Supporting Information

Wu et al. 10.1073/pnas.1215870110

SI Materials and Methods

Tissue Processing. Following tumor tissue isolation by macrodissection or laser capture microdissection, tissue samples were subjected to overnight proteinase K digestion.

For the cDNA-mediated annealing, selection, extension, and ligation (DASL) bead microarray study, total RNA was isolated and purified using the High Pure RNA Isolation Kit (Roche) and cDNA was generated using the single-use cDNA Synthesis Kit (Illumina).

For real-time RT-PCR studies, RNA was isolated using an RNA clean-up kit (Zymo Research). Purified RNA was converted to cDNA using a mixture of random hexamers and oligo-deoxythymine (dT) with SuperScript III RT (Invitrogen). The resultant cDNA then was preamplified using the TaqMan Pre-amp Master Mix (ABI/Life Technologies). PCR assays were performed using TaqMan minor groove binder (MGB) probes and primers obtained from Integrated DNA Technologies.

DASL Bead Microarray Analysis. cDNA sequences for the 1,536 genes were generated, and a custom DASL assay panel was synthesized by Illumina. The Illumina DASL Cancer Panel was not used in this study. The DASL assay was performed at the Mayo Clinic College of Medicine Genotyping Shared Resource (Rochester, MN). The DASL array dataset is available online (www.ncbi.nlm.nih.gov/projects/geo; accession no. GSE44353).

The discovery cohort was divided into a training set and a test set. Sixty samples with long-term follow-up were selected for DASL-mediated gene selection; 30 of the samples were from patients who experienced biochemical failure (BCF) during the follow-up period and 30 were from patients who did not experience BCF. Random forests classification (1) was applied to classify the DASL bead microarray data from the 60 selected samples. Random forests is well suited for classification of large datasets, such as microarrays, because it allows examination of collective expression of sets of genes and maintains accuracy even when most variables are noise and when the number of variables (genes) are greater than the number of observations (samples) (2). Briefly, random forest classifications were run 500 times, and 500 sets of genes were selected from the 1,536-gene set. For every group of selected genes, a risk index was built that classified the sample into high risk or low risk for BCF using the median as a cut point. The -log(P value) was calculated based on a log-rank test. As the gene number in the selected gene sets increased, the -log(P value) tended to increase, and when the gene number reached 32, it stabilized. From the 500 independent random forest runs, 7 runs included 32-gene classifiers—the set of gene classifiers with the most significant predictive value for BCF. These seven sets of 32-gene classifiers were combined, resulting in a 62-gene signature chosen for further development. Finally, using the training set, a risk model was built based on this compact gene set and tested on the test set.

Gene Set Enrichment via Real-Time RT-PCR. The DASL microarray-derived gene expression signature was converted to a real-time RT-PCR platform for two reasons: (i) to eliminate potential false positive signals from the DASL assay because gene expression data are not always generalizable across platforms, and (ii) because a PCR platform is desirable because of its increased sensitivity, larger dynamic range, and facilitation of clinical adoption.

From the training set (Fig. S1), the genes most predictive of BCF were selected using a supervised principal components method (3). A univariate Cox regression model was used, and coefficients for each gene were calculated. Twenty-nine genes whose coefficients were above a cross-validation–determined threshold were selected and used to calculate the principal components. A risk model was constructed based on this gene set and used to predict performance of the test set.

Incremental Value Compared with Nomograms. Predicted freedom from biochemical recurrence scores was generated using three different postoperative nomograms (4–6). To compute predictions from these nomograms, some modifications were made on the original nomograms based on variable coding limitations in the training set. From the Kattan nomogram (4), capsular invasion was collapsed into "yes" or "no" instead of "none," "invades capsule," "focal," and "established," and from the Kattan nomogram calculation (6), the model without surgeon experience was used.

