Cancer Cell, Volume *13*

Supplemental Data

Article

The Role of *SPINK1*

in *ETS* Rearrangement-Negative Prostate Cancers

Scott A. Tomlins, Daniel R. Rhodes, Jianjun Yu, Sooryanarayana Varambally, Rohit Mehra, Sven Perner, Francesca Demichelis, Beth E. Helgeson, Bharathi Laxman, David S. Morris, Qi Cao, Xuhong Cao, Ove Andrén, Katja Fall, Laura Johnson, John T. Wei, Rajal B. Shah, Hikmat Al-Ahmadie, James A. Eastham, Scott E. Eggener, Samson W. Fine, Kristina Hotakainen, Ulf-Håkan Stenman, Alex Tsodikov, William L. Gerald, Hans Lilja, Victor E. Reuter, Phillip W. Kantoff, Peter T. Scardino, Mark A. Rubin, Anders S. Bjartell, and Arul M. Chinnaiyan

Table S1. Meta-COPA Analysis of Seven Prostate Cancer Gene Expression Profiling Data Sets in Oncomine. Genes were ranked by the number of studies in which they scored in the top 100 outliers (ranked by COPA) at any of the three predefined percentile cutoffs (75th, 90th, and 95th). Genes were further ranked by their average COPA rank (Avg. Rank) in studies in which they ranked in the top 100.

Meta COPA Rank	Gene	# of Studies	Avg. Rank
1	*ERG*	7	19.3
2	*SPINK1*	5	29.8
3	*GPR116*	5	46
4	*ORM1*	4	10
5	*ETV1*	4	23
6	*MYL2*	4	26.8
7	*NEB*	4	27
8	*TGM4*	4	30.8
9	*NELL2*	4	33.5
10	*KRT13*	4	49
11	*SLC26A4*	4	63.3
12	*MYL1*	3	8.7
13	*CXCL13*	3	11
14	*HCG3*	3	12
15	*HPGD*	3	20.7
16	*MIPEP*	3	27.3
17	*PLAT*	3	31.7
18	*CST1*	3	32.3
19	*COL2A1*	3	35.3
20	*CPS1*	3	36.3
21	*CRISP3*	3	36.7
22	*CTAG1A*	3	42.3
23	*FGB*	3	42.3
24	*PPFIA2*	3	42.7
25	*CDH2*	3	44
26	*PTPRM*	3	45.3
27	*CYP3A43*	3	51.7
28	*PROM1*	3	53
29	*HDAC9*	3	61.7

Table S2: Identification of *SPINK1* , *ERG* and *ETV1* outlier samples. Gene expression data for *SPINK1*, *ERG* and *ETV1* (reporter indicated) for each data set ("[First author name]_tissue") was downloaded from Oncomine. The total number of benign and cancerous samples are indicated. Expression values were median centered and outlier samples were indicated as described in the **Methods** and **Fig S8,** and the number of benign and cancer outlier samples for each gene is indicated. The number of cancer samples showing exclusive outlier-expression of *SPINK1*, *ERG* and *ETV1* is indicated.

LaTulippe_Prostate

	ERG (914_g_at)	ETV1 (37055_at)	SPINK1 (38582_at)	Exclusive
Benign (3)	0	0	0	NA
Cancer (23)	6	2	5	23

Vanaja_Prostate

	ERG (213541_s_at)	ETV1 (221911_at)	SPINK1 (206239_s_at)	Exclusive
Benign (8)	0	0	1	NA
Cancer (27)	14	1	4	27

Welsh_Prostate

	ERG (914_g_at)	ETV1 (37156_at)	SPINK1 (38582_at)	Exclusive
Benign (9)	0	1	0	NA
Cancer (25)	11	1	6	25

Dhanasekaran_Prostate

	ERG (IMAGE:123755)	ETV1 (no reporter)	SPINK1 (IMAGE:1412481)	Exclusive
Benign (22)	2	NA	0	NA
Cancer (56)	30	NA	4	56

Yu_Prostate

	ERG (914_g_at)	ETV1 (37055_at)	SPINK1 (38582_at)	Exclusive
Benign (23)	0	0	1	NA
Cancer (64)	24	4	3	64

