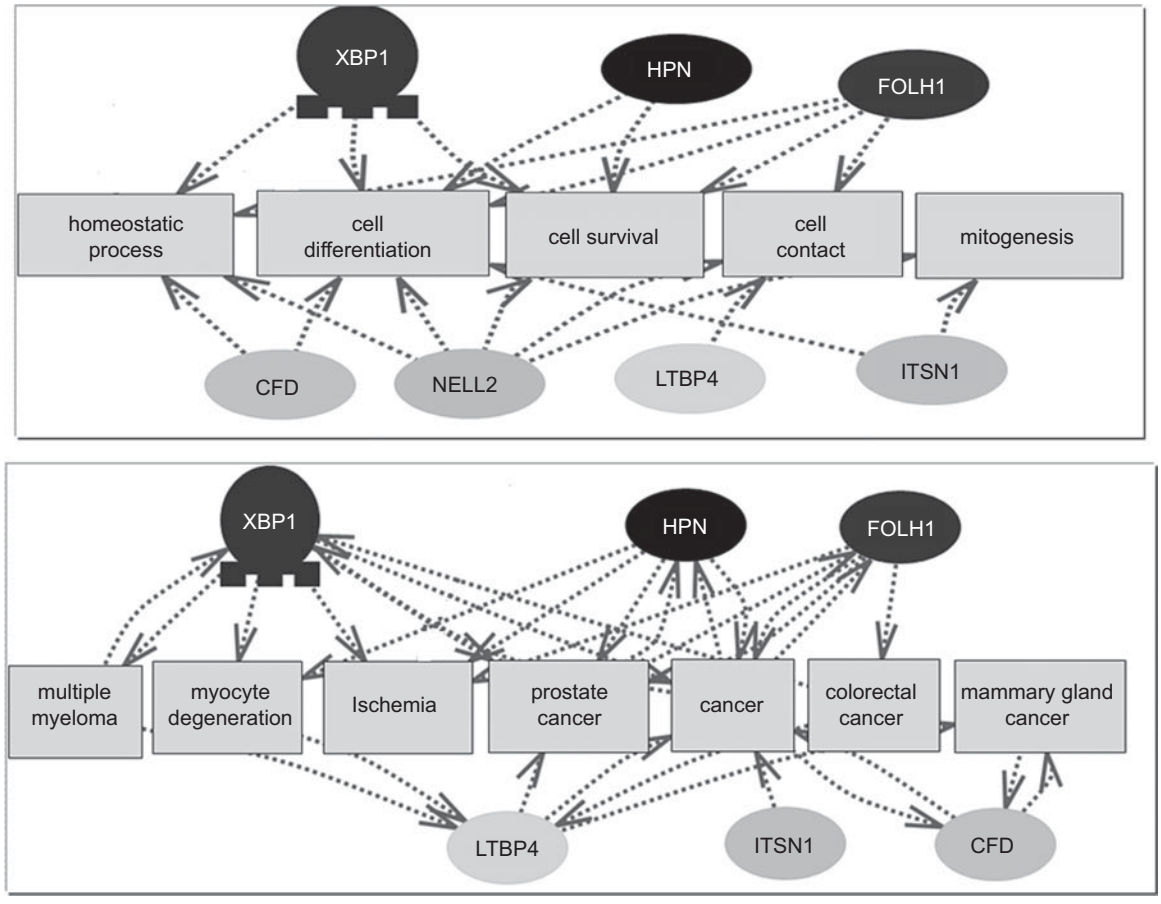

Figure 9. Pathway Studio 5.0 analysis of the literature connections between genes from the presented biomarker panel and cancer-related processes and terms.

In conclusion, high-throughput, i.e. omics measurements, provide a treasure trove of data that can be explored, among other applications, for biomarker discovery. The selection of a biomarker panel from such a large number of features can lead to a discovery of different panels that give a similar level of classification accuracy in a given data set. It is therefore necessary to validate, i.e. test, the quality of obtained biomarker panels against literature information, published gene expression data and independent experimental measurements. In this study we discovered and validated a panel of eight genes that provides highly sensitive and specific diagnosis of prostate tumours. The most accurate diagnosis can be made by using all eight genes in the panel, although a smaller subset of the selected genes also provides valid diagnostic information. The inclusion of both over- and under expressed genes provides an internal reference – the diagnosis can be made based on the relative change of expression of over- and underexpressed genes rather than from the absolute expression values based on some predetermined standard. The consistency of expression change across a large number of samples from ten different microarray gene expression studies was shown and the sample classification accuracy validated in two different data subsets consisting in total of 91 normal and 114 cancer samples. Experimental validation of this biomarker panel was performed on approximately 50 independent samples using a different method, qPCR, for measurement of gene expression. Once again the gene panel showed extremely high specificity and sensitivity for tumour diagnosis. In the future we will further validate our biomarker panel in ongoing preclinical trials and in prostate cells obtained from urine samples. We believe that this biomarker panel can provide an objective, fast and inexpensive addition to the pathologists tools for prostate cancer diagnosis.

