Supplementary Materials

Integration of Urinary EN2 Protein & Cell-Free RNA Data in the Development of a Multivariable Risk Model for the Detection of Prostate Cancer Prior to Biopsy

Shea P. Connell [1], Robert Mills [2], Hardev Pandha [3], Richard Morgan [4], Colin S. Cooper [1], Jeremy Clark [1] and Daniel S. Brewer [1,5,*]; The Movember GAP1 Urine Biomarker Consortium [†]

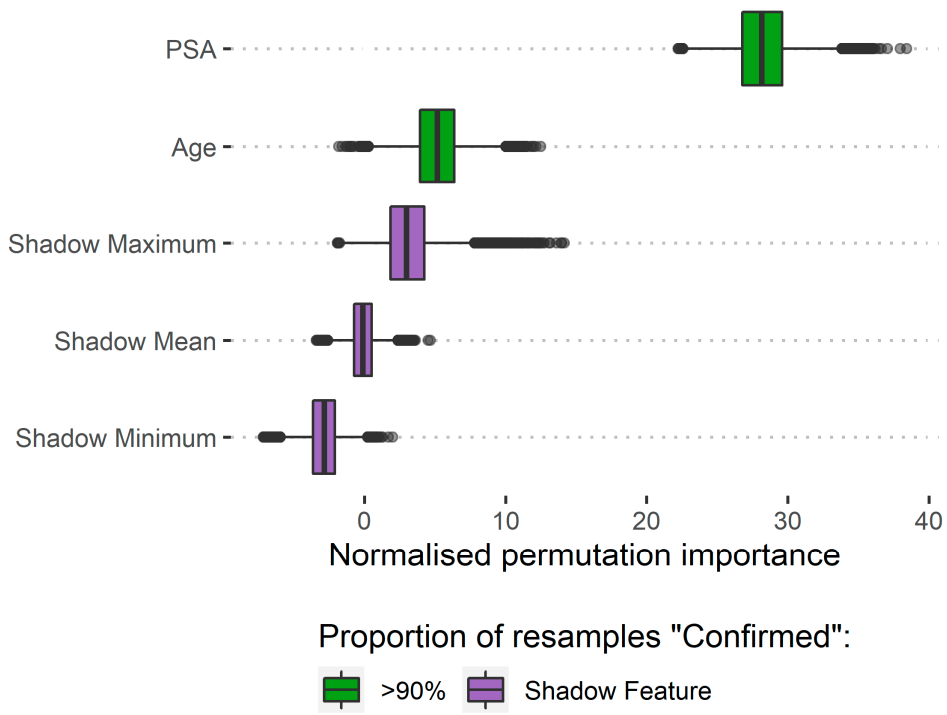

Figure S1. Boruta analysis of variables available for the training of the SoC model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models (Green).