For each nomogram, two separate Cox proportional hazard models were created using the training cohort: one with the nomogram-predicted risk score as the only predictor, and the other including both nomogram risk score and 32-gene risk index score. Restricted cubic splines were used for both continuous predictors to accommodate potential nonlinear relationships. Then, probabilities of freedom from recurrence were calculated for patients in the validation set for each of the six Cox models. The discrimination of each model was quantified by the concordance index, which is identical to the nonparametric area under the ROC curve in the binary setting.

SI Results

Identification of Gene Signature. To identify a prognostic genetic signature for patients with prostate cancer, we first performed gene expression profiling on formalin-fixed and paraffin-embedded radical prostatectomy (RP) tissue samples using a customized 1,536-gene DASL assay. A gene expression signature predictive of BCF was selected using random forest classification methodology. Seven sets of 32-gene classifiers were selected based on their predictive ability and combined, resulting in a 62-gene signature. A composite risk score was built using these 62 genes and coefficients from a univariate Cox regression model on the 124 training subset samples. The risk model was used to predict performance on the test subset.

Individual real-time PCR assays were set up for the 62 genes selected by the random forest method. We enriched the compact gene set by performing a supervised principal components analysis using PCR-based gene expression data. Twenty-nine genes were selected by the supervised principal component analysis. The 29 genes included in the risk model (Table S2) are from multiple functional families, including transcription factors, cell cycle genes, metabolic genes, and genes with unknown functions. In addition, three reference genes with relatively invariant expression across tissue samples were identified and selected for input normalization. Thus, the final assay included 32 genes. A risk model was constructed based on this 29-gene set and used to predict performance of the test subset.

Validation Cohort: Individual Genes Cox Regression Analysis. Cox regression analysis of the individual genes in the risk index demonstrated that 24 of the 29 informative genes individually were statistically significant ($P < 0.05$) predictors of BCF.

Validation Cohort: Subgroup Analyses. Ten-year BCF probabilities in additional subgroups of patients, stratified by Gleason scores, are presented in Fig. S2. Ten-year BCF probabilities in additional

subgroups of patients, stratified by pathologic tumor (pT) scores and surgical margin status, are presented in Fig. S3.

Validation Cohort: Univariate and Multivariate Analysis Excluding Patients with Missing Prostate-Specific Antigen Values. Separate univariate and multivariate analyses excluding the 36 patients with missing presurgery prostate-specific antigen (PSA) values (rather than imputing values) are presented in Tables S3–S5. The 32-gene risk index ($P = 0.0162$) and surgical margin ($P = 0.0052$) were the only significant prognostic factors in this analysis. Demographics and patient characteristics for the full validation cohort ($n = 270$), the validation cohort excluding patients with missing presurgery PSA values ($n = 234$), and the patients with missing PSA values ($n = 34$) are presented in Table S8.

Validation Cohort: Biopsy Subset. Demographics and patient characteristics for the subset of patients in the validation cohort with available needle biopsy tissue samples are presented in Table S9.

Validation Cohort: Multivariate Analysis with the 32-Gene Risk Index and Three Postoperative Nomograms. Multivariate analyses assessing the contribution of the 32-gene risk index with each of the three postoperative nomograms are presented in Table S6. These analyses demonstrate that the 32-gene risk index and the nomograms provide independent prognostic information.

1. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.
2. Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3.
3. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2(4):E108.
4. Kattan MW, Wheeler TM, Scardino PT (1999) Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol* 17(5):1499–1507.
5. Stephenson AJ, et al. (2005) Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *J Clin Oncol* 23(28):7005–7012.
6. Kattan MW, et al. (2009) Preoperative and postoperative nomograms incorporating surgeon experience for clinically localized prostate cancer. *Cancer* 115(5):1005–1010.