Declaration of interest

Funding for this project was provided to ACRI by the Atlantic Innovation Fund (ACOA) and the Dr George L. Dumont Hospital Foundation and to the Institute for Information Technology by the National Research Council, Atlantic Initiative.

References

Belacel N, Cuperlovic-Culf M, Laflamme M, Ouellette R. (2004). Fuzzy J-Means and VNS methods for clustering genes from microarray data. *Bioinformatics*20: 1690–701.

Belacel N, Cuperlovic-Culf M, Ouellette R. (2007). Molecular methods for diagnosis of prostate cancer. US Patent WO2007030919 - 2007-03-22.

Belacel N. (2004). The k-closest resemblance approach for multiple criteria classification problems. In: HoaiLT, Tao PD, eds.*Modelling Computation and Optimization in Information Systems and Management Sciences*. London: Hermes Science Publishing. p. 525.

Belacel N. (2000).Multicriteria assignment method PROAFTN: methodology and medical applications. *Eur J Operational Res* 125:175–83.

Berger R, Feboo PG, Majumder PK et al. (2004). Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells. *Cancer Res* 64:8867–75.

Camp NJ, Cannon-Albright LA, Farnham JM, Baffoe-Bonnie AB, George A et al. (2007). Compelling evidence for a prostate cancer gene at 22q12.3 by the International Consortium for Prostate Cancer Genetics. *Human Mol Genet* 16:1271–8.

Chen Z, Fan Z, McNeal JE, Nolley R et al. (2003).Hepsin and mapsin are inversly expressed in laser capture microdisectioned prostate cancer. *J Urol* 169:1316–19.

Cuperlovic-Culf M, Belacel N, Ouellette R. (2005).Determination of tumour marker genes from gene expression data. *Drug Discovery Today* 10:42.

Dasarathy BV. (1991).*Nearest Neighbour (NN). Norms: NN Pattern Classification Technique*. Los Alamitos, CA:IEEE Computer Society Press.

Dhanasekaran SM, Barrette TR, Ghosh D et al. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* 1412: 2169–80.

Esserman L, Shieh Y, Thompson I. (2009). Rethinking screening for breast cancer and prostate cancer. *JAMA* 302:1685–92.

Gentleman R, Carey V, Huber W , Irizarry R , Dudoit S (2005).Bioinformatics and Computatonal Biology Solutions Using Bioconductor. Springer.

Graif T, Loeb S, Roehl KA et al. (2007).Under diagnosis and over diagnosis of prostate cancer. *J Urol* 178:88–92.

Grutzmann R, Boriss H, Ammerpohl O, et al. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 24:5079–88.

Hessels D, van Gils MPMQ, van Hooij O, Jannink SA, Witjes JA, Verhaegh GW, Schalken JA. (2010).Predicitive value of PCA3 in urinary sediments in determining clinico-pathological characteristics of prostate cancer. *Prostate* 70:10–16.

Holzbeierlein J, Lal P, La Tulippe E, Smith A, Satagopan J, et al. (2004). Gene expression analysis of human prostate carcinoma during hormonal therapy identifies androgen-responsive genes and mechanisms of therapy resistance. *Am J Pathol* 164: 217–27.

Irizarry RA, Boltstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003). Summaries of AffymetrixGeneChip Probe Level Data. *Nucl Acids Res* 31:e15.

Johanneson B, Deutsch K, McIntosh L, Friedrichsen-Karyadi DM, Janer M,Kwon EM, Iwasaki L, Hood L, Ostrander EA, Stanford JL. (2007). Suggestive genetic linkage to chromosome 11p11.2-q12.2 in hereditary prostate cancer families with primary kidney cancer. *Prostate* 67:732–42.

Karan D, Lin M, Johansson SL, Batra SK. (2003). Current status of the molecular genetics of human prostateic adenocarcinomas. *Int J Cancer* 103:285–93.

