Journal of Pathology
J Pathol 2019; **249:** 411–424
Published online 16 October 2019 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.5315

ORIGINAL PAPER

Development of a predictive model for stromal content in prostate cancer samples to improve signature performance

Nadia Boufaied[1], Mandeep Takhar[2], Claire Nash[1], Nicholas Erho[2], Tarek A Bismar[3,4], Elai Davicioni[2] and Axel A Thomson[1]*

[1] *Division of Urology and Cancer Research Program, McGill University Health Centre Research Institute, Quebec, Canada*
[2] *Research and Development, GenomeDx Biosciences, Vancouver, Canada*
[3] *Department of Pathology and Laboratory Medicine, University of Calgary Cumming School of Medicine, Calgary, Canada*
[4] *Department of Oncology, Biochemistry and Molecular Biology, University of Calgary Cumming School of Medicine, Calgary, Canada*

**Correspondence to: AA Thomson, Division of Urology and Cancer Research Program, McGill University Health Centre Research Institute, 1001 Decarie Blvd, Montreal, Quebec H4A 3J1, Canada. E-mail: axelthomson@gmail.com*

Abstract

Prostate cancer is heterogeneous in both cellular composition and patient outcome, and development of biomarker signatures to distinguish indolent from aggressive tumours is a high priority. Stroma plays an important role during prostate cancer progression and undergoes histological and transcriptional changes associated with disease. However, identification and validation of stromal markers is limited by a lack of datasets with defined stromal/tumour ratio. We have developed a prostate-selective signature to estimate the stromal content in cancer samples of mixed cellular composition. We identified stromal-specific markers from transcriptomic datasets of developmental prostate mesenchyme and prostate cancer stroma. These were experimentally validated in cell lines, datasets of known stromal content, and by immunohistochemistry in tissue samples to verify stromal-specific expression. Linear models based upon six transcripts were able to infer the stromal content and estimate stromal composition in mixed tissues. The best model had a coefficient of determination R^2 of 0.67. Application of our stromal content estimation model in various prostate cancer datasets led to improved performance of stromal predictive signatures for disease progression and metastasis. The stromal content of prostate tumours varies considerably; consequently, deconvolution of stromal proportion may yield better results than tumour cell deconvolution. We suggest that adjusting expression data for cell composition will improve stromal signature performance and lead to better prognosis and stratification of men with prostate cancer.

Keywords: prostate cancer; bio-markers; tumour stroma; tissue deconvolution; bio-marker performance

Received 27 November 2018; Revised 27 May 2019; Accepted 13 June 2019

Conflict of interest statement: MT, NE and ED are employees of GenomeDX Biosciences Inc. NB, CN, TAB and AAT have no competing interests to declare.

Introduction

Biomarker discovery in prostate and other cancers has flourished following the introduction of high-throughput technologies such as microarrays, next generation sequencing and bio-informatics. It is now possible to conduct studies on large patient cohorts with comprehensive tumour information to develop markers that predict disease progression. These studies have evolved to identify multi-gene expression signatures instead of single gene biomarkers.

Until recently, cancer biomarker discovery focussed upon molecules expressed by transformed cells *in vitro* and tumour-cell expressed markers. It is now evident that tumour stroma plays a major role in cancer development and progression, and is a complementary source of novel biomarkers. Prostate stroma is a complex tissue composed of cells such as fibroblasts, endothelial cells and immune cells. It undergoes histological and molecular changes during cancer progression shown to be associated with poor outcome [1–4]. Reactive stroma expresses growth factors, chemokines, interleukins, fibroblast growth factors, matrix remodelling factors and other factors involved in growth, survival and angiogenesis to modulate tumourigenesis [5,6]. Several studies from the Mercola group showed that prostate stroma expressed specific molecules associated with cancer that distinguished indolent from aggressive types [7,8]. A classifier based on genes expressed in tumour adjacent stroma had high accuracy (97%) when tested in a cohort of 364 cases [9]. The challenge in developing reliable stromal biomarkers is the ability to de-convolute stromal-specific gene expression profiles from those of tumour cells. Conflicting reports on the contribution of different cell types within gene expression profiles have been published. In leukaemic samples, de Ridder *et al* [10] showed that to obtain reliable microarray gene expression profiles, samples must contain at least 90% of target cells. Microarray analysis of tissues with less than 75% of tumour cellularity led to 25% of erroneously identified genes. In colon cancer samples, De Bruin *et al* [11] showed that a tissue with low epithelial (15%) content contributed 50% of the gene expression profile because the RNA yield of epithelial cells was higher than stroma. In a study across 21 different cancers, Aran *et al* showed that tumour purity was a confounder in genomic analysis. In three bio-informatic analyses routinely applied to cancer studies (correlation analysis, clustering and differential expression), the results were highly obscured by tumour purity because they correlated with tumour purity rather than cancer features [12].

Tissue composition estimated by pathologists on H&E-stained slides can be inaccurate [13]. Conversely, micro-dissection and single cell-based techniques are too cumbersome to be implemented in large cohorts. The alternative is to develop methods to de-convolute transcriptomic data using cell specific markers and estimate their proportions within patient samples. Numerous computational based methods have been developed to extract cell-type specific information from complex tissues or to estimate cell-type proportion; but they are rarely applied in transcriptional or genomic studies. Five classes of computational approaches exist based on the input data required and the type of feature generated. Some methods combine expression profiling of heterogeneous tissues and cell proportion data [14–18], others require either a signature or specific markers of each cell population [19–26] while some methods rely little on proportion or expression profile [27–30]. Most of these methods were developed using haematopoietic malignancies and studies in solid tumours have focussed on tumour (epithelial) cell content estimation. Stroma and other tissue components have largely been neglected. ESTIMATE was devised to calculate tumour purity using stromal and immune signatures [22], while MCP-counter enables quantification of eight immune cell types and two stromal cells (fibroblast and endothelial) in tissues [31]. Among these methods, two have been created and validated in prostate cancer datasets. CellPred is a microarray based de-convolution algorithm that evaluates tumour and stroma content in mixed samples [32]. ISOpure is a statistical method which uses expression profiles from healthy tissues to predict the likely proportion of tumour and normal cells in samples [27,33]. However, a method to quantify stromal content in prostate transcriptomic datasets such as RNAseq data is currently lacking.

We have identified transcripts with stromal expression in the prostate and developed a model to infer stromal contribution within tumour samples. We defined 17 transcripts able to distinguish stromal from epithelial cells that were specific to prostate cancer. We experimentally validated the stromal specificity of these transcripts and used a subset to derive a linear model for estimating stromal content. We used our model to calculate the stromal content of two independent datasets (TCGA and University of Calgary) and demonstrated reliable stromal quantitation of equal or better performance than current models. Finally, we showed that stromal based classifiers performed better when tested in datasets adjusted for stromal content.

Materials and methods

Ethical approval and consent to participate

The expression profiles retrospective patients were extracted from the Decipher GRID registry (NCT02609269). Tissue samples were collected from archival samples processed at the Department of Pathology-Calgary Laboratory Services. All Clinical and pathological data were obtained with approval of the institutional review board at University of Calgary, Cumming School of Medicine. Calgary, Alberta, Canada.

Selection of stromal control transcripts

To derive a prostate stromal quantitation signature, we first selected stromal-specific transcripts from published gene expression datasets and compared this to prostate mesenchyme transcriptional profiling data followed by a series of filtering criteria detailed in Figure 1.