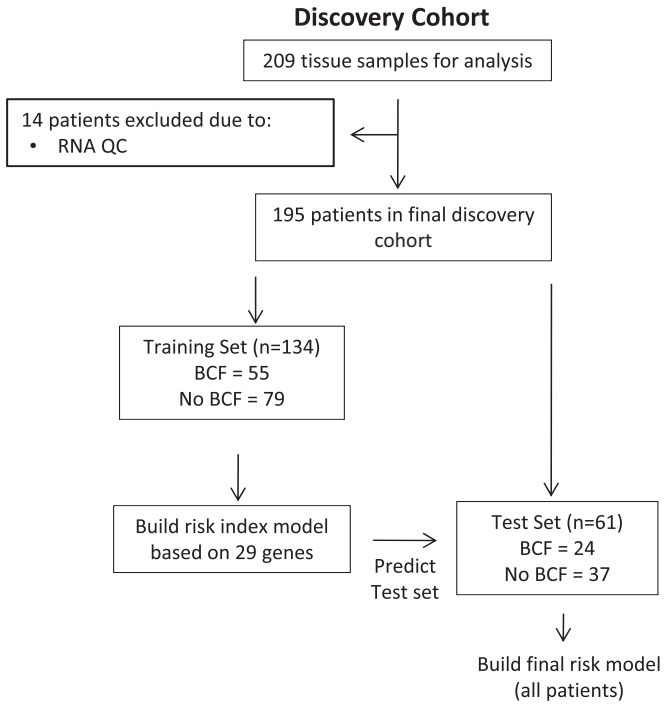

Fig. S1. Training cohort flow chart. Tissue samples for the discovery cohort were obtained from consecutive patients who underwent RP surgery as part of their clinical care at Massachusetts General Hospital from September 1993 to September 1995.

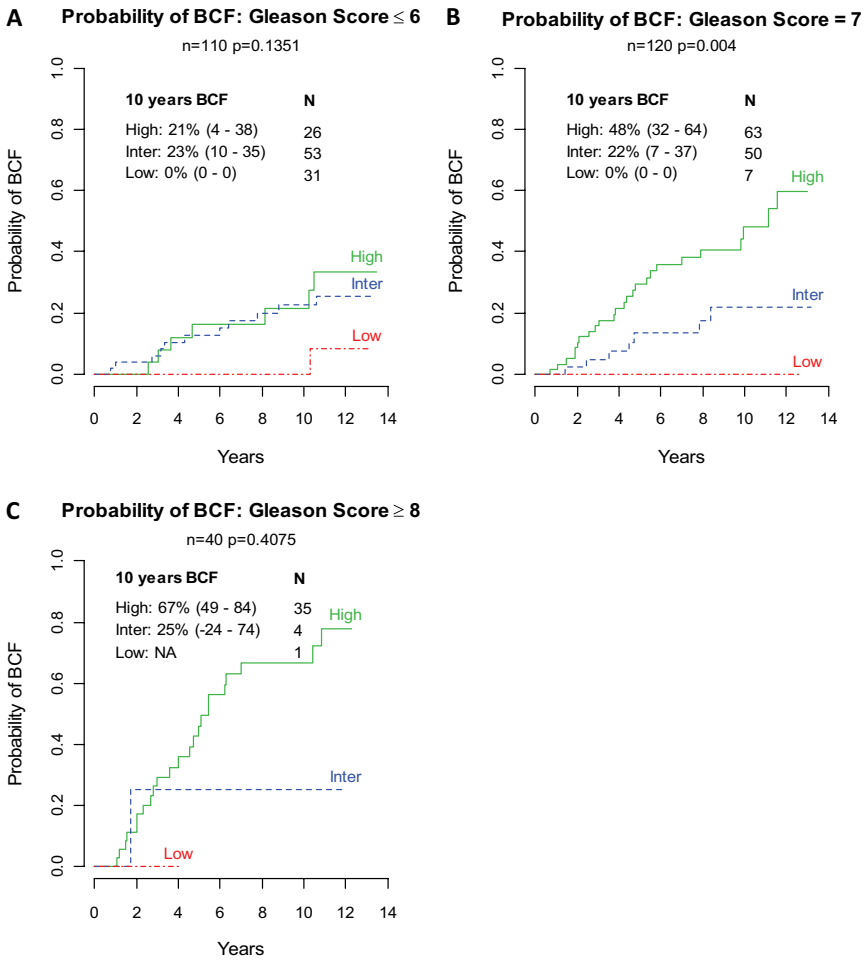

Fig. S2. Cumulative incidence of BCF by Gleason score (validation cohort). Cumulative incidence of BCF in patients with Gleason scores ≤6 (*A*), 7 (*B*), and ≥8 (*C*).