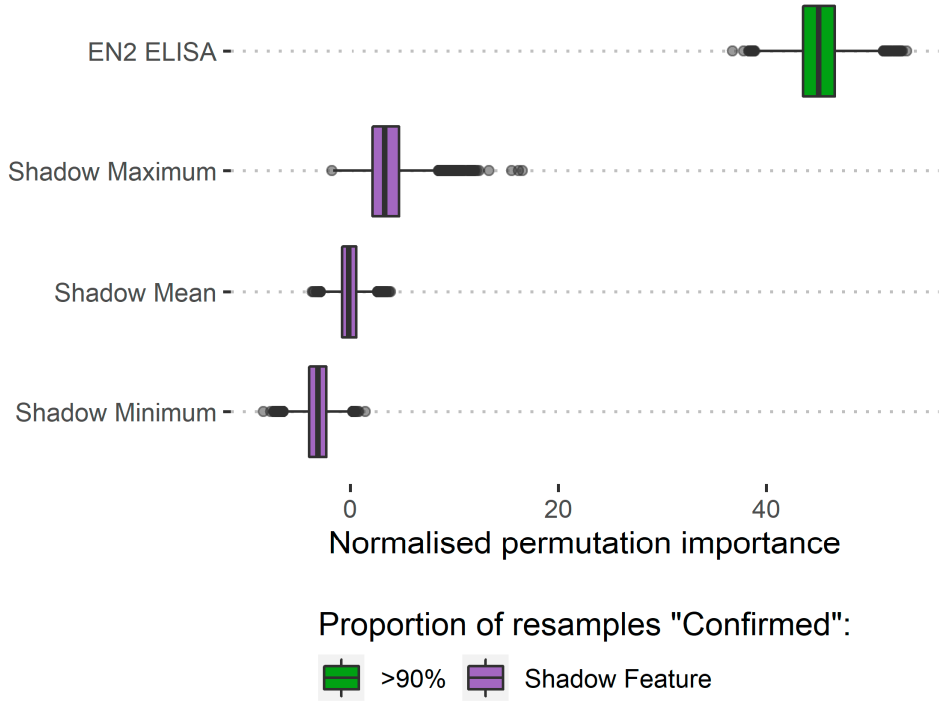

Figure S2. Boruta analysis of variables available for the training of the Engrailed model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models (Green).

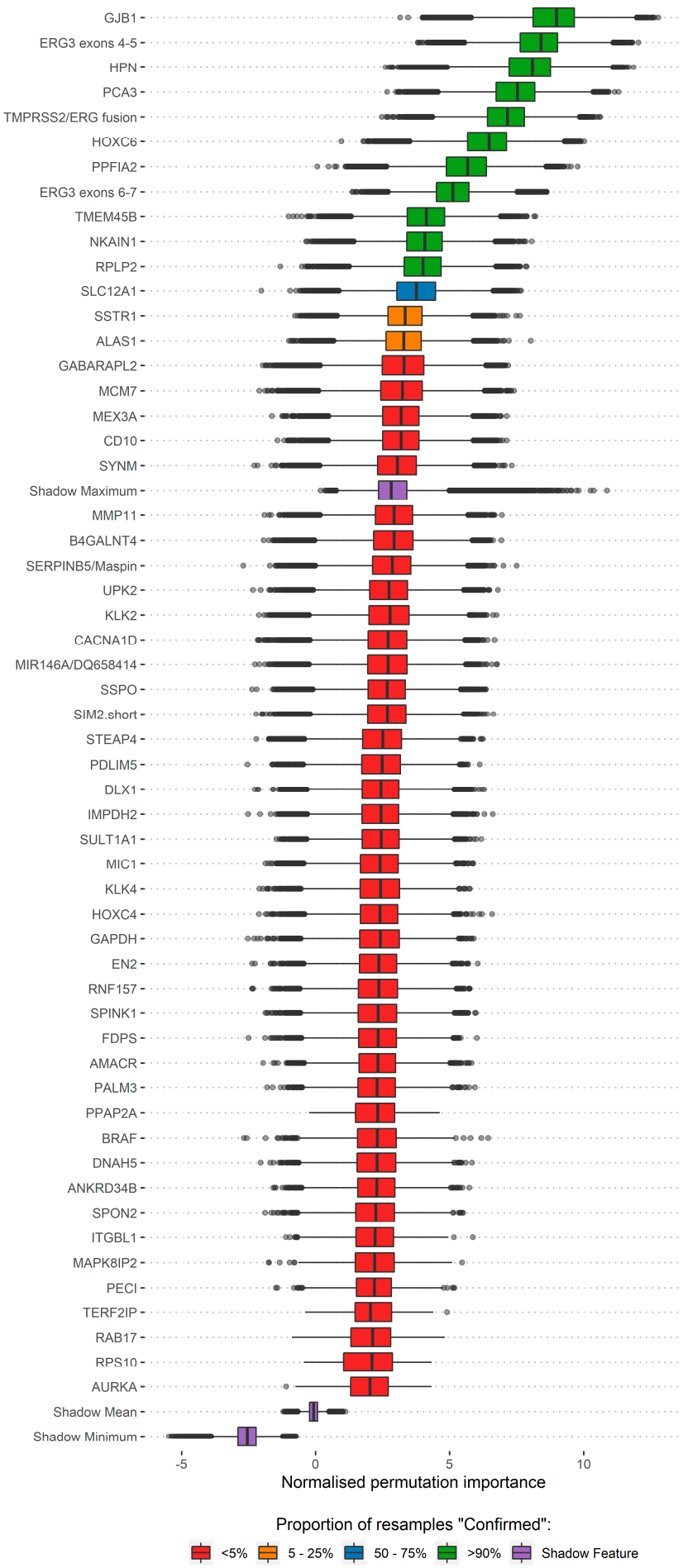
GJB1
ERG3 exons 4-5
HPN
PCA3
TMPRSS2/ERG fusion
HOXC6
PPFIA2
ERG3 exons 6-7
TMEM45B
NKAIN1
RPLP2
SLC12A1
SSTR1
ALAS1
GABARAPL2
MCM7
MEX3A
CD10
SYNM
Shadow Maximum
MMP11
B4GALNT4
SERPINB5/Maspin
UPK2
KLK2
CACNA1D
MIR146A/DQ658414
SSPO
SIM2.short
STEAP4
PDLIM5
DLX1
IMPDH2
SULT1A1
MIC1
KLK4
HOXC4
GAPDH
EN2
RNF157
SPINK1
FDPS
AMACR
PALM3
PPAP2A
BRAF
DNAH5
ANKRD34B
SPON2
ITGBL1
MAPK8IP2
PECI
TERF2IP
RAB17
RPS10
AURKA
Shadow Mean
Shadow Minimum
-5
0
5
10
Normalised permutation importance
Proportion of resamples "Confirmed":
<5%
5 - 25%
50 - 75%
>90%
Shadow Feature