Gene expression profile datasets

Datasets were retrieved from Gene Expression Omnibus (GEO). Datasets with micro-dissected prostate tissues: GSE26910 [34] contains 24 micro-dissected normal and reactive stroma samples from prostate and breast analysed with Affymetrix Human Genome U133 plus 2.0 array. GSE20758 [35] includes paired stroma and epithelial micro-dissected prostate tissues analysed using Affymetrix U133 2A array. GSE6099 [36] includes micro-dissected stroma and epithelial prostate tissues from various origin (normal, BPH, HGPIN, localised cancer and metastasis). Epithelial samples were further subcategorised into normal epithelium adjacent to prostate cancer (PCa), atrophic epithelium, normal epithelium and BPH epithelium. Stromal samples were subcategorised to stroma adjacent to PCa, normal stroma and BPH stroma. Expression profiling was conducted using the Chinnaiyan Human 20 K Hs6 array. We used datasets with pathologist's estimation of tissue composition: GSE8218 [32] which is composed of 136 samples; 65 samples with high tumour cellularity and 71 tumour samples micro-dissected to obtain tumour-adjacent stroma processed on the Affymetrix Human Genome U133 plus 2.0 array. This dataset has pathologist estimation of four different tissue components (tumour, stroma, BPH and atrophic glands). GSE1431 [15] containing 88 prostate tissue samples

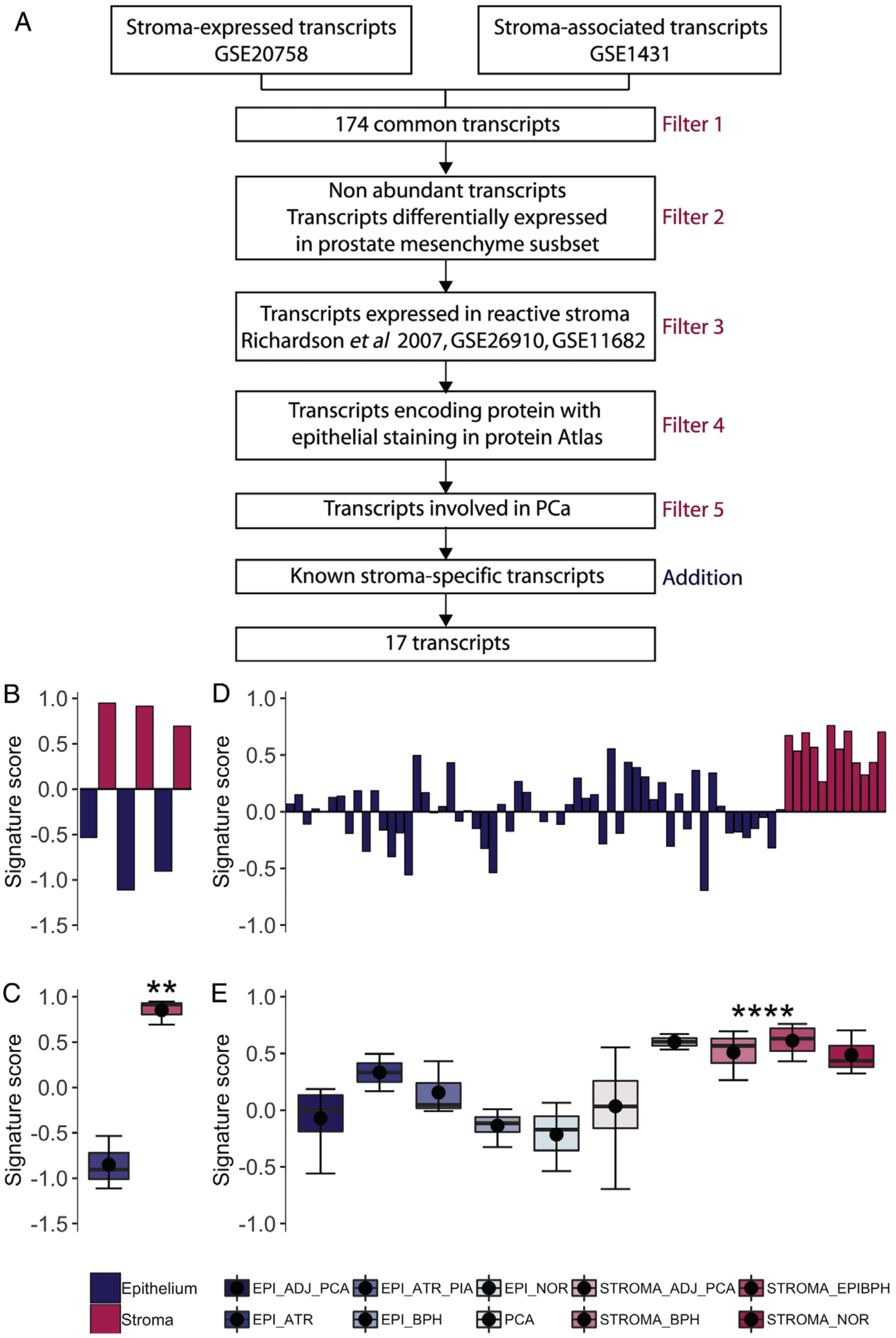

Figure 1. Selection of stromal markers. (A) Schematic of the criteria used to select transcripts for stromal content estimation. These were subsequently examined in micro-dissected tissue datasets, to compare expression in stromal versus epithelial cells. (B) Normalised expression of each of the 17 transcripts was centred and scaled and then used to create a signature score (average) in GSE20758 [35]. Bar plot showing signature score for each micro-dissected sample. (C) Box plot showing the range of signature score in stromal samples versus epithelial samples; the score was higher in stromal samples (*P* value = 3.42e−3). (D) Normalised expression of each of 13 transcripts present on the array was centred and scaled and then used to create a signature score (average) in GSE6099 [36]. Bar plot showing signature score for 12 stroma micro-dissected samples and 59 micro-dissected epithelial samples. (E) Signature scores range in different cell population as identified in Tomlins *et al* [36]. The signature score of stromal samples was significantly higher than epithelial samples (*P* value = 8.71e−10). Black circle represents the signature score mean.

scored for tumour, stroma and BPH content; profiling was conducted using the Affymetrix U95Av2 array. Other prostate datasets used in this analysis included: GSE11682 [1], Richardson *et al* [37], GSE21031 [38] and GSE46691 [39]. Additionally, we used ovarian dataset GSE38666 [40] composed of eight normal stroma and adjacent epithelia plus seven tumour stroma and adjacent tumour tissues. Gene expression analysis was conducted using the Affymetrix Human Genome U133 plus 2.0 array. University of Calgary stromal dilution cohort contained 39 samples obtained from six unique patients. Expression profiles were extracted from the Decipher GRID registry (NCT02609269). Institutional Review Board approval was obtained from the participating institution prior to initiating the current study. RNA extraction and microarray hybridisation to Affymetrix Human Exon 1.0ST arrays (Affymetrix, Santa Clara, CA, USA) for the University of Calgary cohort were performed in a Clinical Laboratory Improvement Amendments-certified laboratory (GenomeDx Biosciences, Vancouver, Canada) and has been described previously [39,41,42].

Microarray data processing

Expression levels from Affymetrix array based studies were normalised using SCAN [43], batch corrected using ComBat (SVA package) [44], probes were mapped to unique transcripts with biomaRt [45]. Multiple probes for the same gene were collapsed by using collapseRows function of WGCNA [46–48]. Relative gene expression was calculated by subtracting from the gene estimate the mean expression value across all patients in the dataset, and then dividing it by its standard deviation across all patients (z-score). Signature score was calculated by averaging the relative expression of the genes in the signature.

Tissue cell-type composition analysis

Three different methods were used to assess tissue composition. ESTIMATE stroma scores and ESTIMATE tumour purity were calculated using the ESTIMATE R package [4]. Tumour purity was calculated using the IsoPure R package [33] and tumour and stromal percentage were estimated with CellPred (http://www.webarraydb.org) [32].

TCGA dataset

TCGA expression data and patient clinicopathological data were downloaded with TCGAbiolinks R package (Bioconductor; http://www.bioconductor.org) [49] and H&E slides were downloaded from the NCI GDC legacy Archive. Estimation of stromal area was performed using ImageScope viewer and analysis software (Aperio Technology Inc., Vista, CA, USA). The Aperio Positive Pixel count algorithm was used to calculate tissue area (*Atotal*). A new algorithm (stroma Algorithm) was created to calculate the stromal area: the Aperio Positive Pixel count algorithm input parameters were set to obtain the identification of pixels related to the stroma as weak positive (*Nwp* in yellow) and tuned to make non-specific pixels (in blue) define epithelia. Stroma was then defined by Atotal × Nwp/Ntotal. Stromal percentage was calculated as stromal area/tissue area × 100 (see supplementary material, Figure S1). To validate the accuracy of software estimation, stromal percentage was reviewed by a pathologist (blinded to the predicted stromal content). Only validated cases were retained for analysis [50] and we excluded slides with tissue not properly spread ($n = 111$). The stromal percentage was averaged for samples with two slides ($n = 84$). TCGA RNAseq data (HTSeq-count) for prostate cancer were downloaded using TCGAbiolinks R package [49]. Genes with low read counts were removed. Read counts were normalised by library size with the 'DESeq2' package [51] and voom transformed using the limma R package [52].

Cell culture

Normal prostate fibroblast cells, PrSC (Lonza), cancer-associated fibroblasts (CAF) [53], WPMY-1 [54], BHPrS [55] normal prostate epithelial cells, RWPE-1 (ATCC), BPH cells, BPH-1 [56] and prostate cancer epithelial cell lines, PC-3, LNCaP, DU-145 (ATCC) were maintained in DMEM (Wisent, Quebec, Canada) supplemented with 10% FBS. Six different cell mixtures of CAF or BHPrS and PC-3 were prepared (100, 70, 50, 30 and 0% fibroblasts). The concentration of each cell line was measured by haemocytometer and cell lines were mixed at the desired ratio.