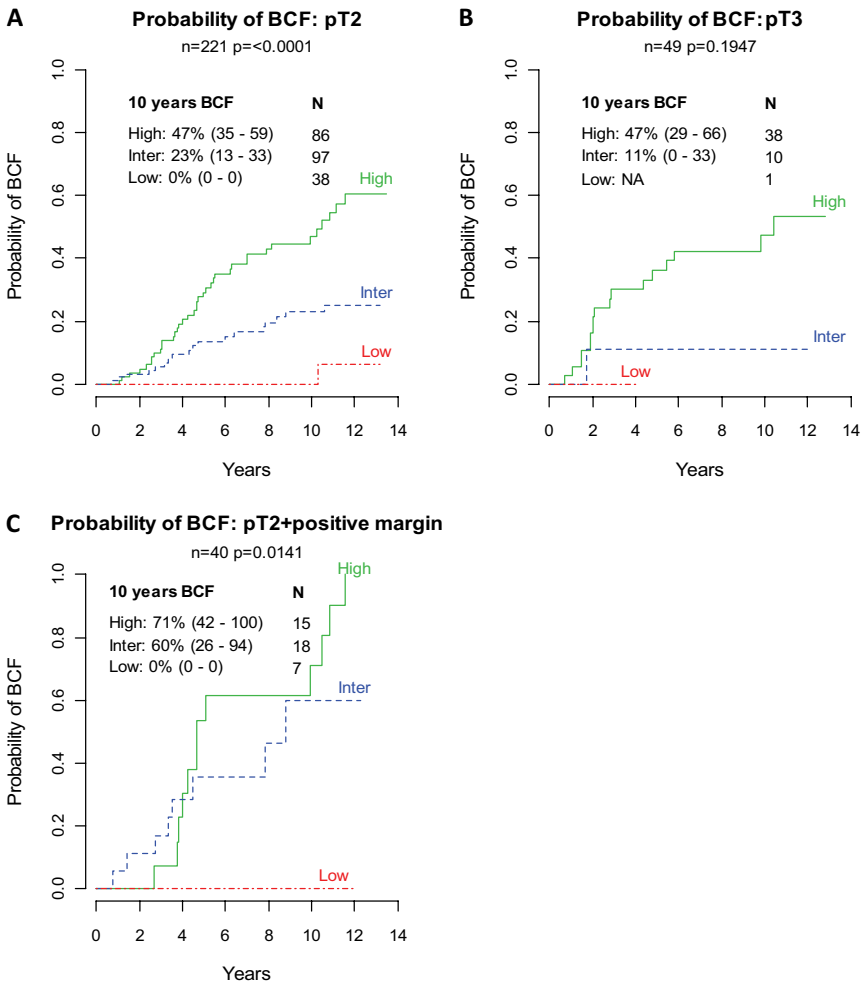

Fig. S3. Cumulative incidence of BCF by pathologic stage and surgical margin (validation cohort). Cumulative incidence of BCF in patients with pT2 (*A*), pT3 (*B*), and pT2 with a positive surgical margin (*C*).