Glinsky_Prostate

	ERG (213541_s_at)	ETV1 (221911_at)	SPINK1 (206239_s_at)	Exclusive
Benign (0)	0	0	0	NA
PCA (79)	37	8	10	78

Su_Multi-cancer

	ERG (914_g_at)	ETV1 (37156_at)	SPINK1 (38582_at)	Exclusive
Benign (0)	0	0	0	NA
Cancer (24)	11	1	5	24

Bittner_Multi-cancer (GSE2109)

	ERG (213541_s_at)	ETV1 (221911_at)	SPINK1 (206239_s_at)	Exclusive
Benign (0)	0	0	0	NA
Cancer (16)	5	0	3	16

Yang_Prostate*

	ERG (213541_s_at)	ETV1 (221911_at)	SPINK1 (206239_s_at)	Exclusive
Benign (71)	5	0	2	NA
Cancer (62)	27	3	16	59

*The Yang_Prostate data set was downloaded from GEO (GSE8218) and uploaded into a private version of Oncomine.

Table S3. Multivariate Cox Proportional-Hazards Regression Analysis for the Association of SPINK1 with Biochemical Recurrence.

Glinsky *et al.* microarray dataset	Recurrence Ratio	95% CI		P
SPINK1 (Positive vs. Negative)	2.522	1.067	5.960	**0.035**
Gleason (>=7 vs. <7)	2.768	0.806	9.500	0.106
Lymph node (Positive vs. Negative)	2.537	0.722	8.916	0.147
Surgical Margin (Positive vs. Negative)	2.284	0.975	5.353	0.057
Pre-operative PSA	1.040	1.001	1.081	**0.047**
Age	1.042	0.979	1.108	0.194

Note: Sample size is 79 with 37 recurrences. Tumor size was missing in the clinical table and thus not included in the model.

University of Michigan IHC cohort	Recurrence Ratio	95% CI		P
SPINK1 (Positive vs. Negative)	4.099	1.431	11.741	**0.009**
Gleason (>=7 vs. <7)	1.957	0.756	5.071	0.167
Tumor size (>=2 cm vs. <2 cm)	1.365	0.620	3.005	0.439
Surgical Margin (Positive vs. Negative)	1.342	0.566	3.180	0.504
Pre-operative PSA	1.050	0.986	1.118	0.127
Age	1.061	1.006	1.119	**0.028**

Note: Sample size is 75 with 28 recurrences.

MSKCC IHC cohort	Recurrence Ratio	95% CI		P
SPINK1 (Positive vs. Negative)	2.02	1.37	2.99	**0.0004**
Pre-operative PSA	1.02	1.01	1.03	**0.0002**
Seminal Vesical Invasion (Positive vs. Negative)	3.07	2.14	4.40	**1.1E-09**
Surgical Margin (Positive vs. Negative)	1.57	1.15	2.14	**0.004**
Lymph Node Involvement (Pos vs. Neg)	2.89	1.79	4.68	**1.5E-05**
Extracapsular Extension (Yes vs. No)	1.96	1.42	2.71	**4.6E-05**
Pathology Gleason (>=7 vs. <7)	1.68	1.18	2.40	**0.004**

Note: Sample size is 817 with 200 recurrences.

Table S4. *SPINK1* outlier status across cohorts. For the six cohorts evaluated in this study, the total number of prostate cancer samples analyzed, the number and percentage of *SPINK1* positive samples (as defined in the Methods for each assay), and the percentage of samples showing mutually exclusive over-expression of *SPINK1* and *ERG, ETV1* or *TMPRSS2-ERG* (if measured). The six cohorts are: the in silico microarray data, quantitative PCR (qPCR) on tissue samples, University of Michigan (UM) immunohistochemistry (IHC) /fluorescence in situ hybridization (FISH), Swedsih Watchful Waiting (SWW) IHC/FISH, Memorial Sloan Kettering Cancer Center (MSKCC) IHC and qPCR on urine samples.