RNA extraction and quantitative PCR

RNA from cell lines and cell mixtures was extracted using Trizol (Invitrogen, Carlsbad, CA, USA). One microgram of RNA was reverse transcribed using iScript™ (Bio-Rad, Hercules, CA, USA) and qPCR was performed using SYBR Green Real time Master Mix (Bio-Rad, Hercules, CA, USA). Cq-values were determined using the iQ5 software (Bio-Rad, Hercules, CA, USA), gene expression was normalised with the geometric mean of three reference transcripts (*GAPDH*, *TPB* and *RNA18S*).

Immunohistochemistry

Antibodies were chosen using ProteinAtlas (www.proteinatlas.org) and selected for those with stromal-specific staining. Formalin-fixed prostate cancer tissues were stained with C1S (NBP1-86439) diluted 1:50, RABGAP1L (H00009910-M05) diluted 1:600, RPBMS (NBP2-33810) (Novus Biologicals, Littleton, CO, USA) diluted 1:40 and VIM (Sigma-Aldrich, St. Louis, MO, USA; HPA001762) diluted 1:1000. After deparaffinising and rehydration, tissue sections were antigen-retrieved with citrate buffer pH 9 in a pressure cooker for 5 min. Slides were then treated with 0.5% H_2O_2 for 30 min to quench peroxidase activity and blocked for 4 h with IHC select (Millipore, Billerica, MA, USA). Antibodies were hybridised overnight at 4 °C. Primary antibodies were omitted to serve as negative controls, and staining patterns were consistent with those observed in ProteinAtlas.

Stromal model

We used linear regression (LM) to predict stromal content of a tissue from its transcriptome profile. GSE8218 [32], where the stromal content was defined, was randomly divided into a training ($n = 88$) and a testing dataset ($n = 44$) and used for training and validating the algorithm. The algorithm performance was also measured in two independent datasets: TCGA and Calgary stroma dilution cohort. Twenty-three features were

Table 1. Description of the new features created using expression levels of the 17 transcripts

Features	Summary
PIMS	Mean *C1S*, *CALD1*, *FHL1*, *MYLK*, *RBPMS*
PIMS2	Mean *CALD1*, *FHL1*, *METTL7A*, *RBPMS*
CSC	Sum *C1S*, *CALD1*, *SVIL*
RP	Sum *PTPRD*, *RABGAP1L*
BC	Sum *BTG3*, *COL4A2*
CM	Sum *C1S*, *METTL7A*

included as putative predicators for stromal content model. In addition to the 17 transcript expression level, we created six new features based on expression levels as described in Table 1.

We built the LM model using the mlr package in R [57] with the 23 predicators, and then a chi-squared feature selection method was applied with generateFilterValuesData function of the FSelector package in R [58]. The model was evaluated by computing the coefficient of determination (R^2), the root mean squared error (RMSE) of prediction and Spearman correlation. The stromal percentage generated by the model was used to adjust gene expression in stroma portion of a given tissue as follows; gene expression × stroma percentage/100.

Statistical analysis

All statistical analysis was performed using R and the Bioconductor suite (http://www.r-project.org). A two-tailed Student's *t*-test was used to evaluate differences between two groups. For experiments involving multiple comparisons, we performed one-way ANOVA with the Tukey *post hoc* test to evaluate differences. The Spearman and Pearson pairwise correlation method was used for all correlations analysis (car R package) [59]. The cor.test function from the R Stats package was used to calculate *P* values for each correlation. Hierarchical clustering was performed on either row-normalised gene expression or on the stromal normalised data using Cluster [60] and factoextra [61]. Tumours were assigned to different clusters based on the stromal classifier described previously [9] or stroma derived metastasis signature (SDMS) [62]. Heatmaps of the stroma classifier genes were produced using gene expression and stromal normalised gene expression using the R package ComplexHeatmap [63]. Kaplan–Meier survival analysis and log-rank tests were used to analyse recurrence-free survival using R package survival [64] and receiver operating characteristic (ROC) curves were used to compare the sensitivity and specificity of the risk prediction for recurrence-free survival using R package plotROC [65].

Results

Identification of transcripts with stromal specificity

We selected a list of 174 stromal-specific transcripts by combining stromal-associated transcripts identified in two previous studies [15,35]. We then identified the best candidates among this list using several criteria (Figure 1A). Transcripts with low Tag count (TPM) in our prostate fibroblast Tag profiling study [66] were removed to ensure transcripts could be easily detected and reliably quantitated. In order to define stromal-specific expression and expression across cell subsets, we used transcriptome data of prostate mesenchymal subsets (VMP and SU) [67,68] to select transcripts ubiquitously expressed in mesenchyme. Transcripts shown to be dysregulated in prostate cancer, or differentially expressed between reactive and normal stroma were removed [1,34,37]. Transcripts encoding proteins with epithelial localisation by immunohistochemistry in the Human Protein Atlas (www.proteinatlas.org) [69] were removed. We included transcripts previously identified as expressed in prostate fibroblasts by our group [53,66]. These criteria yielded 17 stromal-specific transcripts derived from prostate stroma that were used for further analysis (Figure 1A).

To determine whether the 17 transcripts could distinguish stromal versus epithelial tissue, we generated a signature score by averaging their expression in micro-dissected prostate samples that separated stroma from epithelia (GSE20758). The stromal score in the micro-dissected stroma enriched fraction was high, but was low in the epithelial fraction suggesting that these 17 transcripts could distinguish stromal versus epithelial tissues (Figure 1B,D). To confirm the utility of the signature score to distinguish the two tissue compartments, we used an independent dataset (GSE6099) that was not used to identify the stroma transcripts initially (Figure 1C,E). This gene expression dataset was generated using an in-house microarray platform; four of our stromal transcripts were absent from the microarray. The signature score was calculated by averaging the remaining 13 transcripts. Overall, there was a significant difference between the signature score in micro-dissected stromal samples compared to tumour cells. All micro-dissected samples ($n = 12$) showed a positive signature score while signature scores for epithelial samples had a wide range of values (negative and positive values). Normal epithelium and epithelium from BPH (EPI_NOR, EPI-BPH) had the lowest signature scores whereas epithelium from atrophic lesions had the highest signature score (EPI_ATR, EPI_ATR_PIA).

Prostate specificity of the stromal signature

Previous studies have shown that prostate stromal tissue differs from stroma of other organs, and it is documented that reactive stroma has a distinct gene expression profile compared to normal stroma [1,34,70]. We selected transcripts that were expressed ubiquitously in the stroma, and not differentially expressed between normal and reactive stroma. To verify whether our transcripts met these criteria, we compared the signature score of micro-dissected normal prostate stroma to reactive prostate stroma (GSE26910) [34] (Figure 2A,B).