Figure S3. Boruta analysis of variables available for the training of the ExoRNA model. Variable importance was determined over 1,000 bootstrap resamples of the available data and the decision reached recorded at each resample. Colour indicates the proportion of the 1,000 resamples a variable was confirmed to be important in. Variables confirmed in at least 90% of resamples were selected for training predictive models. Those variables rejected in every single resample are not shown here, but the full list of inputs for the ExoRNA model can be seen in Table S1.

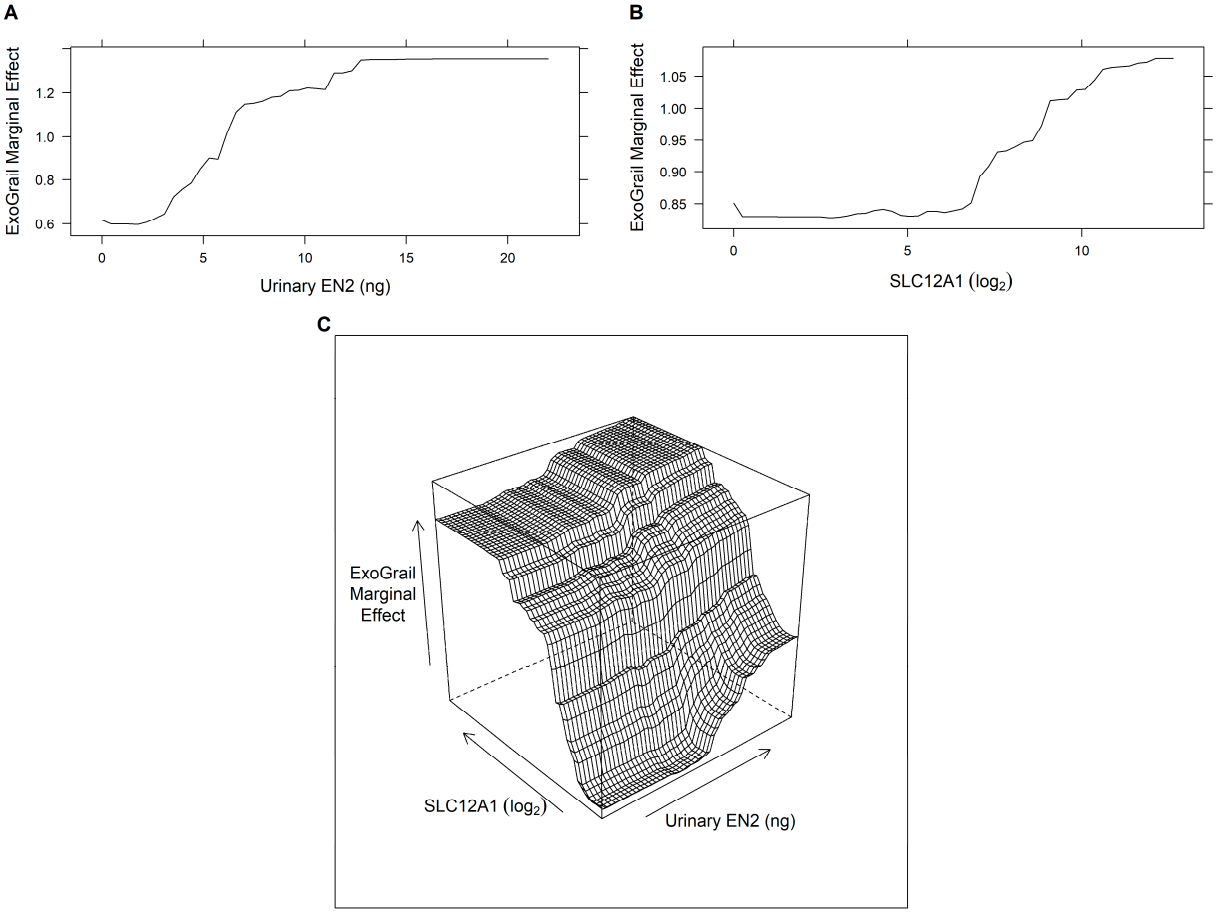

Figure S4 Partial dependency plots detailing the marginal effects and interactions of SLC12A1 and urinary EN2 on predicted ExoGrail Risk Score. **A** - Partial dependency of ExoGrail on urinary EN2, **B** - Partial dependency of ExoGrail on SLC12A1, **C** - Partial dependency of ExoGrail on both SLC12A1 and urinary EN2.

