OPEN

Prostate Cancer and Prostatic Disease (2014) 17, 81–90
© 2014 Macmillan Publishers Limited All rights reserved 1365-7852/14

www.nature.com/pcan

ORIGINAL ARTICLE

An expression signature at diagnosis to estimate prostate cancer patients' overall survival

Z Peng[1], L Skoog[1,2], H Hellborg[3], G Jonstam[4], I-L Wingmo[2], M Hjälm-Eriksson[1,4], U Harmenberg[1,4], GC Cedermark[1,4], K Andersson[5], L Ährlund-Richter[6], S Pramana[7], Y Pawitan[7], M Nistér[1,2], S Nilsson[1,4] and C Li[1,4]

BACKGROUND: This study aimed to identify biomarkers for estimating the overall and prostate cancer (PCa)-specific survival in PCa patients at diagnosis.
METHODS: To explore the importance of embryonic stem cell (ESC) gene signatures, we identified 641 ESC gene predictors (ESCGPs) using published microarray data sets. ESCGPs were selected in a stepwise manner, and were combined with reported genes. Selected genes were analyzed by multiplex quantitative polymerase chain reaction using prostate fine-needle aspiration samples taken at diagnosis from a Swedish cohort of 189 PCa patients diagnosed between 1986 and 2001. Of these patients, there was overall and PCa-specific survival data available for 97.9%, and 77.9% were primarily treated by hormone therapy only. Univariate and multivariate Cox proportional hazard ratios and Kaplan–Meier plots were used for the survival analysis, and a *k*-nearest neighbor (kNN) algorithm for estimating overall survival.
RESULTS: An expression signature of VGLL3, IGFBP3 and F3 was shown sufficient to categorize the patients into high-, intermediate- and low-risk subtypes. The median overall survival times of the subtypes were 3.23, 4.00 and 9.85 years, respectively. The difference corresponded to hazard ratios of 5.86 (95% confidence interval (CI): 2.91–11.78, $P<0.001$) for the high-risk subtype and 3.45 (95% CI: 1.79–6.66, $P<0.001$) for the intermediate-risk compared with the low-risk subtype. The kNN models that included the gene expression signature outperformed the one designed on clinical parameters alone.
CONCLUSIONS: The expression signature can potentially be used to estimate overall survival time. When validated in future studies, it could be integrated in the routine clinical diagnostic and prognostic procedure of PCa for an optimal treatment decision based on the estimated survival benefit.

Prostate Cancer and Prostatic Disease (2014) **17,** 81–90; doi:10.1038/pcan.2013.57; published online 7 January 2014

Keywords: overall and cancer-specific survival; castration therapy; embryonic stem cells; gene expression signature

INTRODUCTION

An estimation of overall survival with or without treatment at the time of prostate cancer (PCa) diagnosis is of the utmost importance for selecting the most appropriate treatment.[1–4] The currently available clinical prognostic tools demonstrate an accuracy of 70–80% for the prediction of biochemical or PSA recurrence, but these tools are less effective at predicting cancer-specific survival and even less accurate at predicting overall survival.[1–6] It is known that PSA recurrence-free survival cannot be used as a reliable surrogate for PCa-specific or overall survival, as the clinical outcomes of recurrence can be highly variable. This is largely due to the variability in the survival benefit conferred by hormone or castration therapy.[2]

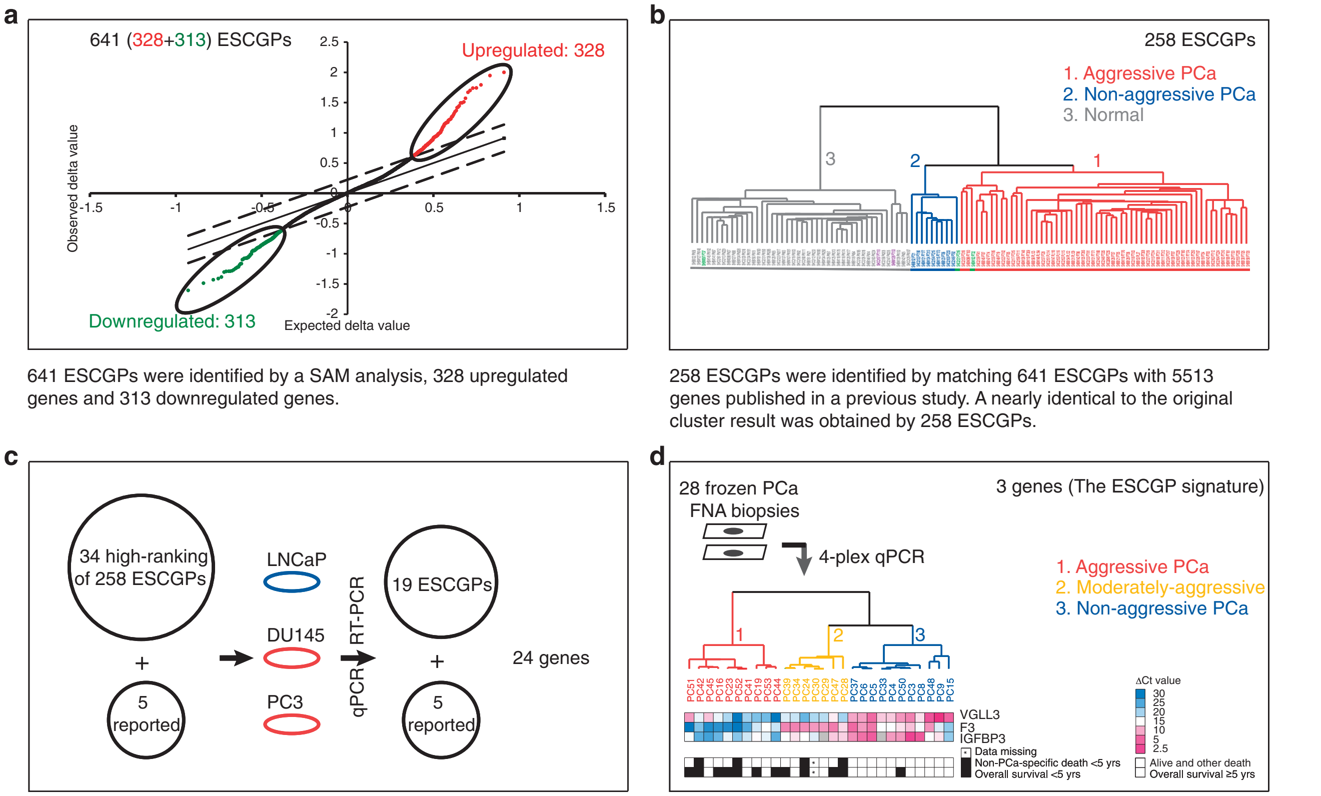
Currently, a primary treatment decision after the diagnosis is based on an overall evaluation of both tumor and patient risk factors. Tumor factors include serum PSA level, biopsy Gleason score and clinical stage, and patient factors include age, performance status and other diseases, that is, comorbidity. The challenge is to identify the most effective treatment that the patient can tolerate. Even though currently available clinical prognostic tools for the prediction of biochemical recurrence are valuable in clinical practice, a tool for estimating overall survival would improve the treatment decision.

Whole-genome expression analyses of tumor samples may identify new biomarkers that could improve the accuracy of survival prediction. However, most previous studies have identified biomarkers that only predict PSA recurrence-free survival, mainly due to limited clinical follow-up. Only a few recent studies have identified genomic markers associated with lethal forms of PCa.[7–11] This inability to predict overall survival is also due to the fact that the primary focus has been on tumor biological aggressiveness or tumor risk factors. However, it has been shown that about 50% of the patients can die of causes other than PCa. The observation underlines that the patient risk factors have the same importance as the tumor risk factors.