Cohort	SPINK1 +	Total	% SPINK1 +	% Exclusive
Microarray	56	376	15%	98.9%
qPCR	4	61	7%	100%
UM IHC	10	75	13%	100%
SWW IHC	23	312	7%	100%
MSKCC IHC	75	823	9%	NA
Urine	11	148	7%	NA
Total:	**179**	**1795**	**10%**	**99.7% ***

* Individual percentages were averaged

Table S5. Differentially expressed genes upon siRNA knockdown of *SPINK1* in 22RV1 cells. 22RV1 cells were transfected with siRNA against *SPINK1* (siRNA *SPINK1*) or non-targeting control siRNA (NT siRNA). Total RNA was isolated and expression profiling was performed using the Agilent Whole Human Genome Oligo Microarray (GPL4133). Hybridizations (siRNA *SPINK1* /NT siRNA) were performed in duplicate with duplicate dye flips. Differentially expressed features (see Methods) are indicated, including the probe name, gene name, representative sequence, and averge fold change across the four hybridizations, corrected for the dye flip (siRNA *SPINK1* /NT siRNA). Genes validated by qPCR (see **Fig S4**) are indicated in blue

Feature ID	ProbeName	Gene	Sequence	siRNA *SPINK1*/ NT siRNA
36382	A_24_P119201	MBD2	NM_015832	-5.97
25530	A_23_P214079	SPINK1	NM_003122	-4.60
31329	A_23_P218058	KLRC4	NM_013431	-2.28
31695	A_23_P20068	GRM3	NM_000840	-2.17
9794	A_24_P56484	BRMS1L	NM_032352	-2.08
531	A_23_P77043	C14orf161	NM_024764	-2.07
41521	A_23_P306211	FAM84A	NM_145175	-1.95
17730	A_24_P937405	PRSS23	NM_007173	-1.90
44387	A_23_P150789	PRSS23	NM_007173	-1.85
13006	A_23_P251647	LOC317671	NM_173362	-1.78
20835	A_24_P413920	FAM84A	NM_145175	-1.67
4415	A_23_P145761	ARL4A	NM_005738	-1.62
17818	A_23_P250164	HGD	NM_000187	-1.62
12968	A_24_P220947	AKR1C1	NM_001353	-1.57
36945	A_32_P75581	BHLHB5	NM_152414	3.33
28438	A_24_P166663	CDK6	NM_001259	3.30
4139	A_32_P194563	THC2281660	THC2281660	3.21
5000	A_23_P201628	LAMC1	NM_002293	3.07
2464	A_23_P116414	HRASLS3	NM_007069	2.96
8126	A_23_P116235	MDK	NM_001012334	2.84
11956	A_23_P409386	SLC25A22	NM_024698	2.63
11319	A_32_P125338	FAM43B	NM_207334	2.60
27988	A_32_P218707	THC2314643	THC2314643	2.55
25309	A_23_P416608	LAMP2	NM_013995	2.53
44598	A_23_P141394	WIPI1	NM_017983	2.51
42489	A_24_P393811	TMCO1	NM_019026	2.44
6490	A_23_P209625	CYP1B1	NM_000104	2.42
37302	A_24_P276583	TMCO1	NM_019026	2.37
23878	A_23_P371266	DNM3	NM_015569	2.31
43554	A_23_P52793	PPP2R1B	NM_002716	2.31
18132	A_23_P130974	KIAA1683	NM_025249	2.30
6090	A_23_P355067	TMCO1	NM_019026	2.29
36903	A_24_P323084	C17orf55	NM_178519	2.27
10939	A_23_P150609	IGF2	NM_001007139	2.23
6181	A_23_P97990	HTRA1	NM_002775	2.22
11433	A_23_P138725	MARVELD1	NM_031484	2.16
17892	A_23_P380318	EGR4	NM_001965	2.14
26152	A_23_P253561	C20orf121	NM_024331	2.13
9644	A_24_P270033	ENST00000278949	ENST00000278949	2.12
28209	A_32_P191262	ACR	NM_001097	2.09