Table S1. Demographics and patient characteristics for the discovery cohort

Characteristic	Discovery cohort (n = 195)
Age, y	
Mean (SD)	61.80 (6.05)
Range	45.0–77.0
Presurgery PSA, ng/mL	
Mean (SD)	8.39 (6.24)
Range	0.50–37.20
PSA unknown, no. (%)	24 (12.3)
Gleason score at RP, no. (%)	
≤6	76 (39)
7	94 (48)
≥8	25 (13)
pT stage, no. (%)	
2	148 (76)
3	47 (24)
Surgical margin, no. (%)	
Negative	115 (59)
Positive	80 (41)
Salvage therapy after BCF, no. (%)	
No	131 (67)
Yes	64 (33)
BCF event, no.	
No	116
Yes	79
Time to BCF event, y, median (range)	3.10 (0.06–12.58)
Follow-up time, no BCF event, y, median (range)	9.39 (0.23–15.81)
Metastasis event, no.	
No	171
Yes	22
Unknown	2
Time to metastasis event, y, median (range)	7.83 (0.15–14.70)
Follow-up time, no metastasis event, y, median (range)	12.89 (0.25–16.27)
Death (all-cause) event, no.	
No	168
Yes	27
Time to death (all-cause), y, median (range)	11.41 (3.65–15.67)
Follow-up time, no death event, y, median (range)	13.65 (0.25–16.27)

Table S2. Genes included in the risk index (excludes three normalization genes)

Symbol	Full name
ACTR3B	ARP3 actin-related protein 3 homolog B
APOC1	Apolipoprotein C-I
ATP8A1	Aminophospholipid transporter (APLT), class I, type 8A, member 1
C10orf116 (aka APM2)	Chromosome 10 ORF 116
Cdk1 (aka CDC2)	Cyclin-dependent kinase 1
CDON	Cdon homolog, cell adhesion molecule-related/down-regulated by oncogenes
DEGS1	Degenerative spermatocyte homolog 1, lipid desaturase
SYMN (aka DMN)	Intermediate filament protein
DPP4	Dipeptidyl-peptidase 4
F12	Coagulation factor XII
FEV	FEV (ETS oncogene family)
GATA3	GATA binding protein 3
GSTM3	GST mu 3
HIST1H3H	Histone cluster 1, H3h
HOXC4	Homeobox C4
TMEM132A (aka HSPA5BP1)	Transmembrane protein 132A
IGSF1	Imunoglobulin superfamily, member 1
IGSF6	Imunoglobulin superfamily, member 6
INHBA	Inhibin, beta A
KRT15	Keratin 15
LDHB	Lactate dehydrogenase B
KIF18B (aka LOC146909)	Kinesin family member 18B
NCAPG2 (aka LUZP5)	Non-SMC condensin II complex, subunit G2
MAOA	Monoamine oxidase A
MT1F	Metallothionein 1F
OIP5	Opa interacting protein 5
PPP3CB	Protein phosphatase 3, catalytic subunit, beta isozyme
QPRT	Quinolinate phosphoribosyltransferase
TPX2	TPX2, microtubule-associated, homolog

Table S3. Univariate Cox regression analysis of BCF in the validation cohort excluding patients with missing presurgery PSA values (n = 270)

	HR (95% CI)	*P* value
32-Gene risk index	7.57 (3.55–16.15)	<0.0001
Gleason score		<0.0001
7 vs. ≤6	1.91 (1.08–3.38)	0.03
≥8 vs. ≤6	5.07 (2.76–9.31)	<0.0001
pT stage (3 vs. 2)	1.62 (0.95–2.75)	0.08
Margin (pos. vs. neg.)	2.40 (1.50–3.83)	0.0002
Log(1+baseline PSA) (n = 234)	1.70 (1.20–2.40)	0.002

The hazard ratio (HR) calculated for the 32-gene index is based on a five-unit change. CI, confidence interval.