Several studies have demonstrated the use of embryonic stem cell (ESC) gene expression signatures for determining subtype classification and prognosis of various cancers, including PCa, as discussed in the review by Glinsky.[8] We have further developed this concept into a hypothesis of ESC gene predictors (ESCGPs) with the following reasoning: (1) Embryonic stem cells are the origin of tissue differentiated cells, tissue stem cells and cancer

[1]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden; [2]Department of Clinical Pathology/Cytology, Karolinska University Hospital, Stockholm, Sweden; [3]Regional Oncologic Center, Karolinska University Hospital, Stockholm, Sweden; [4]Department of Clinical Oncology, Karolinska University Hospital, Stockholm, Sweden; [5]Department of Oncology, Radiology and Radiation Sciences, Uppsala University, Uppsala, Sweden; [6]Department of Women and Children's Health, Karolinska Institutet, Stockholm, Sweden and [7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. Correspondence: Dr C Li, Department of Oncology-Pathology, Karolinska Institutet, Cancercentrum Karolinska, R8:04, Solna, Stockholm 17176, Sweden.
E-mail: chunde.li@ki.se

Received 28 August 2013; revised 14 October 2013; accepted 11 November 2013; published online 7 January 2014



641 ESCGPs were identified by a SAM analysis, 328 upregulated genes and 313 downregulated genes.

258 ESCGPs were identified by matching 641 ESCGPs with 5513 genes published in a previous study. A nearly identical to the original cluster result was obtained by 258 ESCGPs.

34 high-ranking out of 258 ESCGPs and 5 reported genes were verified their expression levels in 3 prostate cancer lines. 19 ESCGPs and 5 reported genes were selected for the clinical samples testing.

A 3-gene combination (The ESCGP signature) out of the 24 genes tested in a pilot set (Subset 1) of prostate cancer samples showed a promising power to classify tumor subtypes with obvious survival difference. Further survival difference was observed in the complete set (Figure 2).

Figure 1. Outline of a stepwise gene selection process. (**a**) Identification of 641 embryonic stem cell gene predictors (ESCGPs) by bioinformatic analysis. Previously published data sets of whole-genome complementary DNA microarrays derived from five human ESC lines and 115 human normal tissues from various organs were retrieved from the Stanford Microarray Database (SMD). After a data-centering process, a sub-data set with expression profile of 24 361 genes in the ESC lines was isolated from the combined whole data set. A single-class significance analysis of microarray (SAM) was performed and a SAM plot was generated. The 328 genes with the highest expression levels and 313 genes with the lowest expression levels were identified, in total 641 ESCGPs. (**b**) Identification of 258 ESCGPs in prostate cancer (PCa). PCa ESCGPs were identified by matching the list of the 641 ESCGPs and the list of 5513 genes published by Lapointe *et al.*[9] When clustering the 112 PCa tissue samples and comparing the cluster results when using all 5513 genes and when using only the 258 ESCGPs present in the data set, nearly identical results were obtained. Sample labeling: PL, lymph node metastasis; PN, normal prostate tissue; PT, prostate tumor. Three cases (marked green) were placed in different classification positions and two cases (purple) were consistently misclassified. (**c**) Selection of important candidate ESCGPs for clinical survival correlation. Of 258 PCa ESCGPs, 34 genes were selected by their high-ranking order in the SAM analysis identifying significant genes for the subtype classification or for the discriminating between tumor and normal samples. Of these 34 ESCGPs, 19 were selected based on their markedly different expression patterns and robust performances in RT-PCR reactions (Supplementary Figure S1). The 19 ESCGPs and the 5 reported genes were included in the optimization of the 4-plex qPCR method using RNAs from PCa cell lines. (**d**) Identification of the ESCGP signature in Subset 1. After the 4-plex qPCR optimization, the method was used to analyze 36 fresh–frozen fine-needle aspiration (FNA) biopsies taken from PCa patients (Subset 1). RNAs could be extracted in 28 biopsies. A series of cluster analyses using different gene combinations revealed that the ESCGP signature VGLL3, IGFBP3 and F3 classified Subset 1 samples into three subtypes with strong survival correlations. The level of gene expression increases from blue to red, whereas the delta Ct value decreases from blue to red. Gray areas represent missing qPCR data.

stem cells. (2) Genes that are important in maintaining ESC status and regulating differentiation are also important in maintaining cancer stem cell status and abnormal differentiation (dedifferentiation). (3) Genes with significant expression variations among different ES cell lines are not important in this respect. (4) Genes that show consistently high or consistently low expression levels across various ES cell lines are equally important in maintaining ESC status. Different expression patterns of these genes determine the development of different normal or cancer tissue. These genes are here named as ESCGPs. (5) These ESCGPs may be expressed not only in cancer stem cells but also cancer cells and their expressions can be measured by microarray, reverse transcription-polymerase chain reaction (RT-PCR) or quantitative PCR (qPCR). (6) Different expression patterns of these ESCGPs measured in the cancer tissues can reflect cancer's biological aggressiveness, and predict the efficacy of treatment and patient survival.[12–14]

To evaluate this hypothesis experimentally, we analyzed fine-needle aspiration (FNA) biopsy samples from 189 PCa patients with nearly complete follow-up data. The cohort is unique in that it contains high-quality fresh–frozen tumor samples and complete survival data. The majority of the patients were not treated radically, such as radical surgery or radiation. The clinical outcomes of these patients may closely resemble the natural course of development and outcome of PCa (i.e. natural history of PCa). We report that an ESCGPs expression signature at diagnosis could indeed estimate overall, PCa-specific and non-PCa-specific survival in this cohort. If our findings can be validated with concurrent cohorts dominated with radical surgery or radiation therapy, the signature may become an important complement in

Table 1. Characteristics of the patients

	Subset 1	*Subset 2*	*Subset 3*	*Complete set*
FNA biopsies, n	36	65	88	189
Mortality, n				
Overall	35	64	86	185
Death due to prostate cancer	13	40	45	98
Death due to other causes	19	21	25	65
Alive	3	3	16	22
Missing	1	1	2	4
Survival (years), median (range)	7.7 (0.1–17.8)	4.0 (0.2–15.7)	4.3 (0.2–15.1)	4.3 (0.1–17.8)
Age (years), mean ± s.d.	70.4 ± 7.8	72.1 ± 8.7	73.8 ± 8.9	72.6 ± 8.7
Missing, n	1	1	2	4
PSA (ng ml^{-1})[a], n *(%)*				
$\leqslant 20$	9 (32.1)	16 (31.2)	23 (28.7)	48 (29.8)
> 20 and $\leqslant 50$	9 (32.1)	14 (26.4)	23 (28.7)	46 (28.6)
> 50	10 (35.7)	23 (43.4)	34 (42.5)	67 (41.6)
Missing	8	12	8	28
Clinical stage[b], n *(%)*				
Localized	19 (59.4)	27 (45.8)	33 (39.3)	79 (45.1)
Advanced	13 (40.6)	32 (54.2)	51 (60.7)	96 (54.9)
Missing	4	6	4	14
WHO tumor grade, n *(%) (level of differentiation)*				
Well/moderately	22 (61.1)	31 (50.0)	33 (37.9)	86 (46.5)
Poorly	14 (38.9)	31 (50.0)	54 (62.1)	99 (53.5)
Missing	0	3	1	4
Treatment, n *(%)*				
Never treated	6 (19.4)	2 (3.3)	4 (4.9)	12 (7.0)
Hormone, orchiectomy	19 (61.3)	53 (88.3)	62 (76.5)	134 (77.9)
Radiation	5 (16.1)	2 (3.3)	11 (13.6)	18 (10.5)
Radical prostatectomy	1 (3.2)	3 (5.0)	4 (4.9)	8 (4.7)
Missing	5	5	7	17

Abbreviations: FNA, fine-needle aspiration; WHO, World Health Organization.
Advanced clinical stage was defined as $T \geqslant T3$ or N1 or M1 or PSA > 100 ng ml^{-1}.
[a] PSA levels in serum were measured at the time of diagnosis (before treatment).
[b] Localized clinical stage was defined as $T < T3$ and N0/Nx and M0/MX and PSA $\leqslant 100$ ng ml^{-1}.

the process of selecting therapeutic modality for each individual patient.

MATERIALS AND METHODS

Written approval from the local ethics committee was obtained for the molecular analysis of biological samples from PCa patients. This study was conducted in a stepwise manner. The procedure for selecting and verifying genes is outlined in Figure 1 and described in detail as follows.