43210	A_24_P390583	THC2336533	THC2336533	2.08
42543	A_23_P164089	RFFL	NM_057178	2.07
37592	A_24_P80204	MALL	NM_005434	2.05
43445	A_23_P112452	BC045756	BC045756	2.01
14227	A_24_P521994	KLHL24	NM_017644	1.97
15160	A_23_P11915	GDAP2	NM_017686	1.96
9597	A_23_P353717	C16orf75	NM_152308	1.94
6602	A_32_P919718	TMEM105	NM_178520	1.94
38052	A_24_P211565	C1QTNF6	NM_031910	1.94
39302	A_23_P168651	CDK6	NM_001259	1.93
13321	A_23_P502797	WDFY1	NM_020830	1.92
33557	A_23_P147185	CANX	NM_001746	1.91
38826	A_23_P100711	PMP22	NM_000304	1.90
15973	A_23_P99540	ZFP36L1	NM_004926	1.90
44682	A_23_P6362	DERL3	NM_198440	1.87
11337	A_23_P107421	TK1	NM_003258	1.86
5137	A_23_P110837	IRX4	NM_016358	1.85
34291	A_23_P90273	CHST8	NM_022467	1.85
2773	A_23_P360804	CPNE5	NM_020939	1.82
34687	A_23_P206724	MT1E	NM_175617	1.82
20703	A_24_P199655	VANGL1	NM_138959	1.81
21796	A_24_P379512	PIGK	NM_005482	1.81
34561	A_23_P434900	C16orf34	NM_144570	1.81
43891	A_32_P216520	WIF1	NM_007191	1.81
6827	A_23_P207131	MAP3K3	NM_203351	1.80
23135	A_23_P111995	LOXL2	NM_002318	1.80
17812	A_23_P211631	FBLN1	NM_006486	1.79
18767	A_23_P52697	CD248	NM_020404	1.78
8319	A_24_P649624	ENST00000377093	ENST00000377093	1.77
21189	A_23_P2554	KIAA0152	NM_014730	1.76
19586	A_24_P592544	BC107568	BC107568	1.76
20512	A_23_P14734	RPS27L	NM_015920	1.76
26298	A_23_P8640	GPR30	NM_001505	1.75
708	A_23_P336513	GEMIN5	NM_015465	1.74
560	A_24_P921260	AK022030	AK022030	1.74
9436	A_23_P10385	DTL	NM_016448	1.70
24715	A_32_P206541	AK128714	AK128714	1.70
7603	A_23_P215431	POM121	AB014518	1.69
26526	A_23_P38894	FLJ11286	NM_018381	1.69
37428	A_32_P201521	TMEM97	NM_014573	1.69
9607	A_23_P162846	LAMP1	NM_005561	1.69
20121	A_23_P10385	DTL	NM_016448	1.69
10990	A_23_P10385	DTL	NM_016448	1.67
13501	A_23_P10385	DTL	NM_016448	1.67
25496	A_32_P104334	AW972815	AW972815	1.65
9342	A_24_P56270	CR612226	CR612226	1.64
20136	A_23_P398637	RG9MTD2	NM_152292	1.64
34560	A_24_P389612	ENST00000377275	ENST00000377275	1.62
32009	A_23_P109122	RP5-860F19.3	BC054347	1.61
23338	A_23_P23924	CAPN2	NM_001748	1.60