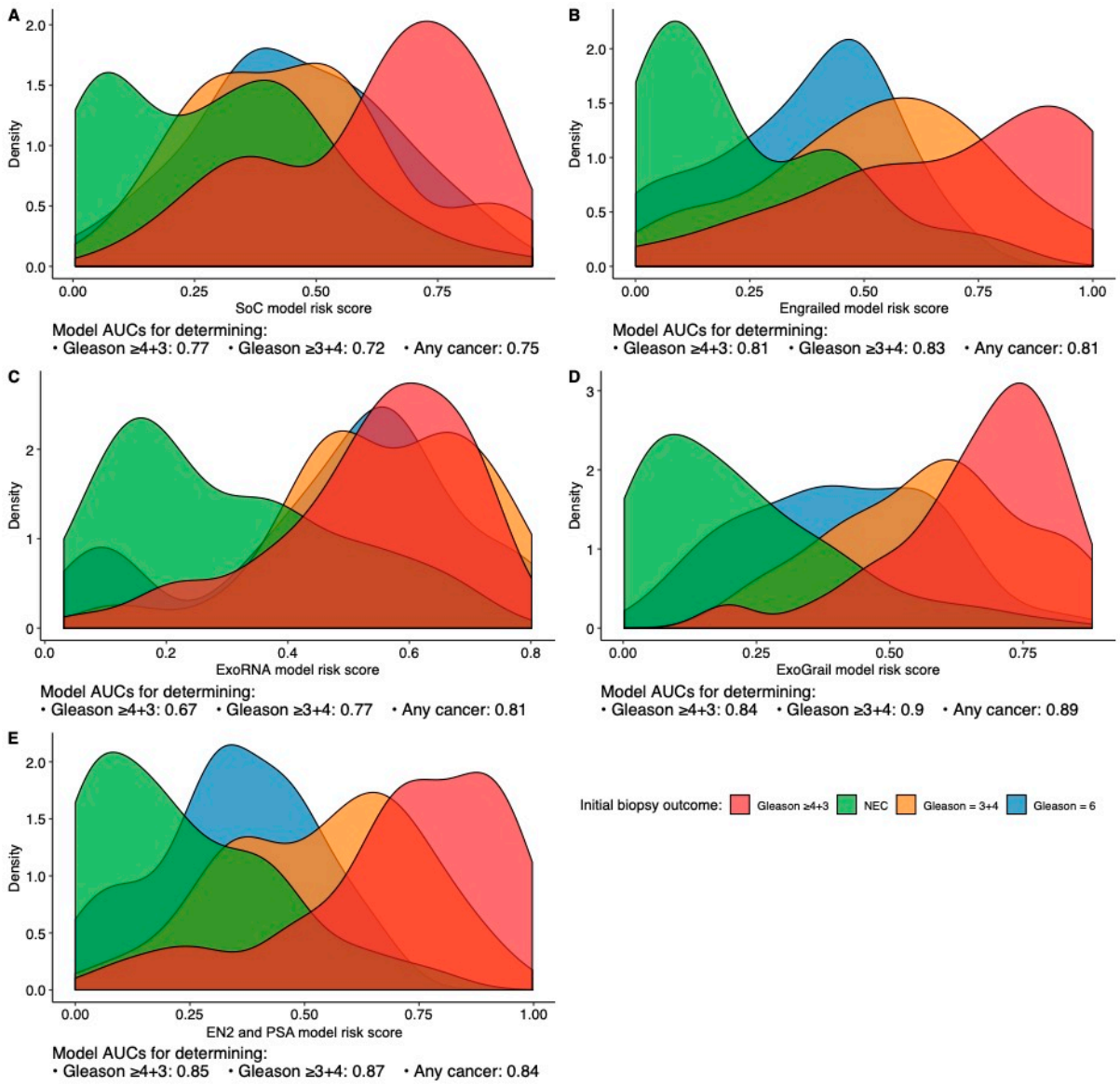

Figure S5. Risk score distributions of the trained models including the EN2 and PSA model, calculated as the out-of-bag predictions and represented as density plots. AUCs for each model's predictive ability for clinically relevant outcomes are detailed underneath each panel. Each random forest model was fit using differing input variables; A - SoC clinical risk model, including Age and PSA, B - Engrailed model, C -ExoRNA model and D - ExoGrail model, combining predictors from all three modes of analysis E -EN2 and PSA model. The full list of variables in each model is available in Table 1. Fill colour shows the risk score distribution of patients with respect to biopsy outcome: No evidence of cancer (Green), Gleason 6 (Blue), Gleason 3+4 (Orange), Gleason ≥ 4+3 (Red). Both the Engrailed model and the EN2&PSA model featured a bimodal risk score distribution in samples from men with no evidence of prostate cancer (NEC) (green density plots), falsely identifying some patients with NEC and low-grade disease as similar to those with more clinically significant G3+4 disease (orange density plots). The bimodal distribution of NEC is not a feature of the ExoGrail model and separation is improved. In the EN2+PSA model a bimodal distribution is now seen in Gleason 3+4 patients (E, orange density plot), falsely identifying some patients with high grade disease as similar to those with lower grade disease. Finally, the ExoGrail model also appears to improve delineation of the G≥4+3 patients manifesting as a more distinct peak in the red density plot of D compared to EN2 and EN2&PSA plots.