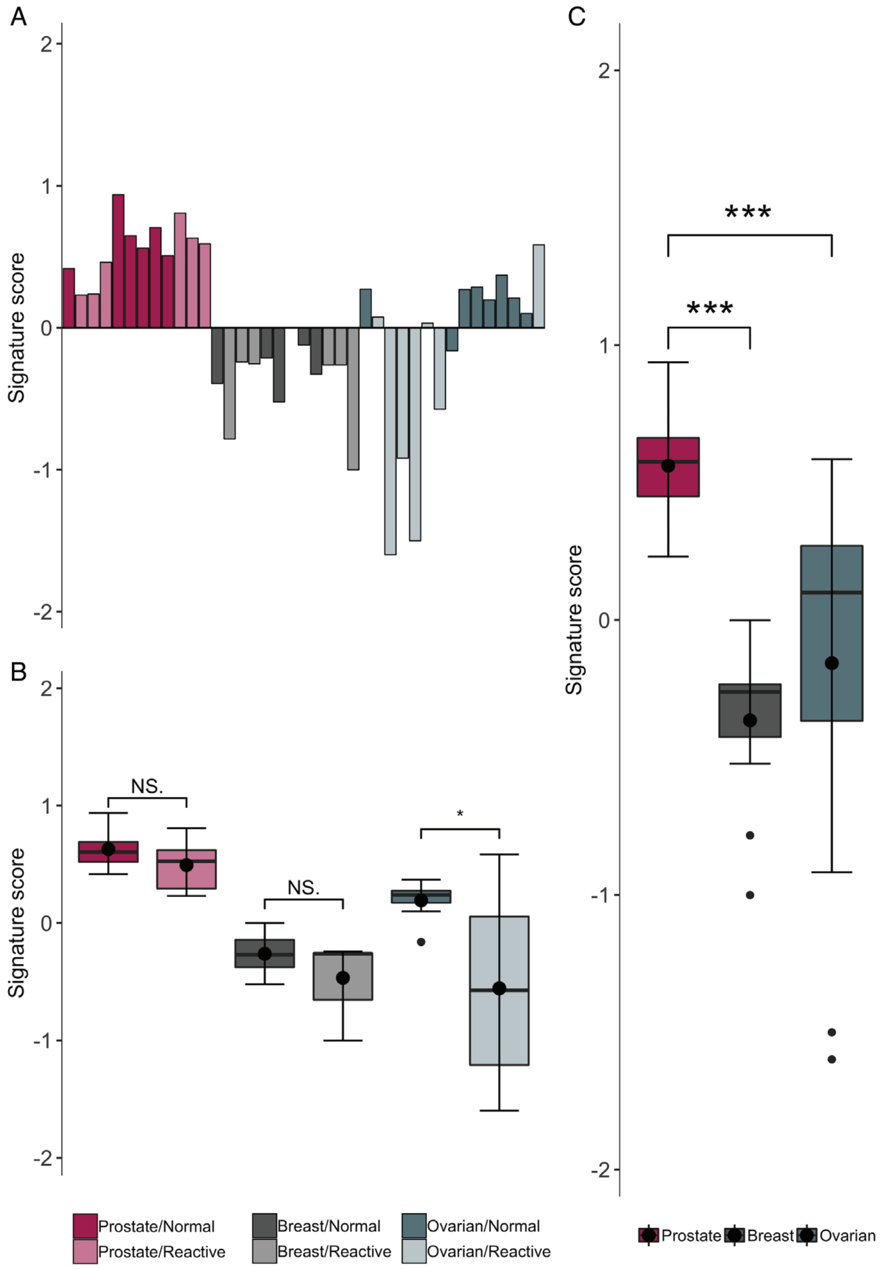

Figure 2. Comparison of stromal signature score between normal and reactive stroma in prostate, breast and ovarian tissue. Signature score was calculated for micro-dissected normal and reactive stroma for prostate and breast tissues using GSE26910 [34] and for ovarian tissue using GSE38666 [40]. (A) Prostate stroma samples had a positive signature score; breast stroma samples had a negative signature score and ovarian stroma had a mixed signature. (B) Signature score for normal and reactive stroma are similar in prostate and breast tissues but significantly different in ovarian tissue (*P* value = 1.41e–02). (C) Signature score for prostate stroma is significantly different from breast signature score (*P* value = 6.40e–05) and ovarian signature score (*P* value = 8.98e–04). NS, not significant.

The signature score for normal stroma micro-dissected samples was similar to the signature score obtained for reactive micro-dissected stroma. We observed similar results in breast tissue samples, however, the signature score generated in ovarian micro-dissected normal stroma samples was significantly higher than signature score of reactive stroma samples in GSE38666 [40] (Figure 2B). When comparing the prostate signature score, we observed a significant difference between prostate signature score and breast or ovarian signature score suggesting that these transcripts showed specificity for prostate stroma (Figure 2C).

Experimental validation of stromal transcript expression

To assess the ability of these stromal genes to infer the stromal content, we created a mixture of tumour cells (PC-3) and normal fibroblasts (BHPrS) and a mixture of tumour cells (PC-3) and cancer fibroblasts (CAF) in varying proportions ranging from 0 to 100% and measured the expression of a subset of our stromal transcripts by RT-qPCR. *C1S*, *FHL1*, *MYLK*, *RBPMS* and *VIM* showed a decrease in gene expression correlating with reduced fibroblast proportion (see supplementary material, Figure S1A,B). To verify stromal expression

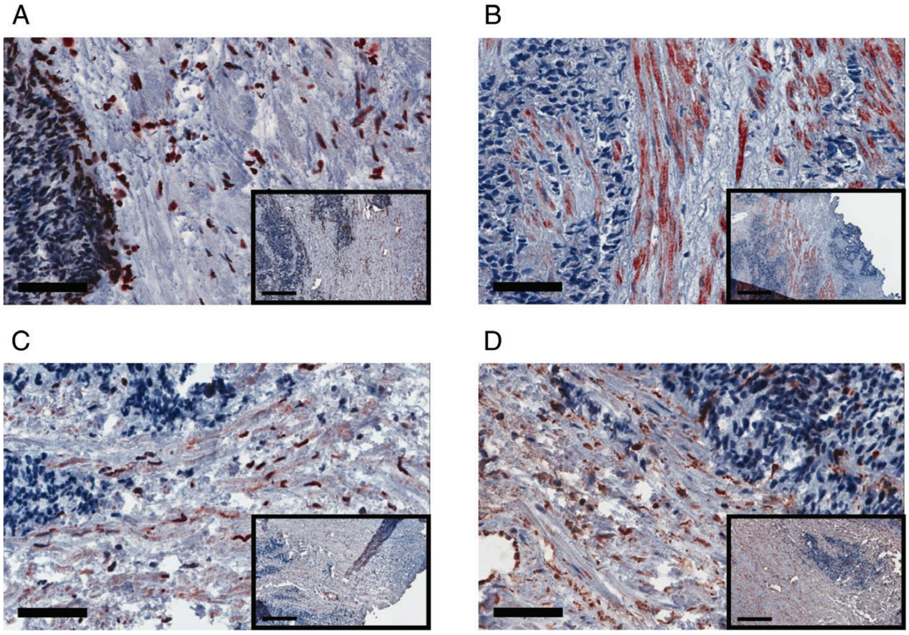

Figure 3. Expression of selected stromal markers in prostate tissue. Immunohistochemistry of RABGAP1L (A), C1S (B), RBPMS (C) and VIM (D) in prostate cancer tissue. RABGAP1L and RBPMS showed nuclear expression in prostate stroma while C1S and VIM showed cytoplasmic expression in prostate stroma. Main image ×40 magnification; scale bars, 200 μm; inset ×10 magnification; scale bars, 60 μm.

of selected transcripts via protein localisation, we performed immunohistochemistry for four of the 17 transcripts on prostate cancer tissues. As shown in Figure 3, RABGAP1L and RBPMS were expressed in the nuclei of stromal cells while C1S and VIM were localised to stromal cytoplasm.

Model construction for stromal quantitation

To build a model for stromal content estimation, a gene expression dataset with stromal percentage defined by a pathologist was divided into testing and training datasets [9]. We used each of the stromal-specific 17 transcripts individually and created new features by combining subset of transcripts. This included summing or averaging the relative expression of the transcripts as described in the section ‘Material and methods’ to give a total of six features. To reduce the number predicator and select the most useful for predicting stromal content using the testing dataset, we performed a chi-squared test and selected four most predictive according to their chi-squared values (Figure 4A). Then, we constructed a linear model (see supplementary material, Figure S2). Graphical representation of model coefficients is presented in Figure 4B. Predicted and observed stromal content relationship in testing and training dataset is presented in Figure 4C,D. The stromal estimation model had an R^2 of 0.46 and Spearman correlation of 0.71 in the training dataset, while in the testing dataset an R^2 of 0.67 and a Spearman correlation of 0.79 were observed. In the testing dataset, the expression level of the four predictive transcripts showed high correlation (Pearson’s $R^2 > 0.5$) with stromal composition (see supplementary material, Figure S3). We have successfully combined a minimal number of stromal transcripts and created a linear model to estimate tissue stromal content. This model performed as well as the CellPred 250-gene model [32] (see supplementary material, Figure S4).