Table S4. Multivariate Cox regression analysis (with 32-gene risk index) of BCF in the validation cohort excluding patients with missing presurgery PSA values (n = 234)

	HR (95% CI)	*P* value
32-Gene risk index	3.74 (1.28–10.96)	0.0162
Gleason score		0.1894
7 vs. ≤6	1.44 (0.75–2.76)	0.2684
≥8 vs. ≤6	2.22 (0.94–5.23)	0.0683
pT stage (3 vs. 2)	0.82 (0.45–1.49)	0.5086
Margin (pos. vs. neg.)	2.15 (1.26–3.69)	0.0052
Log(1+baseline PSA)	1.23 (0.85–1.78)	0.2683

The hazard ratio (HR) calculated for the 32-gene index is based on a five-unit change. CI, confidence interval.

Table S5. Multivariate Cox regression analysis (without 32-gene risk index) of BCF in the validation cohort excluding patients with missing presurgery PSA values (n = 234)

	HR (95% CI)	*P* value
Gleason score		0.0003
7 vs. ≤6	1.59 (0.84–3.03)	0.1574
≥8 vs. ≤6	3.89 (1.94–7.79)	0.0001
pT stage (3 vs. 2)	0.92 (0.51–1.67)	0.7927
Margin (pos. vs. neg.)	1.98 (1.17–3.36)	0.0109
Log(1+baseline PSA)	1.35 (0.94–1.94)	0.1048

The hazard ratio (HR) calculated for the 32-gene index is based on a five-unit change. CI, confidence interval.

Table S6. Multivariate Cox regression analyses of BCF in the validation cohort with the 32-gene risk index and three postoperative nomograms

Variable	HR (95% CI)	*P* value
32-Gene risk index	4.09 (1.64–10.19)	0.0025
Nomogram 1999 (4)	2.31 (1.27–4.19)	0.0060
32-Gene risk index	3.53 (1.27–9.82)	0.0159
Nomogram 2005 (5)	2.03 (1.14–3.63)	0.0168
32-Gene risk index	3.38 (1.31–8.73)	0.0118
Nomogram 2009 (6)	2.53 (1.38–4.61)	0.0025

Validation cohort excludes patients with missing presurgery PSA values (n = 234). The hazard ratio (HR) calculated for the 32-gene index is based on a five-unit change. The HR calculated for the nomograms was based on 50% change. CI, confidence interval.

Table S7. Net reclassification improvement comparing the 32-gene risk index with the three postoperative nomograms

Event	Nomogram*	NRI (95% CI)	*P* value
Metastasis	1999 (4)	0.223 (0.005, 0.440)	0.045
	2005 (5)	0.146 (−0.073, 0.364)	0.191
	2009 (6)	0.005 (−0.215, 0.224)	0.968
BCF	1999 (4)	0.141 (−0.244, 0.306)	0.095
	2005 (5)	0.104 (−0.065, 0.273)	0.226
	2009 (6)	0.027 (-0.151, 0.205)	0.765

CI, confidence interval; NRI, net reclassification improvement.
*Compared with 32-gene risk index.