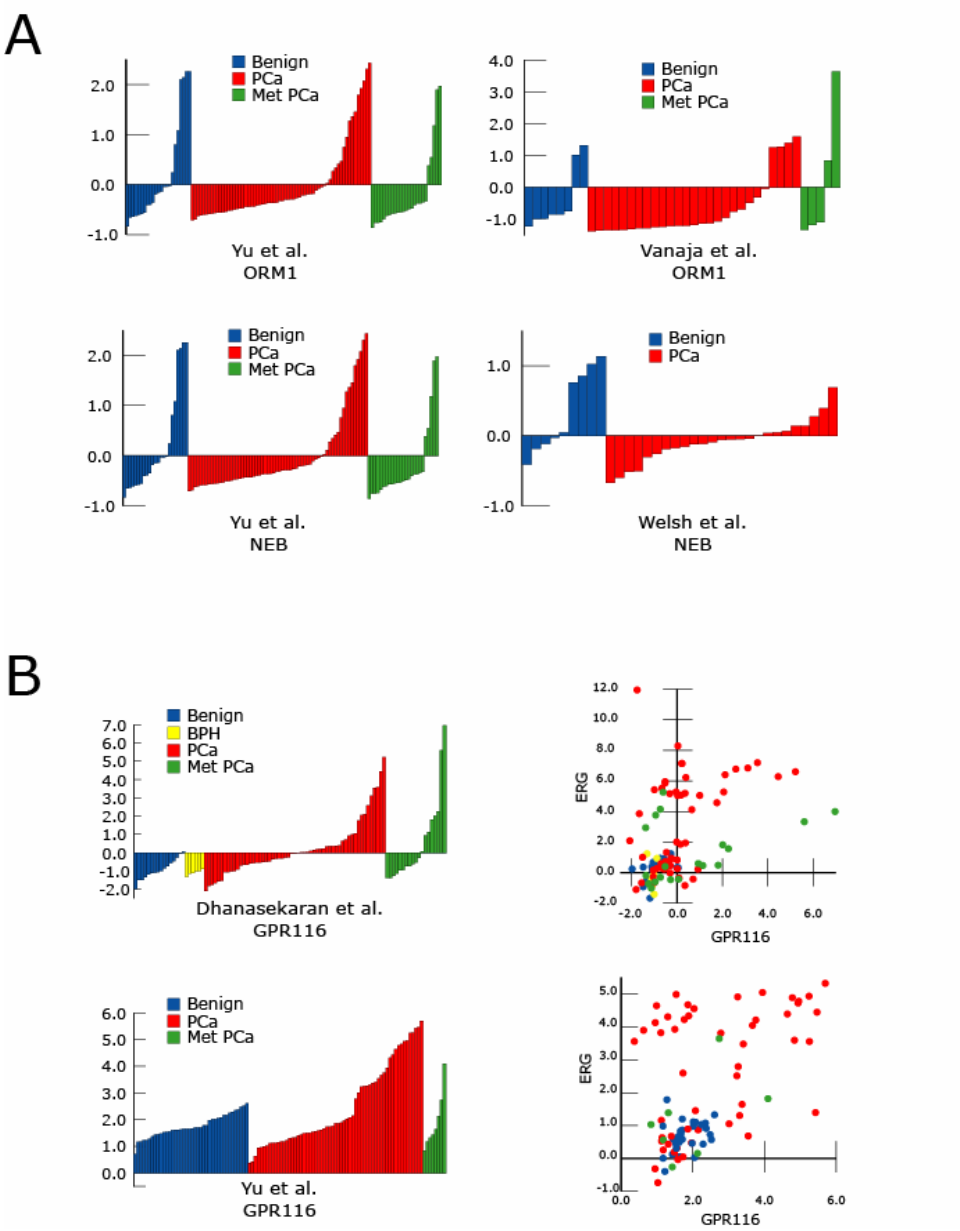


Figure S1. Meta-Outlier Genes Showing Overexpression in Benign Prostate Tissue and *ETS*-Positive Prostate Cancers

Meta-outliers, as indicated in Table 1, were analyzed for exclusive overexpression in prostate cancer compared to benign prostate tissue and mutually exclusive overexpression with *ERG* and *ETV1*.

(A) The expression of meta-outlier genes ORM (ranked 4th) and *NEB* (ranked 7th) in normalized expression units are shown from the indicated studies, according to the sample classes described in Figure 1, revealing outlier-expression in multiple benign samples.

(B) The expression of the 3rd ranked meta-outlier gene *GPR116* (left panels) and scatter plots of *GPR116* vs. *ERG* (right panels) for all profiled samples in two studies shows co-outlier expression of *GPR116* and *ERG* in multiple samples.