Identification of candidate ESCGPs

Previously published data sets of whole-genome cDNA microarrays derived from five human ESC lines[15] and 115 human normal tissues from various organs[16] were retrieved from the SMD (Stanford Microarray Database) (Figure 1a). Initially we combined these two retrieved data sets, the data set of normal tissues was used to normalize the subset of ESC lines by a data-centering process[15,16], afterwards a sub-data set with whole-genome expression profile of the 24 361 genes in the ESC lines was isolated from the combined whole data set. A single-class significance analysis of microarrays (SAM)[17] was performed using the subset of the ESC lines only, whereby all genes were ranked according to the consistency (without significant variations) of their expression levels across the ESC lines, assuming that genes with significant expression variation between ESC lines would not be critical for maintaining stem cell-like status. Significant genes were selected at delta 0.23 and *q*-value $\leqslant 0.05$.

Selection of candidate ESCGPs in PCa

An independent data set[9] with 112 prostate tissue samples was used to verify the ESCGP findings and to select ESCGPs for PCa (Figure 1b). The list of genes in the published data set was matched to the list of the candidate ESCGPs identified in Step 1. This resulted in a shorter list of genes and the expression data of these genes were used to repeat the cluster analysis. The result was compared with the original cluster.

Refining ESCGPs selection using RT-PCR and multiplex qPCR analyses of three PCa cell lines

A SAM analysis was performed using the list of PCa ESCGPs identified in Step 2 (Figure 1c). The high-ranking ESCGPs and an additional five reported genes were selected for further analysis. A 4-plex qPCR method was optimized for the quantification of these genes by using RNAs from three PCa cell lines (LNCaP, DU145, PC-3). The procedures used for the isolation of total RNA from cell cultures and FNA cytology smears, cDNA synthesis, RT-PCR and multiplex qPCR analysis are described in the Supplementary Information file, Supplementary Tables S1–S3 and Supplementary Figures S1 and S2. For the qPCR analysis, the expression level of each gene in a sample was normalized to that of GAPDH (glyceraldehyde 3-phosphate dehydrogenase) and was presented as the delta Ct value, which is inversely correlated to the gene expression level.[18,19] This delta Ct value was centered by the median value across all samples, and the centered delta Ct value was then used for the statistical and *k*-nearest neighbor (kNN) analyses.

Establishing of the clinical importance

FNA samples. The FNA samples were collected according to the routine procedure used at the Clinical Pathology/Cytology unit at Karolinska University Hospital in Stockholm, Sweden[20] (Figure 1d). Multiple cytology smears were obtained by prostate FNA procedure in each patient at the time of diagnosis. The representative smear was identified by examination of Giemsa-stained slides and was used for the clinical cytological diagnosis. The remaining fresh smears on glass slides that were duplicates of the Giemsa-stained slide used for diagnosis were freshly frozen and kept at −70 °C. We found that at least 80% of all cells collected from most of the PCa FNA samples were cancer cells. Of the 241 FNA samples that were collected from the patients, we obtained good-quality total RNAs from 193 samples; 189 of these samples were from patients with a diagnosis of PCa. The researcher who performed the 4-plex qPCR analyses of the FNA samples was not informed of the relevant clinical data until the complete data set was constructed from both the qPCR results and the clinical data.

Table 2. Number of patients' samples for gene expression profiling

	N[a]			
Feature	*Subset 1*	*Subset 2*	*Subset 3*	*Complete set*
Total	36	65	88	189
CTGF	36	—	67	103
FBP1	36	—	46	82
EGR1	26	65	88	179
CYR61	36	—	46	82
WNT5B	36	—	56	92
LRP4	28	—	—	28
CDH1	36	—	—	36
BASP1	28	65	88	181
PTN	28	—	—	28
COL12A1	28	64	88	180
VGLL3	28	40	88	156
METTL7A	36	—	—	36
F3	28	—	67	95
GREM1	36	—	—	36
ERBB3	36	—	56	92
LRNN1	36	62	88	186
THBS1	28	—	—	28
IGFBP3	26	59	88	173
WNT11	28	65	88	181
c-MAF-a	26	64	88	178
c-MAF-b	26	—	46	72
AZGP1	—	63	88	151
AMACR	—	63	88	151
MUC1	—	58	88	146
EZH2	—	59	88	147
The ESCGP signature[b]	28	—	67	95

Abbreviation: ESCGP, embryonic stem cell gene predictors.
[a] The number of samples varies between genes because not all genes were profiled across all samples.
[b] The ESCGP signature includes the expression levels of VGLL3, IGFBP3 and F3.

Clinical characteristics of the cohort. The 189 PCa patients were diagnosed between 1986 and 2001. During this time period in Sweden, PCa diagnoses were mainly confirmed by prostate FNA cytology rather than by performing a multiple-core biopsy of the prostate.[20] Elderly men without lower urinary tract symptoms were seldom tested for their serum PSA levels. The average PSA level at the time of PCa diagnosis was therefore higher during this study than the level currently observed. In this cohort, very few patients had an indolent cancer, over 50% of patients had high-grade, advanced cancer, and hormone therapy was the primary treatment for 77.9% of the patients (Table 1). With regard to comorbidity, 40% of the patients had cardiovascular disease and 9% diabetes. An internship doctor who was not informed of the results of the molecular analyses collected the relevant clinical data under the supervision of an oncologist. Information with regard to the date of diagnosis, the date of death and the cause of death for all patients was first obtained from regional or national registries and was then verified by examining the medical journals. Diagnosis and cause of death were coded according to the International Classification of Diseases (ICD9 and ICD10) recommended by the World Health Organization.[21] PCa-specific mortality was assigned to cases where PCa or metastases were the primary or secondary cause of death. Death causes of patients are described in Supplementary Table S8. By 31 December 2008, 22 of the original 189 patients were still alive, 163 were deceased and 4 could not be found in the registries (Table 1).

Statistical analysis

Sample size and design of the subsets. The details for these procedures are provided in the Supplementary Information file. The patient cohort was divided into three subsets, according to the diagnoses and the experimental time order (Table 1). For Subset 1, we evaluated the strongest candidate genes from the ESCGP list. For Subset 2, we evaluated the most significant genes from Subset 1 and selected genes (reported genes) from the literature. For Subset 3, we tested the genes that demonstrated significance in Subset 2 and a limited number of genes that showed significance in Subset 1 but were not tested in Subset 2. A summary of the genes tested in the different subsets and complete set is shown in Table 2. Two major factors determined that not all candidate genes could be tested in every sample, that is, insufficient amount of total RNAs isolated from most FNA samples and fixed gene combinations by 4-plex qPCR.

Definition of important parameters. The details and references for the parameters are provided in the Supplementary Information file.

Survival analysis. The univariate and multivariate hazard ratios were calculated according to the Cox proportional hazard model using Stata statistics software (version 10.1; StataCorp LP, College Station, TX, USA). The Kaplan–Meier plots and statistical box plots were made using JMP statistics software (version 8.0.1; SAS Institute, Cary, NC, USA).

Cluster analysis. Gene expression data were evaluated using the unsupervised hierarchical clustering method and the gene median-centered delta Ct values, results were visualized using Treeview software (Eisen Lab, University of California at Berkeley, CA, USA).[22] Unsupervised hierarchical clustering is based on similarity measures and identifies clusters as groups of patterns.