Lacroix M, Leclercq G. (2004). About *GATA3*, *HNF3A* and *XBP1*,three genes co-expressed with the oestrogen receptor-a gene (*ESR1*). in breast cancer. *Mol Cell Endocrinology* 219:1–7.

Landers KA, Burger MJ, Tebay MA, Purdie DM, Scells B, Samaratunga H, Lavin MF, Gardiner RA. (2005). Use of multiple biomarkers for a molecular diagnosis of prostate cancer. *Int J Cancer* 114:950–6.

Lapointe J, Li C, Higgins JP et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *ProcNatlAcadSci USA* 101:811–16.

La Tulippe E, Satagopan J, Smith A, Scher H, Scardino P, Reuter V, Gerald WL. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* 62:4499–506.

Li S, Bhamre S, Lapointe J, Pollack JR, Brooks JD.(2006). Application of genomic technologies to human prostate cancer. *Omics* 10:261–75.

Li J, Fine JP. (2008). ROC analysis with multiple tests and multiple classes: methodology and applications in microarray studies. *Biostatistics* 9:566–76.

Luo J, Duggan DJ, Chen Y, et al. (2001). Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 60:858–63

Magee JA, Araki T, Patil S, Ehrig T et al. (2001). Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res* 61:5692–6.

Mhawech-Fauceglia P, Zhang S, Terracciano L, Sauter G, Chadhuri A, Herrmann FR, Penetrante R. (2007). Prostate-specific membrane antigen (PSMA).Protein expression in normal and neoplastic tissues and its sensitivity and specificity in prostate adenocarcinoma. *Histopathology* 50:472–83.

Nelson PS. (2004). Predicting prostate cancer behavior using transcript profiles. *J Urol* 172:828–33.

Ogdie A, Li J, Dai L, YuX, Daiz-Torne C, Schumacher HR, Pessler F. (2010). Identification of broadly applicable tissue biomarkers of synovitis with binary and multi-category receiver operating characteristic analysis. *Biomarkers* 15:183–90.

Pal P, Kaushal R, Sun G, Jin CH, et al. (2006). Variants in the HEPSIN gene are associated with prostate cancer in men of European origin. *Human Genet* 120:187–92.

Peehl DM. (2005). Primary cell cultures as models of prostate cancer development. *EndocrinolRel Cancer* 12:19–47.

Pfaffl MW. (2001). A new mathematical model for relative quantification in real-time RT PCR. *Nucl Acid Res* 29:e45.

Rhodes DR, Barrette TR, Rubin MA et al. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 62:4427–33.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. (2007). Oncomine 3.0: genespathwaysand networks in a collection of 18 000 cancer gene expression profiles. *Neoplasia* 9:166–80.

Rose A, Xu Y, Chen Z, et al. (2005).Comparative gene and protein expression in primary cultures of epithelialcells from benign prostatic hyperplasia and prostate cancer.*Cancer Lett* 227:213–22.

Singh D, Febbo PG, Ross K, et al. (2002). Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell* 2:203–9.

Stephen C, Yousef GM, Scorilas A, Jung K, et al. (2004).Hepsin is highly over expressed in and a new candidate for a prognostic indicator in prostate cancer. *J Urol* 171:187–91.

Sterner-Kock A, Thorey IS, Koli K, Wempe F,Otte J, Bangsow T, Kuhlmeier K, Kirchner T, Jin S, Keski-Oja J, von Melchner H. (2002).Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP4). causes abnormal lung development, cardiomyopathy, and colorectal cancer. *Genes Dev* 16:2264–73.

Takahashi S, Suzuki S, Inaguma S, Ikeda Y, et al. (2002). Down-regulation of human X-box binding protein I expression correlates with tumor progression in human prostate cancers. *Prostate* 50:154–61.

Tanguay S. (2000).The role of complexed PSA and percent free PSA in prostate cancer detectionprostate update.*The Canadian Prostate Health Council* 5.

Tomlins SA, Mehra R, Rhodes DR, Cao X et al. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39:41–51.

Tusher VG, Tibishirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *ProcNatlAcadSci USA* 98:5116–21.

Vanaja DK, Cheville JC, Iturria SJ, Young CY. (2003). Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. *Cancer Res* 63:3877–82.

Varambally S, Dhanasekharan S, Zhou M, et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 419:624–9.

Welsh JB, Sapinosos LM, Su AI, et al. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 61:5974–8.

Yu YP, Landsittel D, Jing L, et al. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J ClinOncol* 22: 2790–9.