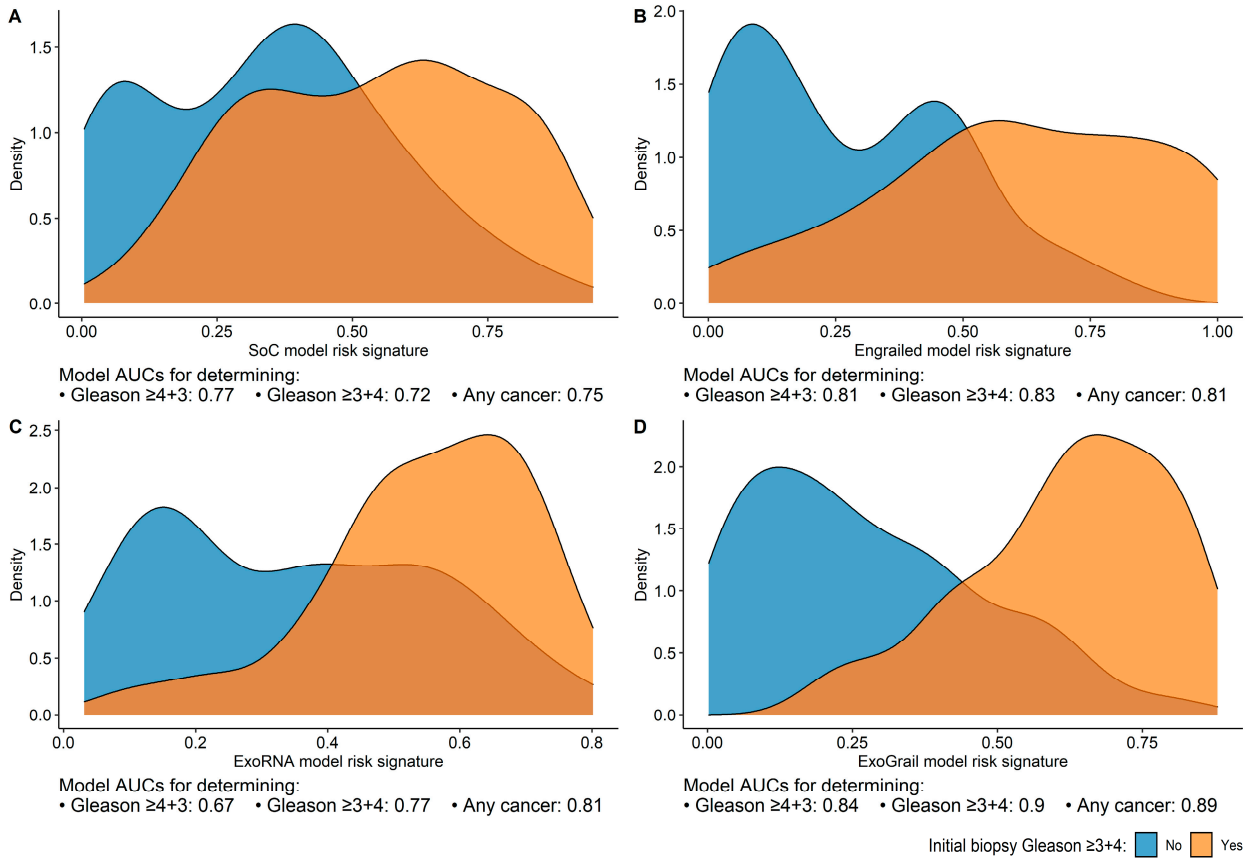

Figure S6. Density plots detailing risk score distributions generated from four trained models. Models A to D were trained with different input variables; **A** - SoC clinical risk model, including Age and PSA, **B** - Engrailed model, **C** -ExoRNA model and **D** - ExoGrail model, combining the predictors from all three previous models. The full list of variables in each model is available in Table 1. Fill colour shows the risk score distribution of patients with a significant biopsy outcome of Gs ≥ 3+4 (Orange) or Gs ≤ 6 (Blue)

Table S1. List of all features available for selection as input variables for each model prior to bootstrapped Boruta feature selection.

SoC	Engrailed	ExoRNA	ExoGrail
PSA	EN2 ELISA	March5	PSA
UrineVol		AATF	UrineVol
DRESize		ABCB9	DRESize
Age		ACTR5	Age
		AGR2	EN2 ELISA
		ALAS1	March5
		AMACR	AATF
		AMH	ABCB9
		ANKRD34B	ACTR5
		ANPEP	AGR2
		APOC1	ALAS1
		ARexon9	AMACR
		ARexons4.8	AMH
		ARHGEF25	ANKRD34B
		AURKA	ANPEP
		B2M	APOC1
		B4GALNT4	ARexon9
		BRAF	ARexons4.8
		BTG2	ARHGEF25
		CACNA1D	AURKA
		CADPS	B2M
		CAMK2N2	B4GALNT4
		CAMKK2	BRAF
		CASKIN1	BTG2
		CCDC88B	CACNA1D
		CD10	CADPS
		CDC20	CAMK2N2
		CDC37L1	CAMKK2
		CDKN3	CASKIN1

CKAP2L
CLIC2
CLU
COL10A1
COL9A2
CP
CTA.211A9.5.MIATNB
DLX1
DNAH5
DPP4
EIF2D
EN2
ERG3.exons.4.5
ERG3.exons.6.7
ERG5
FDPS
FOLH1.PSMA.NAALAD1
GABARAPL2
GAPDH
GCNT1
GJB1
GOLM1
HIST1H1C
HIST1H1E
HIST1H2BF
HIST1H2BG
HIST32HA
HMBS
HOXC4
HOXC6
HPN

CCDC88B
CD10
CDC20
CDC37L1
CDKN3
CKAP2L
CLIC2
CLU
COL10A1
COL9A2
CP
CTA.211A9.5.MIATNB
DLX1
DNAH5
DPP4
EIF2D
EN2
ERG3.exons.4.5
ERG3.exons.6.7
ERG5
FDPS
FOLH1.PSMA.NAALAD1
GABARAPL2
GAPDH
GCNT1
GJB1
GOLM1
HIST1H1C
HIST1H1E
HIST1H2BF
HIST1H2BG

HPRT
IFT57
IGFBP3
IMPDH2
ISX
ITGBL1
ITPR1
KLK2
KLK3.PSA.exons1.2
KLK3.PSA.exons2.3
KLK4
LASS1
LBH
MAK
MAPK8IP2
MCM7
MCTP1
MDK
MED4
MEMO1
Met
MEX3A
MFSD2A
MGAT5B
MIC1
MIR146A.DQ658414
MIR4435.1HG.lOC541471
MKi67
MMP11
MMP25
MMP26