Model evaluation and validation in TCGA and GenomeDx datasets

In order to evaluate the performance of the stromal quantitation model using independent datasets, we estimated the stromal proportion within TCGA samples by image analysis. Prostate cancer histologic H&E stained sections from the TCGA Network were analysed using an Aperio stromal algorithm (see supplementary material, Figure S5). The stromal content of 308 samples included in the analysis ranged between 5 and 97%. The dataset showed a higher inclusion of normal tissues compared with cancer tissues but showed no statistically significant difference in Gleason scores (see supplementary material, Figure S5D). The stromal content predicted by the stromal quantitation model showed poor correlation with the image-based estimation of stromal area (Spearman’s $R^2 = 0.42$; Figure 5A). However, a better correlation was observed between the stromal quantitation model and the stromal score of the ESTIMATE algorithm (Spearman’s $R^2 = 0.668$; Figure 5B). We also calculated the correlation between tumour content (tumour purity) estimates by ISOpure and our stromal quantitation and observed a highly inverse correlation ($R^2 - 0.778$; Figure 5C), as expected. This suggested that transcriptomic-based deconvolution methods show greater concordance among different algorithms versus histopathology-based estimates. To investigate the divergence in model performance in TCGA testing and training dataset, we examined the correlation between histological-based tissue composition

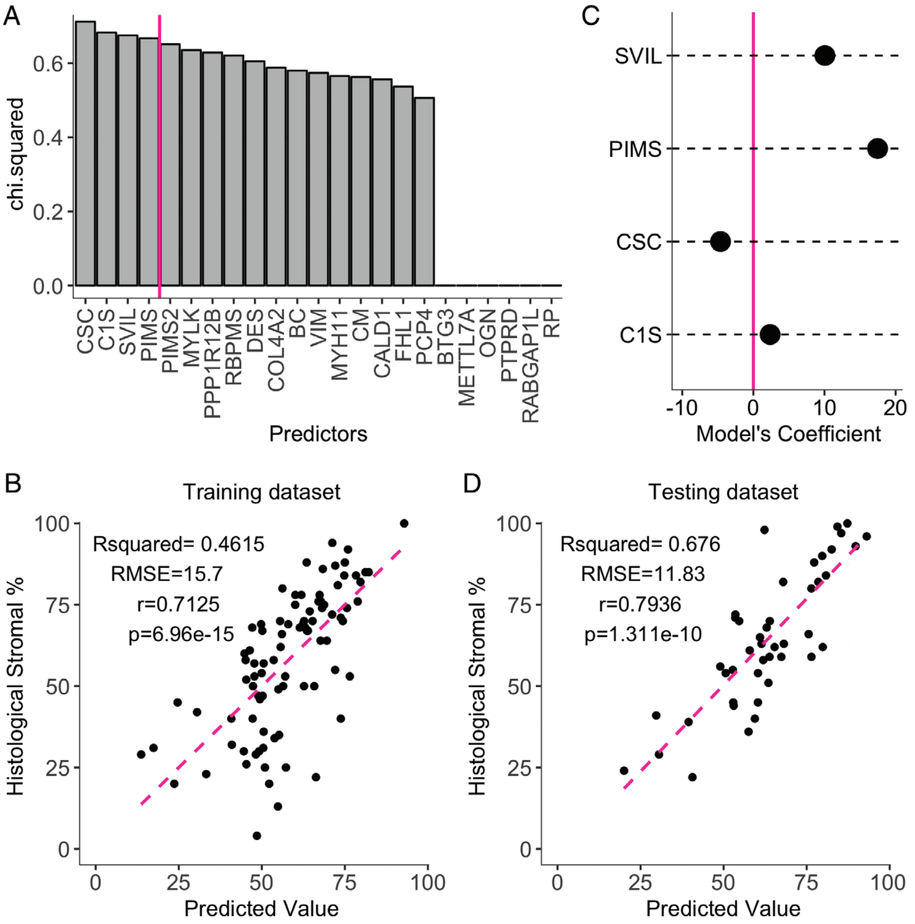

Figure 4. Development of a model to predict stromal content. (A) Importance of predicator variables. Graph showing the chi-squared value calculated for the 23 features selected to use for modelling. The four predicators with the highest chi-squared were retained. (B) Graphical representation of our stromal model. The graph shows the coefficients of the four predicators included in the model: *SVIL*, *C1S* and PIMS (mean of *FHL1*, *MYLK*, *RBPMS*, *C1S* and *CALD1*) and CSC (sum of *C1S*, *SVIL* and *CALD1*). (C,D). Stromal model with four predicators. Panels C and D show the predicted stromal content versus the observed one in both training dataset (R^2 = 0.4614 and Spearman = 0.7125) and testing dataset (R^2 = 0.676 and Spearman = 0.7936).

and transcriptomic-based method (CellPred, ESTIMATE and ISOpure) in TCGA and GSE8218. As previously, shown, the two types of methods had low correlation in TCGA [12]. Interestingly, in GSE8218, the correlation coefficients between all methods were higher than in TCGA (see supplementary material, Figure S6). Taken together, our organ-selective stromal signature performed better than existing models in predicting stromal content.

To test the performance of our stromal quantitation model, we applied it to a Calgary cohort containing samples from six patients with known stromal content ranging from 5 to 60%; each sample was diluted to produce standard curves of stromal percentage for each patient. The correlation between the stromal signature score and stromal histological-based method was low (Spearman's $R^2 = 0.31$, Pearson's R^2 with adjustment for patient = 0.48) while the correlation with ESTIMATE was considerably higher (Spearman's $R^2 = 0.76$) (Figure 5D,E). Next, we examined correlation between our stromal signature score and stromal dilution in each patient of the cohort individually and found a high association between them (see supplementary material, Figure S7). This showed that our signature could detect a wide range of stromal content.

Effect of 'stromalisation' on prognostic performance of a stromal classifier

To measure the effect of stromal normalisation on the performance of a stroma-based prognostic classifier, we assessed the performance of a 15 gene classifier developed by Jia *et al* [9], using the Taylor dataset (GSE21032) [38] with or without stromal quantitation (stromalisation). We used our stromal quantitation model to define the stromal content of 160 samples in GSE21032. The stromal percentage ranged from 2 to 100% (see supplementary material, Figure S8A). Then, we used the predicted stromal content to adjust gene expression values. Hierarchical clustering using the 15 genes included in the classifier divided the tumour samples into three groups (see supplementary material, Figure S8B–E). Heatmaps of the markers in non-stromalised data showed modest differences between the three patient clusters (Figure 6A), but differences were clear after stromalisation of the data