Parametric model design. A first-order polynomial model using the selected genes was designed based on the assumption of the Weibull distribution in Stata statistics software (StataCorp LP). Of the 95 patients for whom there was data with regard to expression pattern of the ESCGP signature, 87 had expression data and clinical information that could be used for Weibull regression survival prediction. Two models were made, one using clinical parameters alone and one combining clinical parameters and ESCGP signature. The clinical parameters included PSA level (>50 vs $\leqslant 50\,\mathrm{ng\,ml^{-1}}$), clinical disease stage (advanced vs localized), tumor grade (poorly vs well/moderately differentiated) and age at the time of diagnosis. A first-order polynomial model using the selected genes was designed in Stata statistics software (StataCorp LP).

kNN modeling. The data set was randomly divided into a training set (70% of the data set, $n = 139$) for model development and a test set (30% of the data set, $n = 50$) for model verification. Four different kNN models estimating overall survival were designed and optimized on the training set data (Table 5). One of the models had only clinical parameters, one had only the ESCGP signature and two had combinations of clinical parameters and ESCGP signature. In all cases, models were applied only to cases without missing data, Euclidian distance measures were used and the average survival time for the three nearest neighbors was calculated as output. The scaling of the parameters of each model was determined through an exhaustive search of all combinations of the scaling factors 1, 3 and 9. A random number generator with a similar distribution as the overall survival in the data set was used as the reference for random guess of overall survival. For all models and the random number generator, the prediction performance of the kNN models was evaluated by comparing the average and variance of the absolute prediction error. kNN is a pattern based classification tool that assigns an unknown case to the same group as the most similar reference cases, meaning that kNN can be capable of classifying data sets where there is no simple univariate relationship between gene expression and patient outcome.

RESULTS

Identification of the ESCGP signature by a stepwise process

Identification of candidate ESCGPs. The SAM analysis of public data identified 328 genes with consistently high levels of expression and 313 genes with consistently low levels of expression in ESCs (Figure 1 and Supplementary Table S1), that is, 641 genes in total (Figure 1a).

Selection of candidate ESCGPs in PCa. The ability of the 641 genes to classify tumor subtype was verified on an independent data set of PCa samples,[9] wherein the clustering result was almost identical when comparing the complete original data set of 5513 genes and the 258 PCa-related ESCGPs isolated from the same original data set. Therefore, the 641 genes were defined as ESCGPs (Figure 1b).

Table 3. Cox proportional hazards analysis of ESCGPs and various clinical parameters (univariate analysis)

Feature	*n (%)*	N[a]	*Overall survival*		*PCa-specific survival*	
			Hazard ratio (95% CI)	*P-value*	*Hazard ratio (95% CI)*	*P-value*
PSA (ng ml^{-1})						
⩽50	94 (58%)	161	1.00 (reference)		1.00 (reference)	
>50	67 (42%)	161	2.34 (1.65–3.31)	<0.0001	2.61 (1.68–4.05)	<0.0001
WHO tumor grade (level of differentiation)						
Well/moderately	85 (47%)	181	1.00 (reference)		1.00 (reference)	
Poorly	96 (53%)	181	1.59 (1.16–2.18)	0.004	1.94 (1.28–2.94)	0.002
Clinical stage[b]						
Localized	79 (45%)	175	1.00 (reference)		1.00 (reference)	
Advanced	96 (55%)	175	1.70 (1.23–2.35)	0.001	2.20 (1.44–3.38)	<0.0001
Age[c]		185	1.04 (1.02–1.06)	<0.0001	1.03 (1.00–1.05)	0.035
PSA (ng ml^{-1})[d]		161	1.00 (1.00–1.00)	0.005	1.00 (1.00–1.00)	0.004
F3[e]		92	1.11 (1.04–1.17)	0.001	1.14 (1.06–1.22)	<0.0001
WNT5B[e]		89	1.14 (1.04–1.25)	0.004	1.26 (1.11–1.42)	<0.0001
VGLL3[e]		152	1.09 (1.04–1.15)	<0.0001	1.08 (1.02–1.15)	0.014
c-MAF-a[e]		174	1.09 (1.02–1.16)	0.008	1.09 (1.01–1.19)	0.036
CTGF[e]		100	1.13 (1.03–1.23)	0.008	1.15 (1.02–1.29)	0.023
IGFBP3[e]		169	1.05 (0.99–1.12)	0.078	1.10 (1.02–1.18)	0.013
c-MAF-b[e]		69	1.13 (0.96–1.33)	0.134	1.28 (1.04–1.57)	0.019
EZH2[e]		144	0.93 (0.83–1.04)	0.208	0.85 (0.74–0.97)	0.018
AMACR[e]		148	1.09 (1.02–1.16)	0.009	1.08 (1.00–1.17)	0.049
MUC1[e]		143	1.07 (1.01–1.14)	0.025	1.06 (0.99–1.15)	0.109

Abbreviations: CI, confidence interval; ESCGP, embryonic stem cell gene predictors; PCa, prostate cancer; WHO, World Health Organization.

[a] The number of samples varies between ESCGPs because not all ESCGPs were profiled across all samples.

[b] Localized clinical stage was defined as T<T3 and N0/Nx and M0/MX and PSA ⩽100 ng ml^{-1}. Advanced clinical stage was defined as T⩾T3 or N1 or M1 or PSA>100 ng ml^{-1}.

[c] Age was modeled as a continuous variable. The hazard ratio is for each 1.0 year increase in age.

[d] PSA was modeled as a continuous variable. The hazard ratio is for each 1.0 ng ml^{-1} PSA increase in serum.

[e] The centered delta Ct value for gene expression was modeled as a continuous variable. It is inversely correlated to the gene's expression level. The hazard ratio is for each increase of 1.0 unit in centered delta Ct value.

Table 4. Cox proportional hazards analysis of the ESCGP signature and various clinical parameters (univariate and multivariate analyses)

Feature	*n (%)*	N[a]	*Overall survival*				*PCa-specific survival*			
			Univariate analysis		*Multivariate analysis*		*Univariate analysis*		*Multivariate analysis*	
			Hazard ratio (95% CI)	*P-value*	*Hazard ratio (95% CI)*	*P-value*	*Hazard ratio (95% CI)*	*P-value*	*Hazard ratio (95% CI)*	*P-value*
The ESCGP signature[b]										
Group 3	26 (30%)	87	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Group 2	32 (37%)	87	3.45 (1.79–6.66)	<0.0001	2.51 (1.21–5.21)	0.013	3.99 (1.65–9.64)	0.002	2.96 (1.11–7.87)	0.030
Group 1	29 (33%)	87	5.86 (2.91–11.78)	<0.0001	4.77 (2.27–10.01)	<0.0001	7.67 (3.04–19.36)	<0.0001	7.12 (2.56–19.85)	<0.0001
PSA (ng ml^{-1})										
⩽50	48 (55%)	87	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
>50	39 (45%)	87	2.93 (1.76–4.86)	<0.0001	2.09 (1.10–3.94)	0.023	3.33 (1.73–6.41)	<0.0001	1.76 (0.77–4.03)	0.183
WHO tumor grade (level of differentiation)										
Well/ moderately	35 (40%)	87	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Poorly	52 (60%)	87	1.65 (1.03–2.66)	0.039	1.17 (0.69–2.00)	0.556	1.93 (1.04–3.57)	0.038	1.20 (0.61–2.39)	0.596
Clinical stage[c]										
Localized	37 (43%)	87	1.00 (reference)		1.00 (reference)		1.00 (reference)		1.00 (reference)	
Advanced	50 (57%)	87	2.13 (1.32–3.45)	0.002	1.68 (0.91–3.08)	0.097	3.87 (1.94–7.70)	<0.0001	3.62 (1.55–8.45)	0.003
Age[d]	87	1.06 (1.03–1.09)	<0.0001	1.03 (1.00–1.06)	0.048	1.06 (1.02–1.10)	0.003	1.03 (0.99–1.08)	0.108	

Abbreviations: CI, confidence interval; ESCGP, embryonic stem cell gene predictors; PCa, prostate cancer; WHO, World Health Organization.

[a] Eighty-seven out of the 95 clustered samples had all clinical information including age at diagnosis, PSA value, WHO tumor grade and clinical stage. Univariate and multivariate analyses included these 87 samples.

[b] The ESCGP signature includes the expression levels of VGLL3, IGFBP3 and F3, and classified samples into three tumor subtypes (Group 1, Group 2 and Group 3) by Cluster analysis (Figure 2a). It was modeled as a non-continuous variable with three categories according to the tumor subtype.

[c] Localized clinical stage was defined as T<T3 and N0/Nx and M0/MX and PSA ⩽100 ng ml^{-1}. Advanced clinical stage was defined as T⩾T3 or N1 or M1 or PSA>100 ng ml^{-1}.