HIST32HA
HMBS
HOXC4
HOXC6
HPN
HPRT
IFT57
IGFBP3
IMPDH2
ISX
ITGBL1
ITPR1
KLK2
KLK3.PSA.exons1.2
KLK3.PSA.exons2.3
KLK4
LASS1
LBH
MAK
MAPK8IP2
MCM7
MCTP1
MDK
MED4
MEMO1
Met
MEX3A
MFSD2A
MGAT5B
MIC1
MIR146A.DQ658414

MNX1

MSMB

MXI1

MYOF

NAALADL2

NEAT1

NKAIN1

NLRP3

OGT

OR52A2.PSGR

PALM3

PCA3

PCSK6

PDLIM5

PECI

PPAP2A

PPFIA2

PPP1R12B

PSTPIP1

PTN

PTPRC

PVT1

RAB17

RIOK3

RNF157

RP11.244H18.1.P712P

RP11.97O12.7

RPL18A

RPL23AP53

RPLP2

RPS10

MIR4435.1HG.lOC541471

MKi67

MMP11

MMP25

MMP26

MNX1

MSMB

MXI1

MYOF

NAALADL2

NEAT1

NKAIN1

NLRP3

OGT

OR52A2.PSGR

PALM3

PCA3

PCSK6

PDLIM5

PECI

PPAP2A

PPFIA2

PPP1R12B

PSTPIP1

PTN

PTPRC

PVT1

RAB17

RIOK3

RNF157

RP11.244H18.1.P712P

RPS11
SACM1L
SChLAP1
SEC61A1
SERPINB5.Maspin
SFRP4
SIM2.long
SIM2.short
SIRT1
SLC12A1
SLC43A1
SLC4A1.S
SMAP1.ex.7.8
SMIM1
SNCA
SNORA20
SPINK1
SPON2
SRSF3
SSPO
SSTR1
ST6GALNAC1
STEAP2
STEAP4
STOM
SULF2
SULT1A1
SYNM
TBP
TDRD
TERF2IP

RP11.97O12.7
RPL18A
RPL23AP53
RPLP2
RPS10
RPS11
SACM1L
SChLAP1
SEC61A1
SERPINB5.Maspin
SFRP4
SIM2.long
SIM2.short
SIRT1
SLC12A1
SLC43A1
SLC4A1.S
SMAP1.ex.7.8
SMIM1
SNCA
SNORA20
SPINK1
SPON2
SRSF3
SSPO
SSTR1
ST6GALNAC1
STEAP2
STEAP4
STOM
SULF2

TERT	SULT1A1
TFDP1	SYNM
Timp4	TBP
TMCC2	TDRD
TMEM45B	TERF2IP
TMEM47	TERT
TMEM86A	TFDP1
TMPRSS2.ERG.fusion	Timp4
TRPM4	TMCC2
TWIST1	TMEM45B
UPK2	TMEM47
VAX2	TMEM86A
VPS13A	TMPRSS2.ERG.fusion
ZNF577	TRPM4
	TWIST1
	UPK2
	VAX2
	VPS13A
	ZNF577

Table S2. AUC of all trained models, including a combination of EN2 and PSA, for detecting outcomes of an initial biopsy for varying clinically significant thresholds. Brackets detail 95% confidence intervals of the AUC, calculated from 1,000 stratified bootstrap resamples.

Initial biopsy outcome:	SoC	Engrailed	ExoRNA	ExoGrail	EN2_PSA
Gleason ≥4+3:	0.77 (0.69 - 0.84)	0.81 (0.73 - 0.88)	0.67 (0.60 - 0.75)	0.84 (0.78 - 0.89)	0.85 (0.78 - 0.91)
Gleason ≥3+4:	0.72 (0.65 - 0.79)	0.83 (0.77 - 0.88)	0.77 (0.70 - 0.83)	0.90 (0.85 - 0.94)	0.87 (0.82 - 0.91)
Any Cancer	0.75 (0.68 - 0.82)	0.81 (0.75 - 0.86)	0.81 (0.75 - 0.87)	0.89 (0.85 - 0.93)	0.84 (0.78 - 0.89)