Table S8. Demographics and patient characteristics for three cohorts

Characteristic	Full cohort (n = 270)	Validation cohort	
		Cohort excluding patients with no pre-RP PSA (n = 234)	Patients with no pre-RP PSA (n = 36)
Age, y			
Mean (SD)	61.96 (7.06)	62.29 (7.11)	59.78 (6.42)
Range	37.00–79.00	37.00–79.00	50–77
Presurgery PSA, ng/mL			
Mean (SD)	7.06 (5.68)	7.06 (5.68)	NA
Range	0.80–52.4	0.80–52.4	NA
PSA Unknown, no. (%)	36 (13.3%)	0	36
Gleason score at RP, no. (%)			
≤6	110 (41%)	94 (40%)	16 (44%)
7	120 (44%)	106 (45%)	14 (39%)
≥8	40 (15%)	34 (15%)	6 (17%)
pT stage, no. (%)			
2	221 (82%)	189 (81%)	32 (89%)
3	49 (18%)	45 (19%)	4 (11%)
Surgical margin, no. (%)			
Negative	204 (76%)	177 (76%)	27 (75%)
Positive	66 (24%)	57 (24%)	9 (25%)
Salvage therapy after BCF, no (%)			
No	225 (83.3%)	195 (83.3%)	30 (83.3%)
Yes	45 (16.7%)	39 (16.7%)	6 (16.7%)
BCF event, no.			
No	195	169	26
Yes	75	65	10
Time to BCF event, y, median (range)	4.28 (0.70–11.54)	4.53 (0.70–11.54)	2.47 (0.75–5.46)
Follow-up time, no BCF event, y, median (range)	9.43 (0.53–13.49)	9.55 (0.53–13.49)	5.10 (0.96–12.81)
Metastasis event, no.			
No	253	220	33
Yes	17	14	3
Time to metastasis event, y, median (range)	7.54 (2.09–12.96)	7.30 (2.09–12.96)	9.78 (7.16–11.65)
Follow-up time, no metastasis event, y, median (range)	10.35 (0.47–14.67)	10.60 (0.47–14.67)	7.02 (0.96–13.78)
Death (all-cause) event, no.			
No	229	198	31
Yes	41	36	5
Time to death (all-cause), y, median (range)	9.84 (1.41–13.13)	10.02 (1.41–13.13)	8.42 (3.03–12.00)
Follow-up time, no death event, y, median (range)	12.74 (10.74–14.18)	12.73 (10.74–14.18)	12.85 (11.62–13.73)

NA, not available.

Table S9. Demographics and patient characteristics for the subset of patients in the validation cohort with available needle biopsy tissue samples (n = 79)

Characteristic	Biopsy cohort (n = 79)	Validation cohort (n = 270)
Age, y		
Mean (SD)	61.91 (7.41)	61.96 (7.06)
Range	37.00–79.00	37.00–79.00
Presurgery PSA, ng/mL		
Mean (SD)	6.68 (6.12)	7.06 (5.68)
Range	0.90–49.20	0.80–52.4
PSA unknown, no. (%)	3 (3.8%)	36 (13.3%)
Gleason score at RP, no. (%)		
≤6	32 (41%)	110 (41%)
7	35 (44%)	120 (44%)
≥8	12 (15%)	40 (15%)
Gleason score at biopsy, no. (%)		
≤6	45 (57%)	NA
7	27 (34%)	NA
≥8	7 (9%)	NA
cT stage, no. (%)		
1	70 (89%)	NA
2	9 (11%)	NA
pT stage, no. (%)		
2	64 (81%)	221 (82%)
3	15 (19%)	49 (18%)
Proportion of cores positive, no. (%)		
≤ One-third	49 (62%)	NA
> One-third to < two-thirds	30 (38%)	NA
Highest positive core, no. (%)		
<25%	23 (29%)	NA
25–50%	21 (27%)	NA
51–75%	20 (25%)	NA
>75%	15 (19%)	NA
Surgical margin, no. (%)		
Negative	60 (76%)	204 (76%)
Positive	19 (24%)	66 (24%)
Salvage therapy after BCF, no (%)		
No	62 (87%)	225 (83.3%)
Yes	9 (13%)	45 (16.7%)
BCF event, no.		
No	52	195
Yes	27	75
Time to BCF event, y, median (range)	4.53 (0.70–10.58)	4.28 (0.70–11.54)
Follow-up time, no BCF event, y, median (range)	11.30 (1.14–13.16)	9.43 (0.53–13.49)
Metastasis event, no.		
No	73	253
Yes	6	17
Time to metastasis event, y, median (range)	7.23 (2.74–8.18)	7.54 (2.09–12.96)
Follow-up time, no metastasis event, y, median (range)	10.97 (1.09–14.67)	10.35 (0.47–14.67)
Death (all-cause) event, no.		
No	66	229
Yes	13	41
Time to death (all-cause), y, median (range)	10.44 (4.98–13.13)	9.84 (1.41–13.13)
Follow-up time, no death event, y, median (range)	12.83 (11.56–14.18)	12.74 (10.74–14.18)

cT, clinical tumor; NA, not available.