Figure S2

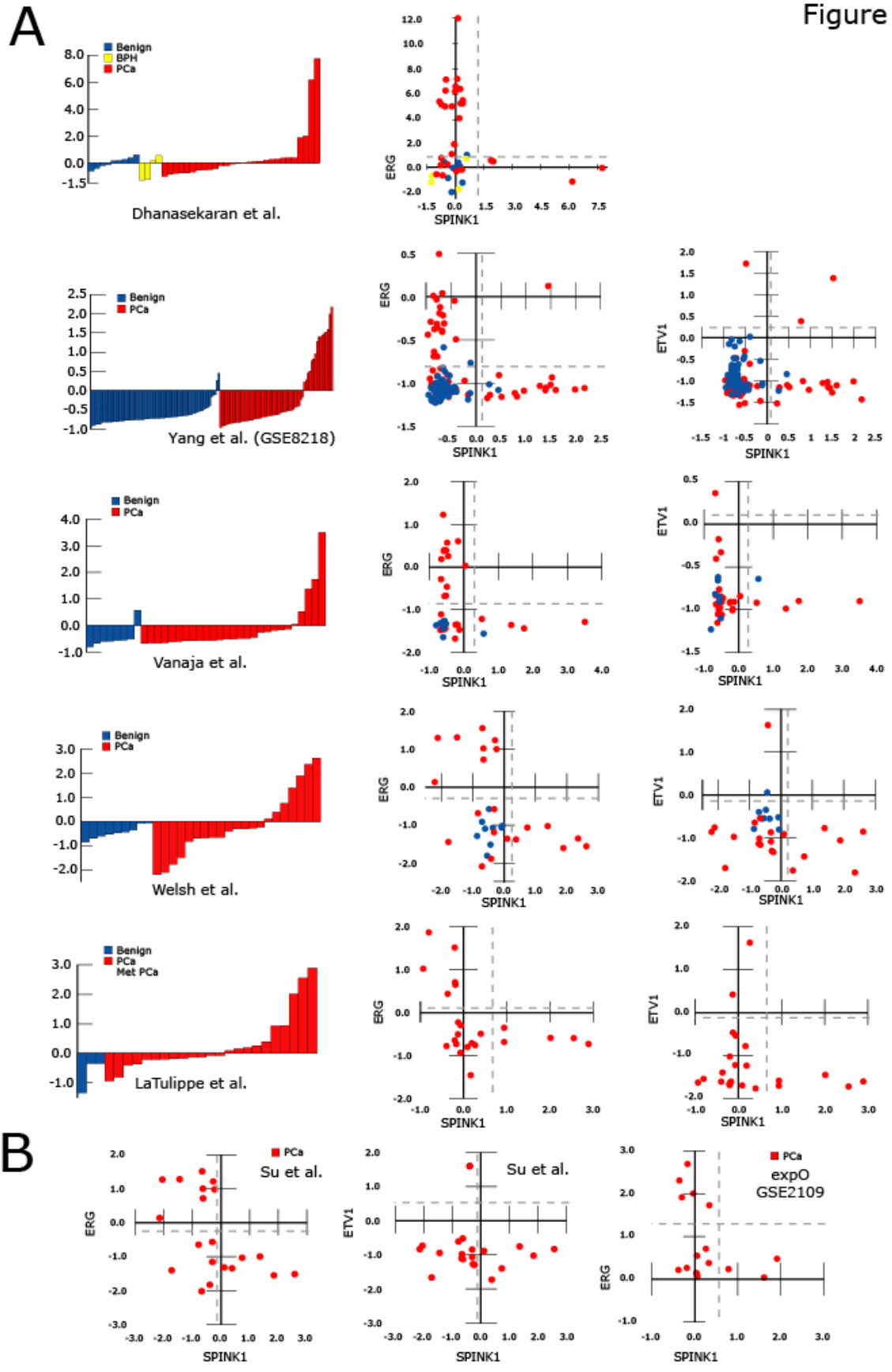

A
Benign
BPH
PCa
Met PCa
Dhanasekaran et al.
Yang et al. (GSE8218)
Vanaja et al.
Welsh et al.
LaTulippe et al.
ERG
ETV1
SPINK1
B
Su et al.
expO
GSE2109

Figure S2. Overexpression of *SPINK1* in Prostate Cancer Compared to Benign Prostate Tissue and Mutually Exclusive Overexpression with *ERG* and *ETV1* in DNA Microarray Studies

The expression of *SPINK1* and scatter plots of *SPINK1* vs. *ERG* and *SPINK1* vs. *ETV1* (if measured) for five studies profiling distinct classes of prostate tissue (A) and two studies profiling prostate cancers as part of multi-cancer studies (B) are shown as in Figure 1. Outlier-expression is delineated by the dashed gray lines (See Experimental Procedures). *SPINK1* vs. *ETV1* is not shown for the expO study, as no samples showed *ETV1* overexpression.