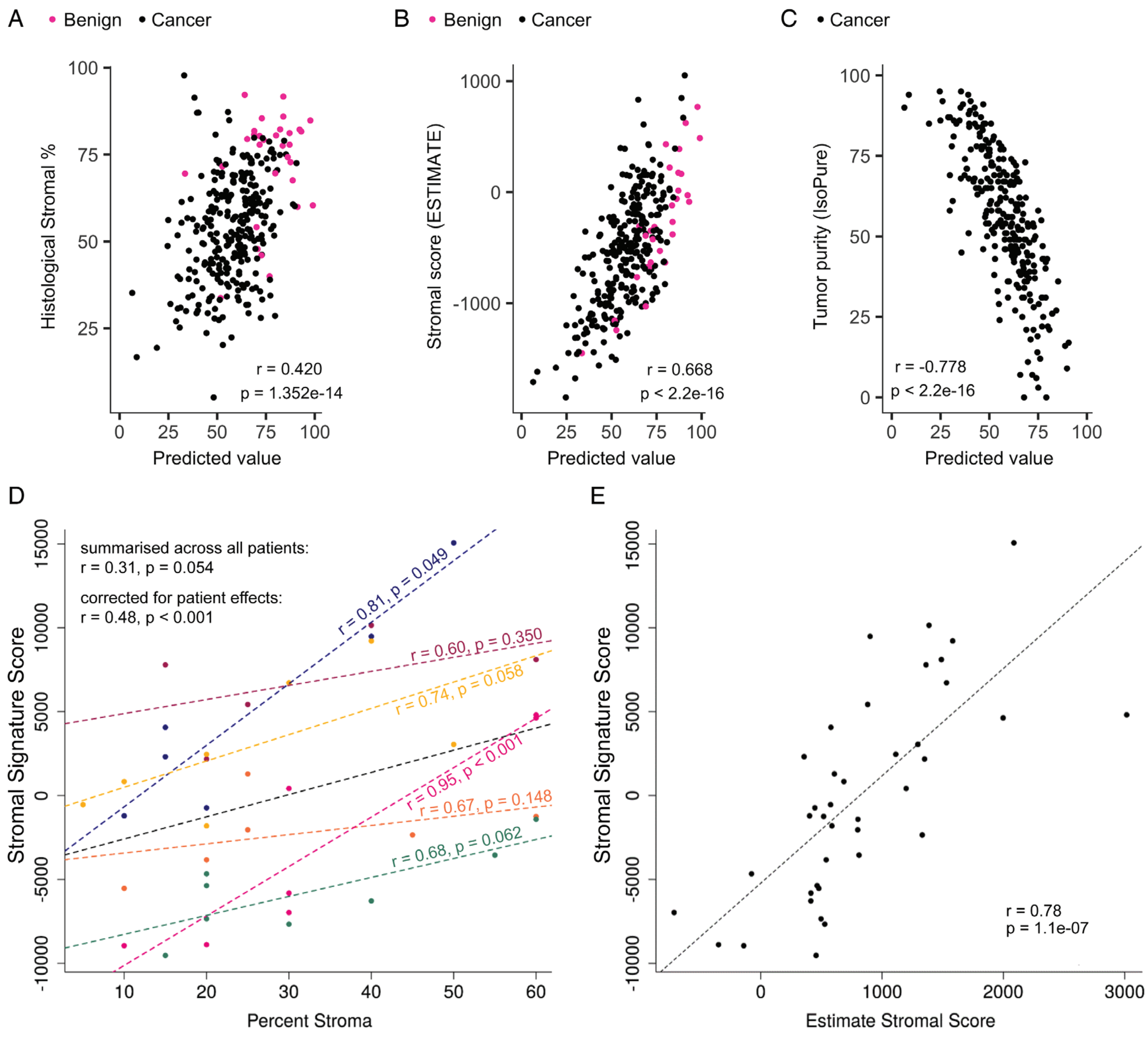

Figure 5. Stromal model validation in TCGA (A–C) and Calgary cohorts (D,E). (A) Scatterplot of predicted stromal content in TCGA samples versus stromal content measured by digital image analysis (histological stromal %) ($r = 0.420$, P value = 1.352e−14). Only validated TCGA samples (308) were used in the analysis. (B) Correlation of stromal predicted values with ESTIMATE stromal score ($r = 0.668$, P value <2.2e−15). (C) Correlation of stromal predicted values with tumour purity calculated with IsoPure ($r = -0.778$, P value = 2.2e−16). (D) Scatter plot of stromal signature score in Calgary cohort versus stromal dilution ($r = 0.48$, P value <0.001 following correction for patient effects). (E) Correlation between stromal signature score and ESTIMATE stromal score ($r = 0.78$, P value = 1.1e−07).

(Figure 6B). Survival analysis did not show a prognostic value for the signature ($p = 0.25$) when using raw tumour data (Figure 6C). However, when hierarchical clustering was applied to data that was stromalised, three groups were observed, with a clear association with biochemical recurrence (BCR) ($p = 0.0047$) (Figure 6D). Individuals in group 2 and 3 showed a very high hazard ratio of 3.03 and 4.15 (95% CI = 1.241–7.397; 95% CI = 1.543–11.189) (Table 2) for the stromal signature [7]. The association remained significant in multivariate analysis (Table 3); hazards ratio (HR) 2.66 and 4.152 (95% CI = 1.066–6.641; 95% CI = 1.505–11.45). This highlights the importance of quantitating the stromal contribution within samples to maximise the performance of stromal prognostic signatures. Using the same strategy, we assessed the performance of the 93-gene stromal derived metastasis signature SDMS [62], in GSE21034 [38] and in GSE 46691 [39]. Stromalisation showed a modest (approximately 10%) improvement of signature performance (see supplementary material, Figures S9 and S10).

Discussion

It has been shown that the stromal microenvironment plays an active role during prostate cancer development and progression. The dynamic interactions between stromal and epithelial compartments are involved in tumour growth, metastasis and patient outcome. To better predict prostate cancer progression, stromal-based prognostic signatures have been

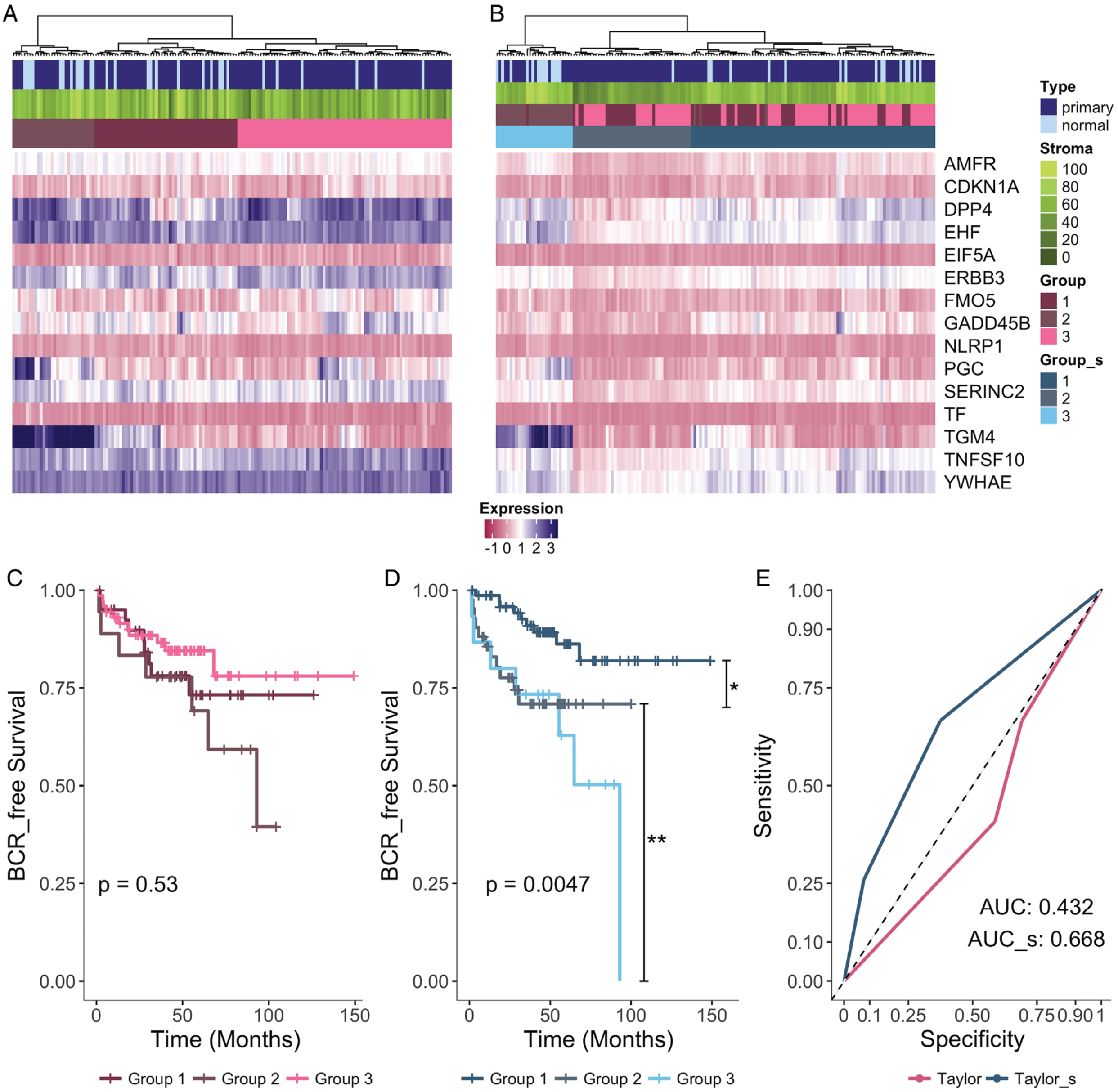

Figure 6. The effect of adjustment for stromal content upon the performance of a 15 gene stromal classifier in the Taylor dataset. Hierarchical clustering heatmap based on expression values of genes in the Jia *et al* [9] stromal classifier using the transcriptional profiles from GSE21034 [38] prior to 'stromalisation' (A) and post 'stromalisation' (B). Three groups of patients were identified, blue and purple indicate high and low expression levels, respectively. (C,D) Kaplan–Meier survival analysis of relapse-free survival according to the Jia stromal classifier of patients in the different groups. Groups were defined using expression data prior to 'stromalisation' showed no difference in progression (C). (D) Groups defined using expression data post 'stromalisation' show differential progression rates with a *P* value = 0.0047. The differences between the curves were assessed by the two-sided log-rank test. Overall *P* values are shown. * denotes *P* value < 0.05; ** denotes *P* value <0.01. (E) ROC curve analysis comparing the sensitivity and specificity of the predictive group defined prior and post stromalisation. Areas under ROC curve (AUCs) were 0.432 prior to stromalisation and 0.668 after stromalisation.

developed. A significant limitation is the variable contribution of stroma within tumour samples, which is a confounding factor reducing stromal-specific signature performance. In this study, we identified prostate stromal-specific transcripts that we used to infer stromal composition. We constructed a model to predict the stromal composition of prostate tissue based on mRNA expression of six stromal-specific transcripts and validated its performance in TCGA and University of Calgary patient cohorts.