[d] Age was modeled as a continuous variable. The hazard ratio is for each 1.0 year increase in age.

Refining ESCGPs selection using RT-PCR and multiplex qPCR analyses of three PCa cell lines. Within the 258 verified PCa ESCGPs, the 34 genes of highest ranking order in SAM analyses for the discrimination between tumor and normal samples and between different tumor subtypes were selected for follow-up analysis. In addition, five reported genes based on previously published

Table 5. Analysis of classification error of kNN model performance.

Model	*Parameters [scale]*	*Training set (n = 139) classification error (overall survival, years)*		*Test set (n = 50) classification error (overall survival, years)*		*P-value (is random)*
		Average error	*Standard deviation*	*Average error*	*Standard deviation*	
RND	Random numbers			3.85	3.18	
1	Age, WHO, CS, log-PSA [1,9,3,9]	2.97	2.14	3.44	2.93	0.2300
2	IGFBP3, VGLL3, F3, min(3G), max(3G) [1,9,1,3,1]	2.82	2.36	3.14	2.71	0.0373
3	Age, log-PSA, IGFBP3, VGLL3, F3, min(3G), max(3G) [1,3,1,9,3,3,1]	2.85	2.05	2.69	2.34	0.0024
4	Age, WHO, CS, log-PSA, IGFBP3, VGLL3, F3, min(3G), max(3G) [1,9,9,3,1,9,1,3,1]	2.81	1.76	2.72	2.41	0.0038

Abbreviations: CS, clinical stage; 3G, IGFBP3, VGLL3, F3; kNN, *k*-nearest neighbor; WHO, World Health Organization tumor grading.

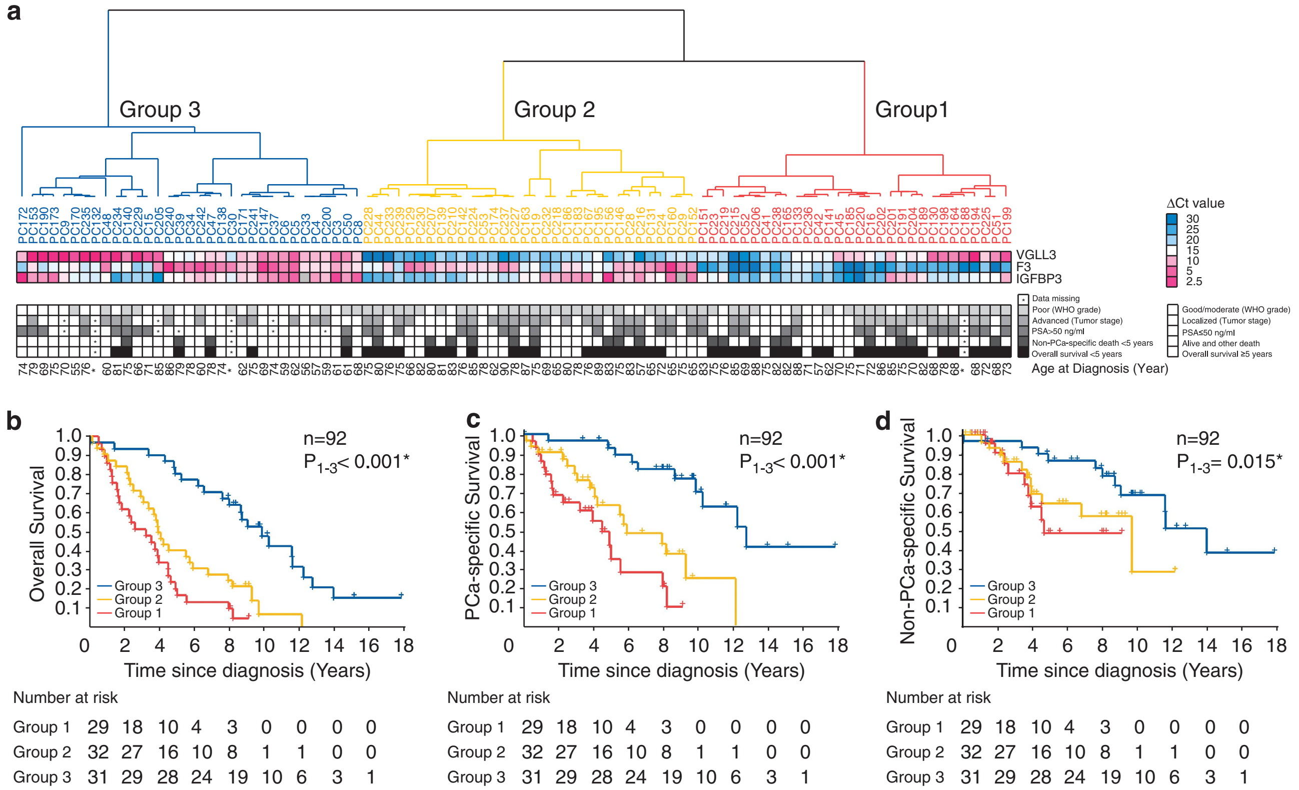
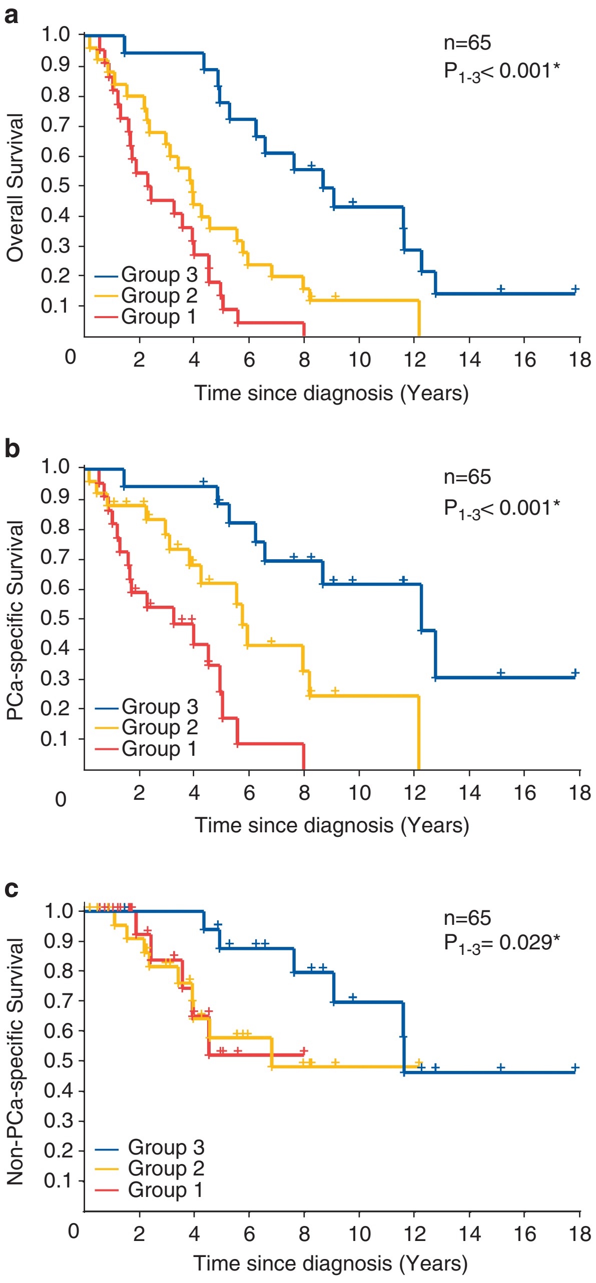