We have previously showed that subsets within prostate mesenchyme can be a useful source of stromal molecules involved in both development and prostate cancer progression [66,67]. We identified 17 stromal-specific transcripts by combining data from micro-dissected prostate tissue and from our studies of mesenchyme during prostate development. These 17 transcripts distinguished prostate stroma from epithelia, and many showed stromal specificity at both RNA and protein level. Although prostate cancer stroma and fibroblasts are heterogeneous, our signature score in the stromal compartment was less variable than in the epithelial compartment. It was not significantly different between normal and reactive prostate stroma indicating

Table 2. Univariate analysis for recurrence-free survival (BCR) in the Taylor dataset (GSE21032) (left), and with 'stromalisation' (right) using the Jia stromal signature

					Stromalised			
	HR	95% CI	*P* value	Adj. *P* value	HR	95% CI	*P* value	Adj. *P* value
Group 2	0.640	0.579–4.208	0.379	0.4336	3.03	1.241–7.397	0.0149	0.0148
Group 3	1.422	0.291–1.698	0.434	0.433	4.154	1.543–11.18	0.00482	0.00964

Taylor dataset (GSE21032) [38].

Table 3. Multivariate analysis for recurrence-free survival (BCR) in the Taylor dataset (GSE21032) (left), and with 'stromalisation' (right) using the Jia stromal signature

					Stromalised			
	HR	95% CI	*P* value	Adj. *P* value	HR	95% CI	*P* value	Adj. *P* value
Group 2	1.441	0.529–3.921	0.474	0.697	2.66	1.066–6.641	0.0361	0.051
Group 3	1.054	0.43–2.579	0.908	0.908	4.152	1.505–11.45	0.006	0.015
PSA	1.342	0.501–3.594	0.558	0.697	1.287	0.476–3.478	0.618	0.618
Gleason	4.993	1.146–21.75	0.032	0.08	4.537	1.064–19.36	0.041	0.051
Stage	0.275	0.123–0.613	0.0016	0.008	0.261	0.117–0.585	0.0011	0.0055

PSA, prostate specific antigen.
Taylor dataset (GSE21032) [38].

that we identified stromal transcripts with stable expression independent of disease or pathology. Our stromal signature distinguished prostate from breast and ovarian stroma, two other hormone-regulated cancers, supporting the concept that prostate stroma is distinct from stroma of other organs. These results also suggest that organ-specific stromal deconvolution signatures may perform better than generic stromal signatures designed for tissue deconvolution in multiple tumour types.

We focussed on identifying the optimal prostate stromal-specific transcripts to build a reliable model for stromal quantitation and were able to improve the correlation coefficient between the predicted stromal composition and pathologist estimation in two datasets. In the testing and validation dataset, compared to the Wang 5-gene model ($R^2 = 0.38$) we observed an improved correlation coefficient of 0.7. Even though the model performed modestly in TCGA data ($R^2 = 0.42$), it was almost double that of ESTIMATE ($R^2 = 0.23$). Additionally, we observed better concordance between visual stromal quantitation and our transcript-based model in the Calgary stromal dilution cohort, albeit with a small sample size. In general, we have observed better correlation of our signature with computationally estimated cell proportions than visual estimation. The discrepancy between these two methods is common and has been attributed to error in pathologist estimation and use of different tissue for histology and RNA extraction. The tissue surface might not accurately represent the full tissue composition of a core, especially when the core extends far from the histological section, which leads to better correlations among transcript based methods versus poor correlation between transcript based methods and visual estimation.

We suggest that there is considerable need for transcript expression data derived from samples of known stromal composition. GSE8218 and Calgary cohorts were developed specifically to study tissue composition and when we closely examined the correlation between computational and pathologist estimation methods, we observed a higher correlation than in our stromally defined subset of TCGA data.

In most existing datasets, samples are included for gene expression profiling only if they contain at least 60% tumour cells; this is sub-optimal for stromal biomarker studies. Jia *et al* [9] developed a 19 probe-based classifier (17 genes expressed preferentially in the microenvironment), that predicted risk with high accuracy (87%). However, the classifier only worked in a dataset enriched in stroma and the performance decreased when applied to datasets with epithelial content greater than 10%. Using our stromal quantitation model, we observed that the signature derived by Jia *et al* could work very well in samples with low stromal content, after adjustment of expression values.

Our model is very simple, and only requires gene expression data of six transcripts; thus it could be used to estimate stroma content of old microarray datasets with limited number of probes as well as in RNAseq datasets that cover the whole genome. It is also easily implemented in Nanostring based measurement of transcript expression where limited number of transcripts can be assessed. The implementation of stromal deconvolution will lead to changes in biomarker discovery, and will support the identification of markers that change as a result of gene regulation rather than changes in cell proportions.

Conclusions

Biomarkers are keys for distinguishing indolent from aggressive prostate cancer, and to stratify patients among different treatment options. Patient samples used for transcript-based biomarker tests are comprised of several cell types including tumour, stroma and immune subtypes. To improve biomarker signature performance, we have developed a prostate-selective stromal quantitation model which outperforms pan-organ models in prediction of stromal content. Importantly, our model led to considerable improvement of stromal signature performance in data adjusted for stromal proportion. The application of optimised stromal signatures will improve patient stratification and can be combined with tumour and immune signatures.

Acknowledgements

We would like to thank Prof David Berman for comments on the manuscript and the Pronto Team for advice and encouragement. This work was funded by Prostate Cancer Canada and the Movember foundation, grant #T2014-01 (NB and AAT), and by the Canadian Cancer Society INNOV14-1 #702423 (CN and AAT). TAB was supported by a PCF, USA Young Investigator award. The funders had no role in the design of the study and collection, analysis and interpretation of the data, or in writing the manuscript.

Author contributions statement

NB performed bio-informatic data analysis and experimental work. MT performed bio-informatic analysis of stromal dilution cohort. CN contributed to bio-informatic analysis and experimental work. TAB developed the stromal dilution cohort. NE and ED provided access to cohorts used in analysis. AAT conceived the study and supervised the work. NB and AAT wrote the paper. All authors read and approved the final manuscript.

Availability of data and material

Most data used in this study are publically available, via TCGA and GEO. Some data that support the findings of this study in regard to data of defined stromal content are available from GenomeDX Inc. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from GenomeDX upon reasonable request.

References

1. Dakhova O, Ozen M, Creighton CJ, *et al.* Global gene expression analysis of reactive stroma in prostate cancer. *Clin Cancer Res* 2009; **15:** 3979–3989.
2. Dakhova O, Rowley D, Ittmann M. Genes upregulated in prostate cancer reactive stroma promote prostate cancer progression in vivo. *Clin Cancer Res* 2014; **20:** 100–109.
3. Ayala G, Tuxhorn JA, Wheeler TM, *et al.* Reactive stroma as a predictor of biochemical-free recurrence in prostate cancer. *Clin Cancer Res* 2003; **9:** 4792–4801.
4. Yanagisawa N, Li R, Rowley D, *et al.* Stromogenic prostatic carcinoma pattern (carcinomas with reactive stromal grade 3) in needle biopsies predicts biochemical recurrence-free survival in patients after radical prostatectomy. *Hum Pathol* 2007; **38:** 1611–1620.
5. Barron DA, Rowley DR. The reactive stroma microenvironment and prostate cancer progression. *Endocr Relat Cancer* 2012; **19:** R187–R204.
6. Hagglof C, Bergh A. The stroma-a key regulator in prostate function and malignancy. *Cancers (Basel)* 2012; **4:** 531–548.
7. Jia Z, Wang Y, Sawyers A, *et al.* Diagnosis of prostate cancer using differentially expressed genes in stroma. *Cancer Res* 2011; **71:** 2476–2487.
8. Chen X, Xu S, McClelland M, *et al.* An accurate prostate cancer prognosticator using a seven-gene signature plus Gleason score and taking cell type heterogeneity into account. *PLoS One* 2012; **7:** e45178.
9. Jia Z, Rahmatpanah FB, Chen X, *et al.* Expression changes in the stroma of prostate cancer predict subsequent relapse. *PLoS One* 2012; **7:** e41371.
10. de Ridder D, van der Linden CE, Schonewille T, *et al.* Purity for clarity: the need for purification of tumour cells in DNA microarray studies. *Leukemia* 2005; **19:** 618–627.
11. de Bruin EC, van de Pas S, Lips EH, *et al.* Macrodissection versus microdissection of rectal carcinoma: minor influence of stroma cells to tumour cell gene expression profiles. *BMC Genomics* 2005; **6:** 142.
12. Aran D, Sirota M, Butte AJ. Corrigendum: systematic pan-cancer analysis of tumour purity. *Nat Commun* 2016; **7:** 10707.
13. Smits AJ, Kummer JA, de Bruin PC, *et al.* The estimation of tumour cell percentage for molecular testing by pathologists is not accurate. *Mod Pathol* 2014; **27:** 168–174.
14. Ghosh D. Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics* 2004; **20:** 1663–1669.
15. Stuart RO, Wachsman W, Berry CC, *et al.* In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc Natl Acad Sci U S A* 2004; **101:** 615–620.
16. Erkkila T, Lehmusvaara S, Ruusuvuori P, *et al.* Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics* 2010; **26:** 2571–2577.
17. Shen-Orr SS, Tibshirani R, Khatri P, *et al.* Cell type-specific gene expression differences in complex tissues. *Nat Methods* 2010; **7:** 287–289.
18. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3:** Article 3.
19. Abbas AR, Wolslegel K, Seshasayee D, *et al.* Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* 2009; **4:** e6098.
20. Gaujoux R, Seoighe C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect Genet Evol* 2012; **12:** 913–921.
21. Qiao W, Quon G, Csaszar E, *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput Biol* 2012; **8:** e1002838.
22. Yoshihara K, Shahmoradgoli M, Martinez E, *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013; **4:** 2612.
23. Kuhn A, Thu D, Waldvogel HJ, *et al.* Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat Methods* 2011; **8:** 945–947.
24. Shoemaker JE, Fukuyama S, Eisfeld AJ, *et al.* Integrated network analysis reveals a novel role for the cell cycle in 2009 pandemic influenza virus-induced inflammation in macaque lungs. *BMC Syst Biol* 2012; **6:** 117.
25. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 2013; **29:** 1083–1085.