Figure 2. Clear survival difference according to tumor subtypes classification based on the embryonic stem cell gene predictor (ESCGP) signature (VGLL3, IGFBP3 and F3). Data were available for evaluation of the ESCGP signature for 95 of the 189 patients. (**a**) Fine-needle aspiration (FNA) samples from the 95 patients were used to create three tumor subtypes (group 1, red tree; group 2, yellow tree; group 3, blue tree) according to the ESCGP signature. The expression data was evaluated using the unsupervised hierarchical clustering method and the gene median-centered delta Ct values; the results were visualized using Treeview software. The gene expression level increases from blue to red, whereas the delta Ct value decreases from blue to red. Missing data are represented by the gray color. The clinical parameters of each patient are marked by various squares. Empty squares represent a longer survival period, lower PSA level, localized PCa clinical disease stage and a well or moderately differentiated tumor grade. Squares with various fill colors represent a shorter survival period, higher PSA level, advanced clinical disease stage and poorly differentiated tumor grade. (**b**–**d**) The overall, PCa-specific and non-PCa-specific survival analyses of the three subtypes were presented by Kaplan–Meier curves. X and Y axis presents actual time as diagnosis and survival rate, respectively. The *P*-values for differences between each of the three tumor subtypes were calculated using a log-rank test, and the *P*-values marked with stars represent statistical significance (*P*-value < 0.05). Besides the most significant difference between subtypes 1 and 3 shown in the figure, the other *P*-values between each two subtypes were $P_{1-2} = 0.063$, $P_{2-3} < 0.001$ (**b**); $P_{1-2} = 0.063$, $P_{2-3} < 0.001$ (**c**); $P_{1-2} = 0.523$, $P_{2-3} = 0.070$ (**d**).

studies[9,23–32] were included in the same set, both to serve as a positive reference and to make possible the investigation if these known genes could improve the predictive power of the final gene signature. Of the 34 ESCGPs, 19 had both robust performances in RT-PCR reactions and clearly different expression patterns in PCa cell lines (Figure 1c and Supplementary Figure S1). The 19 ESCGPs and 5 reported genes (Table 2) were included in an optimization of 4-plex qPCR using RNAs from PCa



Number at risk

Group 1	22	12	6	1	0	0	0	0	0
Group 2	25	20	11	6	4	1	1	0	0
Group 3	18	17	17	13	10	6	4	2	1

Figure 3. Survival difference between the three tumor subtypes classified according to the embryonic stem cell gene predictor (ESCGP) signature in patients primarily treated with castration therapy. Of the 95 patients shown in Figure 2, 65 received castration therapy as their primary treatment. Within this group, clear survival differences could still be observed according to the three tumor subtypes classified based on the ESCGP signature. The overall (upper panel), PCa-specific (middle panel) and non-PCa-specific (lower panel) survival analyses of the three subtypes are shown by the Kaplan–Meier curves. The *P*-values for differences between each of the three tumor subtypes were calculated using a log-rank test. Besides the most significant difference between subtypes 1 and 3 shown in the figure, the other *P*-values between each two subtypes were $P1–2 = 0.037$*, $P2–3 = 0.001$* (**a**); $P_{1–2} = 0.009$*, $P_{2–3} = 0.006$* (**b**); $P_{1–2} = 0.955$, $P_{2–3} = 0.076$ (**c**). The overall survival rates at 5 years of follow-up were 13.6%, 36.0% and 77.8% for groups 1, 2 and 3, respectively.

cell lines. One gene (*MAF*) has two different mRNA transcripts (c-MAF-a and c-MAF-b), and both were included. After the optimization, the 4-plex qPCR method was ready to use for the analysis of FNA samples taken from PCa patients.

Establishing of the clinical importance. Finally, the potential additive or synergistic effects by different combinations of the 10 significant genes identified in the univariate analysis were explored (Table 3). Using the data of 36 patient samples in the pilot subset (Subset 1), a series of cluster analyses were performed using 120 different gene combinations by the selection of k $(2 \leqslant k \leqslant 10)$ different genes each time. Of these 120 different combinations, the ESCGP signature VGLL3, IGFBP3 and F3 was the best for tumor subtype classification in correlation to survival differences (Figure 1d). The risk of false discovery by multiple testing is reduced as survival difference was correlated to the ESCGP signature as observed in cluster analysis, univariate and multivariate analyses also after inclusion of two additional subsets of patients (Subset 2 with 65 patients and Subset 3 with 88 patients, Table 2).

The resulting gene expression data for the complete cohort was subjected to analysis with respect to overall and PCa-specific survival. In the univariate analysis, all of the clinical parameters were significantly correlated with both overall and PCa-specific survival (Table 3 and Supplementary Figures S3 and S4). Of the 25 gene expression markers, 10 (F3, WNT5B, VGLL3, CTGF, IGFBP3, c-MAF-a, c-MAF-b, AMACR, MUC1 and EZH2) were significantly correlated with either overall or PCa-specific survival. Two of these markers (F3 and WNT5B) presented a more significant *P*-value than did PSA when they were used as continuous variables, and this level of significance remained after a stringent Bonferroni correction was performed for the multiple testing of 30 variables ($P < \beta = 0.0016667$; Table 3 and Supplementary Table S5). A multivariate analysis was performed to evaluate the influence of clinical parameters on the significance of each gene variable. The number of patients included in the multivariate analysis was smaller than that included in the univariate analysis because several parameters had missing data. In summary, four markers, F3, IGFBP3, CTGF and AMACR, showed correlations with both overall and PCa-specific survival, which were independent of any of the clinical parameters evaluated (Supplementary Tables S6 and S7). Three of these genes (*F3*, *IGFBP3*, *CTGF*) are ESCGPs.

Of the 189 patients evaluated, 87 had data available for all clinical parameters (mainly patients in Subsets 1 and 3) and could be classified into subtypes according to the expression signatures of VGLL3, IGFBP3 and F3. The multivariate analysis for overall and PCa-specific survival revealed that the tumor subtype classification defined by the ESCGP signature was the most powerful survival indicator and further independent of age, PSA level, tumor grade and clinical stage (Table 4). The median overall survival time was 3.23 years for patients with the high-risk subtype, 4.00 years for the intermediate-risk subtype and 9.85 years for the low-risk subtype (Figure 2), and these values corresponded to hazard ratios of 5.86 (95% confidence interval (CI): 2.91–11.78, $P < 0.001$) for the high-risk subtype and 3.45 (95% CI: 1.79–6.66, $P < 0.001$) for the intermediate-risk subtype compared with the low-risk subtype (Table 4 and Figure 2). Kaplan–Meier plots further indicated a clear survival difference between the three subtypes classified using the ESCGP signature (Figure 2 and Supplementary Figures S4 and S5). The difference in overall survival was attributed to both PCa-specific and non-PCa-specific survival (Figure 2). Interestingly, the survival difference between the three tumor subtypes was maintained when only patients treated with hormone therapy were analyzed, and these differences were independent of all other clinical parameters (Figure 3 and Supplementary Figure S5). Results from separate analysis of the subgroup of patients with cardiovascular disease were in agreement with the results from the complete cohort.

Survival predictions with the combined use of the ESCGP signature and various clinical parameters

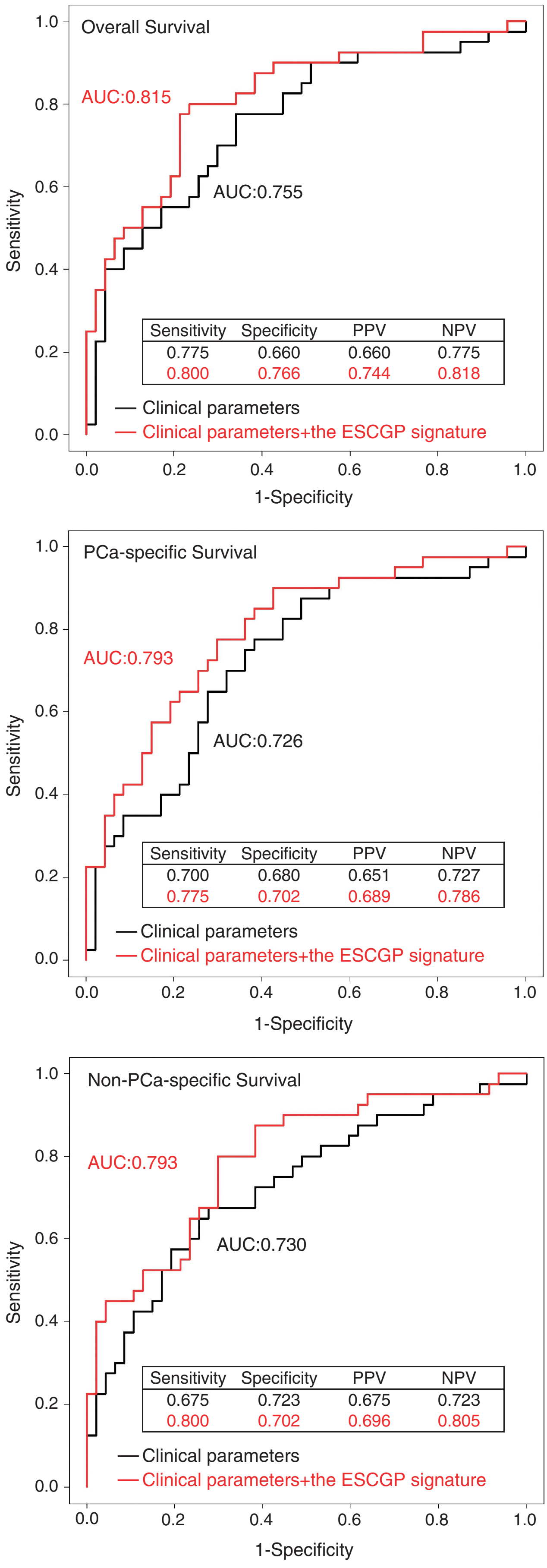
To assess the predictive performance of the selected ESCGP genes, different kNN classification algorithms were developed using the training set to estimate the overall survival.[33] When evaluated on the test set (Table 5), the performance of the kNN model using only clinical parameters was similar to the random model, whereas all kNN models including the selected ESCGP genes were significantly ($P<0.04$) better than the random model. Another illustration of predictive performance was obtained using a parametric model. This model was used to estimate whether the use of tumor subtype classification,[34] according to the expression signature of VGLL3, IGFBP3 and F3, could improve the prediction of survival beyond that estimated using the available clinical parameters (Figure 4). Compared with the prediction model that used only the clinical parameters, the addition of the tumor subtype classification improved sensitivity and specificity of the overall survival prediction from 0.775 to 0.800, and from 0.660 to 0.766, respectively (at 5 years; Figure 4). Receiver operating characteristic curves at 5-year survival were estimated to show the sensitivity and the specificity of survival prediction. The area under the receiver operating characteristic curve value was increased from 0.755 to 0.815 in overall survival prediction, from 0.726 to 0.793 in PCa-specific survival prediction and from 0.730 to 0.793 in non-PCa-specific survival prediction, respectively (Figure 4).

DISCUSSION

This report discusses the ability to estimate the overall and cancer-specific survival using gene expression levels in PCa samples. If such a measure would become available, it would provide an important and orthogonal complement to the currently available data used in the decision process for selecting treatment for individual patients.

Numerous attempts to produce prognostic methods for PCa use surrogate end points like biochemical relapse or even cancer-specific mortality.[35] This is probably due to the fact that data sets for surrogate end points are more easily obtained. However, it has been shown that nearly 50% of the patients die of diseases other than PCa.[36] The identification of biomarkers that correlate with overall survival of PCa is rare. Our results demonstrate that PCa tumor subtypes classified by the gene expression signature of VGLL3, IGFBP3 and F3 at the time of diagnosis are clearly correlated with overall and cancer-specific survival in the evaluated cohort.

The gene expression signature was independent of age, PSA level, World Health Organization tumor grade and clinical stage. Furthermore, as shown through the kNN model and the parametric prediction model, this signature demonstrated clear prognostic value and a potential to further improve the prognostic accuracy of conventional clinical parameters. Following validation on additional cohorts, this ESCGP signature could be particularly beneficial in the clinical management of early-stage PCa. In such



Sensitivity	Specificity	PPV	NPV
0.775	0.660	0.660	0.775
0.800	0.766	0.744	0.818

Sensitivity	Specificity	PPV	NPV
0.700	0.680	0.651	0.727
0.775	0.702	0.689	0.786

Sensitivity	Specificity	PPV	NPV
0.675	0.723	0.675	0.723
0.800	0.702	0.696	0.805

Figure 4. Receiver operating characteristic (ROC) curves for 5-year survival prediction. Prediction of survival time was modeled using a parametric model based on the assumption of the Weibull distribution. ROC curves at 5-year survival prediction show the sensitivity and the specificity of survival prediction. Overall (upper panel), PCa-specific (middle panel) and non-PCa-specific survival (bottom panel) predictions at 5 years were determined by the clinical parameters alone (black lines), and by both clinical parameters and the tumor subtypes classified by embryonic stem cell gene predictor (ESCGP) signature (red lines). The area under the curve (AUC) values of overall, PCa-specific and non-PCa-specific survival predictions were all increased by adding ESCGP signature. Positive predictive value (PPV) and negative predictive value (NPV) both increased.

cases, the accuracy of conventional clinical parameters for the prediction of cancer-specific and overall survival are limited by a relatively low PSA level, localized disease stage and insufficient tumor material for Gleason scoring.[1–4,37] When evaluating such small tumor samples, the ESCGP has a potential to improve the assessment.

Overall survival is the real lifetime determined by the aggressiveness of PCa and patient's other conditions or comorbidities.[2] The ability to estimate overall survival by the ESCGP signature may reflect the biological functions of the three genes. Both F3 and IGFBP3 have been shown associated with metastasis development in prostate and other cancers. They have also been shown important in the development of many non-cancer diseases of the coagulatory, cardiovascular and metabolic systems, diseases that are common causes of death in PCa patients.[28,38] The positive correlation between prolonged survival and increased expression of F3 was unexpected and may suggest that PCa cells with higher levels of F3 are strongly androgen-dependent and sensitive to castration treatment.[9,24,39,40] The functions of VGLL3 have yet to be studied. VGLL3 shows clear correlation with the age at diagnosis (Supplementary Tables S6 and S7), which is an important patient risk factor that strongly influences the patient overall survival and treatment decision. We suggest that the expression of VGLL3 may reflect the patient's biological age that currently can be estimated only by physician's subjective observation. Therefore, the combination of these three genes could provide a molecular classification sufficient to estimate overall survival.

Several reported gene markers (AMACR, EZH2, c-MAF-a, c-MAF-b and MUC1) selected from previous studies were also validated in our FNA cohort (Supplementary Table S5); however, they were not as strong as the ESCGP signature when estimating overall survival. Owing to the limited RNA quantity present in the FNA samples, the previously reported 'stemness' gene signature[8] could not be compared with our ESCGP signature, although this comparison would be warranted in future studies.

The present study was driven by the stem cell hypothesis, whereby ESC gene expression signatures are thought to be associated with the prognosis of various cancers. Our results demonstrate that the 258 ESCGPs could classify an independent PCa data set in a nearly identical manner as compared with using the complete 5513 genes identified in a previous study (Figure 1b). Furthermore, PCa tumor subtypes classified by the ESCGP signature of VGLL3, IGFBP3 and F3 at the time of diagnosis clearly correlated with overall and PCa-specific survival. These two results support the stem cell hypothesis.

The stepwise procedure implemented in the current study has both advantages and disadvantages. We find it advantageous to use an initially wide concept and incrementally narrow the scope through use of independent historic data sets and new measurements, as illustrated in Figure 1. The drawback is that Subset 1 of the patient database was part of selecting the ESCGP signature, leaving a smaller set of patient material for validation. In our case, the limited availability of FNA samples prevented us from dividing them into one discovery set and one validation set. On the other hand, the availability of a series of high-quality, fresh–frozen FNA samples with nearly complete survival data made it possible to complete this study. Currently, evaluation of Gleason score using transrectal ultrasound-guided prostate biopsy samples has become the major diagnostic procedure for PCa. A direct comparison and correlation between this signature and Gleason score needs to be established using biopsy samples. All in all, before implementing the ESCGP signature in clinical practice, it has to be validated in an independent data set using sample material readily available in pathology laboratories. Such a validation study has been initiated in our laboratory.