26. Zhong Y, Wan YW, Pang K, *et al.* Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* 2013; **14:** 89.
27. Quon G, Haider S, Deshwar AG, *et al.* Computational purification of individual tumour gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med* 2013; **5:** 29.
28. Ahn J, Yuan Y, Parmigiani G, *et al.* DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* 2013; **29:** 1865–1871.
29. Clarke J, Seo P, Clarke B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* 2010; **26:** 1043–1049.
30. Gosink MM, Petrie HT, Tsinoremas NF. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics* 2007; **23:** 3328–3334.
31. Becht E, Giraldo NA, Lacroix L, *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* 2016; **17:** 218.
32. Wang Y, Xia XQ, Jia Z, *et al.* In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res* 2010; **70:** 6448–6455.
33. Anghel CV, Quon G, Haider S, *et al.* ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* 2015; **16:** 156.
34. Planche A, Bacac M, Provero P, *et al.* Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PLoS One* 2011; **6:** e18640.
35. Gregg JL, Brown KE, Mintz EM, *et al.* Analysis of gene expression in prostate cancer epithelial and interstitial stromal cells using laser capture microdissection. *BMC Cancer* 2010; **10:** 165.
36. Tomlins SA, Mehra R, Rhodes DR, *et al.* Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 2007; **39:** 41–51.
37. Richardson AM, Woodson K, Wang Y, *et al.* Global expression analysis of prostate cancer-associated stroma and epithelia. *Diagn Mol Pathol* 2007; **16:** 189–197.
38. Taylor BS, Schultz N, Hieronymus H, *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* 2010; **18:** 11–22.
39. Erho N, Crisan A, Vergara IA, *et al.* Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* 2013; **8:** e66855.
40. Lili LN, Matyunina LV, Walker LD, *et al.* Molecular profiling predicts the existence of two functionally distinct classes of ovarian cancer stroma. *Biomed Res Int* 2013; **2013:** 846387.
41. Karnes RJ, Bergstralh EJ, Davicioni E, *et al.* Validation of a genomic classifier that predicts metastasis following radical prostatectomy in an at risk patient population. *J Urol* 2013; **190:** 2047–2053.
42. Ross AE, Johnson MH, Yousefi K, *et al.* Tissue-based genomics augments post-prostatectomy risk stratification in a natural history cohort of intermediate- and high-risk men. *Eur Urol* 2016; **69:** 157–165.
43. Piccolo SR, Sun Y, Campbell JD, *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* 2012; **100:** 337–344.
44. Leek JT, Johnson WE, Parker HS, *et al.* The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012; **28:** 882–883.
45. Durinck S, Spellman PT, Birney E, *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009; **4:** 1184–1191.
46. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. *J Mol Biol* 1990; **215:** 403–410.
47. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9:** 559.
48. Miller JA, Cai C, Langfelder P, *et al.* Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* 2011; **12:** 322.
49. Colaprico A, Silva TC, Olsen C, *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016; **44:** e71.
50. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* 2015; **163:** 1011–1025.
51. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15:** 550.
52. Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43:** e47.
53. Orr B, Riddick AC, Stewart GD, *et al.* Identification of stromally expressed molecules in the prostate by tag-profiling of cancer-associated fibroblasts, normal fibroblasts and fetal prostate. *Oncogene* 2012; **31:** 1130–1142.
54. Webber MM, Trakul N, Thraves PS, *et al.* A human prostatic stromal myofibroblast cell line WPMY-1: a model for stromal-epithelial interactions in prostatic neoplasia. *Carcinogenesis* 1999; **20:** 1185–1192.
55. Franco OE, Jiang M, Strand DW, *et al.* Altered TGF-beta signaling in a subpopulation of human stromal cells promotes prostatic carcinogenesis. *Cancer Res* 2011; **71:** 1272–1281.
56. Hayward SW, Dahiya R, Cunha GR, *et al.* Establishment and characterization of an immortalized but non-transformed human prostate epithelial cell line: BPH-1. *In Vitro Cell Dev Biol Anim* 1995; **31:** 14–24.
57. Bischl B, Lang M, Kotthoff L, *et al.* mlr: Machine Learning in R. *J Mach Learn Res* 2016; **17:** 1–5.
58. Cheng T, Wang Y, Bryant SH. FSelector: a Ruby gem for feature selection. *Bioinformatics* 2012; **28:** 2851–2852.
59. Fox J, Weisberg S. *An {R} Companion to Applied Regression* (3rd edn). Sage: Thousand Oaks, CA, 2019. [Last accessed 11 October 2019]. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
60. Maechler M, Rousseeuw P, Struyf A, *et al. cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.0, 2019. [Last accessed 11 October 2019]. Available from: https://cran.r-project.org/package=cluster.
61. Kassambara A, Mundt F. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5, 2017. [Last accessed 11 October 2019]. Available from: https://CRAN.R-project.org/package=factoextra.
62. Mo F, Lin D, Takhar M, *et al.* Stromal gene expression is predictive for metastatic primary prostate cancer. *Eur Urol* 2018; **73:** 524–532.
63. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016; **32:** 2847–2849.
64. Therneau T. *A Package for Survival Analysis in S*. version 2.38, 2015. Available from: https://CRAN.R-project.org/package=survival.
65. Sachs MC. plotROC: a tool for plotting ROC curves. *J Stat Softw* 2017; **79:** 1–19.
66. Orr B, Grace OC, Brown P, *et al.* Reduction of pro-tumourigenic activity of human prostate cancer-associated fibroblasts using Dlk1 or SCUBE1. *Dis Model Mech* 2013; **6:** 530–536.
67. Vanpoucke G, Orr B, Grace OC, *et al.* Transcriptional profiling of inductive mesenchyme to identify molecules involved in prostate development and disease. *Genome Biol* 2007; **8:** R213.
68. Boufaied N, Nash C, Rochette A, *et al.* Identification of genes expressed in a mesenchymal subset regulating prostate organogenesis

using tissue and single cell transcriptomics. *Sci Rep* 2017; **7:** 16385.

69. Uhlen M, Fagerberg L, Hallstrom BM, *et al.* Proteomics: tissue-based map of the human proteome. *Science* 2015; **347:** 1260419.

70. Ayala GE, Muezzinoglu B, Hammerich KH, *et al.* Determining prostate cancer-specific death through quantification of stromogenic carcinoma area in prostatectomy specimens. *Am J Pathol* 2011; **178:** 79–87.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Expression of stromal transcripts in mixtures of prostate epithelial and fibroblast cells with varying proportions (0–100%)

Figure S2. Diagnostic plots for stromal model performance

Figure S3. Relationship between stromal transcript level and stromal proportion defined visually

Figure S4. ROC curve for our stroma prediction model and CellPred in the GSE17951 dataset (Wang *et al* [32])

Figure S5. Image analysis to define stromal content of samples in the TCGA dataset

Figure S6. Relationship between histological based tissue quantification and computational based tissue estimation methods

Figure S7. Stromal model performance in GenomeDX stromal dilution cohort and comparison with ESTIMATE

Figure S8. The performance of a 15-transcript stromal classifier [9] in GSE21034 [38] with or without stromalisation

Figure S9. The performance of a 93-gene SDMS [62] in GSE21034 [38] with or without stromalisation

Figure S10. The performance of a 93-gene SDMS [62] in GSE46691 [39] with or without stromalisation