In conclusion, the ESCGP signature is a promising biomarker combination suitable for estimating the survival of PCa patients. After validation in an independent large cohort study, it would provide an important and orthogonal complement to the currently clinical parameters routinely used in the process of treatment decision for individual patients, in particular for early-stage PCas.

CONFLICT OF INTEREST

CL is one of the co-founders and a stockowner of Chundsell Medicals AB, an innovative start-up company financed by both private investors and the governmental innovation support agency Almi aiming to develop a molecular test kit based on the gene expression signature identified in this study. ZP started as a scientific consultant of this company after the manuscript was finished. These innovative associations have no influence on the scientific research ethics and quality of this study. The other authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the American Journal Experts and Dr Nigel Tooke for providing professional English editing of the manuscript. This work was supported by The Swedish Cancer Society, the Stockholm County Council, Konung Gustaf V:s Jubileumsfond, Cancer och Allergifonden, the KID funding of Karolinska Institutet and the Swedish Society of Medicine.

REFERENCES

1 Albertsen PC. Treatment of localized prostate cancer: when is active surveillance appropriate? *Nat Rev Clin Oncol* 2010; **7**: 394–400.

2 Hellerstedt BA, Pienta KJ. The current state of hormonal therapy for prostate cancer. *CA Cancer J Clin* 2002; **52**: 154–179.

3 Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. *Cancer* 2008; **113**: 3075–3099.

4 Wilkins A, Parker C. Treating prostate cancer with radiotherapy. *Nat Rev Clin Oncol* 2010; **7**: 583–589.

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

6 Cuzick J, Fisher G, Kattan MW, Berney D, Oliver T, Foster CS *et al.* Long-term outcome among men with conservatively treated localised prostate cancer. *Br J Cancer* 2006; **95**: 1186–1194.

7 Cheville JC, Karnes RJ, Therneau TM, Kosari F, Munz JM, Tillmans L *et al.* Gene panel model predictive of outcome in men at high-risk of systemic progression and death from prostate cancer after radical retropubic prostatectomy. *J Clin Oncol* 2008; **26**: 3930–3936.

8 Glinsky GV. 'Stemness' genomics law governs clinical behavior of human cancer: implications for decision making in disease management. *J Clin Oncol* 2008; **26**: 2846–2853.

9 Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* 2004; **101**: 811–816.

10 Liu W, Xie CC, Thomas CY, Kim ST, Lindberg J, Egevad L *et al.* Genetic markers associated with early cancer-specific mortality following prostatectomy. *Cancer* 2013; **119**: 2405–2412.

11 Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y *et al.* Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genom* 2010; **3**: 8.

12 Clevers H. Stem cells, asymmetric division and cancer. *Nat Genet* 2005; **37**: 1027–1028.

13 Ratajczak MZ, Shin DM, Liu R, Marlicz W, Tarnowski M, Ratajczak J *et al.* Epiblast/germ line hypothesis of cancer development revisited: lesson from the presence of Oct-4+ cells in adult tissues. *Stem Cell Rev* 2010; **6**: 307–316.

14 Visvader JE. Cells of origin in cancer. *Nature* 2011; **469**: 314–322.

15 Sperger JM, Chen X, Draper JS, Antosiewicz JE, Chon CH, Jones SB *et al.* Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. *Proc Natl Acad Sci USA* 2003; **100**: 13350–13355.

16 Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, Sethuraman A *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biol* 2005; **6**: R22.

17 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001; **98**: 5116–5121.

18 Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* 2008; **3**: 1101–1108.

19 Wittwer CT, Herrmann MG, Gundry CN, Elenitoba-Johnson KS. Real-time multiplex PCR assays. *Methods* 2001; **25**: 430–442.

20 Andersson L, Hagmar B, Ljung BM, Skoog L. Fine needle aspiration biopsy for diagnosis and follow-up of prostate cancer. Consensus Conference on Diagnosis

and Prognostic Parameters in Localized Prostate Cancer. Stockholm, Sweden, May 12–13, 1993. *Scand J Urol Nephrol Suppl* 1994; **162**: 43–49, discussion 115-27.

21 World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*. World Health Organization: Geneva, Switzerland, 1992.

22 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.

23 de Melo Martins PC, Parise Junior O, Pereira Hors C, Villela Miguel RE, da Costa Andrade VC, Garicochea B. C8orf4/TC-1 (thyroid cancer-1) gene expression in thyroid cancer and goiter. *J Otorhinolaryngol Relat Spec* 2007; **69**: 127–130.

24 Kasthuri RS, Taubman MB, Mackman N. Role of tissue factor in cancer. *J Clin Oncol* 2009; **27**: 4834–4838.

25 Li C, Berx G, Larsson C, Auer G, Aspenblad U, Pan Y *et al*. Distinct deleted regions on chromosome segment 16q23–24 associated with metastases in prostate cancer. *Genes Chromosomes Cancer* 1999; **24**: 175–182.

26 Lin SH, Cheng CJ, Lee YC, Ye X, Tsai WW, Kim J *et al*. A 45-kDa ErbB3 secreted by prostate cancer cells promotes bone formation. *Oncogene* 2008; **27**: 5195–5203.

27 Mackman N. The many faces of tissue factor. *J Thromb Haemost* 2009; **7**(Suppl 1): 136–139.

28 Mehta HH, Gao Q, Galet C, Paharkova V, Wan J, Said J *et al*. IGFBP-3 is a metastasis suppression gene in prostate cancer. *Cancer Res* 2011; **71**: 5154–5163.

29 Rubin MA, Zhou M, Dhanasekaran SM, Varambally S, Barrette TR, Sanda MG *et al*. Alpha-methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA* 2002; **287**: 1662–1670.

30 Saitoh T, Katoh M. Expression and regulation of WNT5A and WNT5B in human cancer: up-regulation of WNT5A by TNFalpha in MKN45 cells and up-regulation of WNT5B by beta-estradiol in MCF-7 cells. *Int J Mol Med* 2002; **10**: 345–349.

31 Strawbridge RJ, Nister M, Brismar K, Gronberg H, Li C. MUC1 as a putative prognostic marker for prostate cancer. *Biomark Insights* 2008; **3**: 303–315.

32 Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG *et al*. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 2002; **419**: 624–629.

33 Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H *et al*. *k*-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenom J* 2010; **10**: 292–309.

34 David W, Hosmer J, Stanley L, Susanne M. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 2nd edn (Wiley: New York), 2008, pp 244–285.

35 Teeter AE, Presti Jr. JC, Aronson WJ, Terris MK, Kane CJ, Amling CL *et al*. Do nomograms designed to predict biochemical recurrence (BCR) do a better job of predicting more clinically relevant prostate cancer outcomes than BCR? A report from the SEARCH database group. *Urology* 2013; **82**: 53–59.

36 Epstein MM, Edgren G, Rider JR, Mucci LA, Adami HO. Temporal trends in cause of death among Swedish and US men with prostate cancer. *J Natl Cancer Inst* 2012; **104**: 1335–1342.

37 Bast Jr. RC, Lilja H, Urban N, Rimm DL, Fritsche H, Gray J *et al*. Translational crossroads for biomarkers. *Clin Cancer Res* 2005; **11**: 6103–6108.

38 Yeap BB, Chubb SA, McCaul KA, Ho KK, Hankey GJ, Norman PE *et al*. Associations of IGF1 and IGFBPs 1 and 3 with all-cause and cardiovascular mortality in older men: the Health In Men Study. *Eur J Endocrinol* 2011; **164**: 715–723.

39 Brodin E, Vikan T, Hansen JB, Svartberg J. Testosterone, hemostasis, and cardiovascular diseases in men. *Semin Thromb Hemost* 2011; **37**: 87–94.

40 Mitchell S, Abel P, Madaan S, Jeffs J, Chaudhary K, Stamp G *et al*. Androgen-dependent regulation of human MUC1 mucin expression. *Neoplasia* 2002; **4**: 9–18.

Supplementary Information accompanies the paper on the Prostate Cancer and Prostatic Diseases website (http://www.nature.com/pcan)