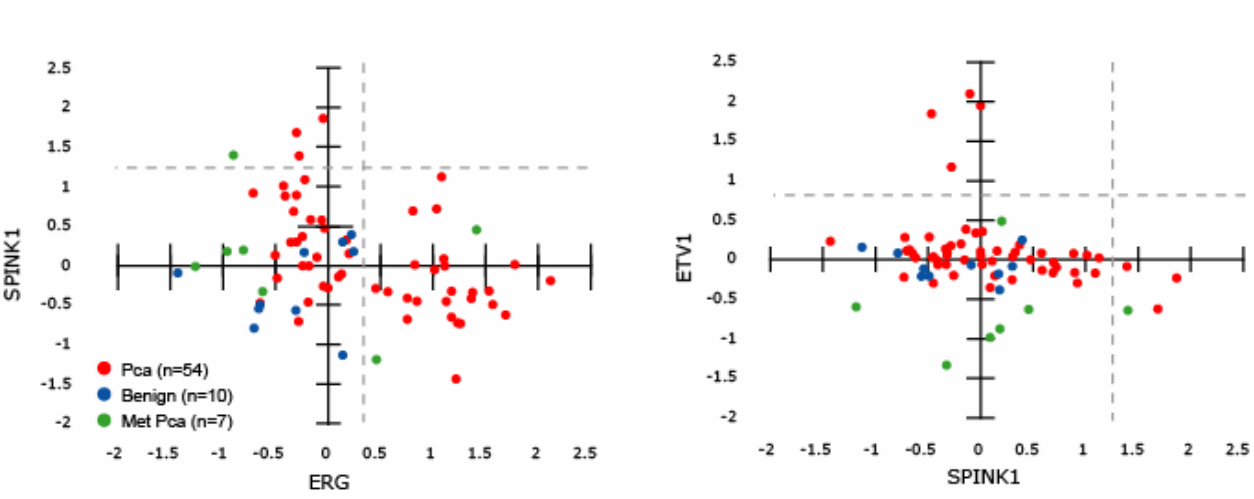


Figure S3. Overexpression of *SPINK1* in Prostate Cancer Compared to Benign Prostate Tissue and Mutually Exclusive Overexpression with *ERG* and *ETV1*

Scatter plots of *ERG* vs. *SPINK1* (left panel) and *ETV1* vs. *SPINK1* (right panel) by qPCR in 10 benign prostate samples (blue), 54 localized prostate cancers (PCa, red) and 7 metastatic (Met) PCa samples (green). Log of Target gene normalized to the average of HMBS+GAPDH are plotted. Outlier-expression is delineated by the dashed gray lines (See Experimental Procedures).

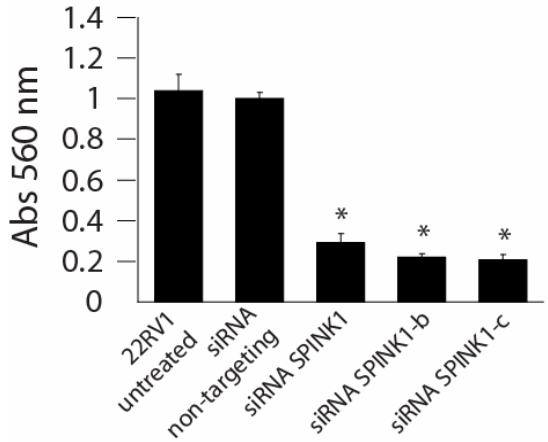


Figure S4. Knockdown of *SPINK1* in 22RV1 Prostate Cancer Cells Attenuates Invasiveness

SPINK1 mediates invasiveness in 22RV1 cells. 22RV1 cells were treated with transfection reagent alone (untreated), or transfected with non-targeting or siRNA against *SPINK1* as in Figure 5, or two additional siRNAs directed against *SPINK1* (SPINK1-b, -c), and cells were assayed for invasion. Mean (n = 3) + SEM are shown, and P values < 0.05 are indicated by asterisks.

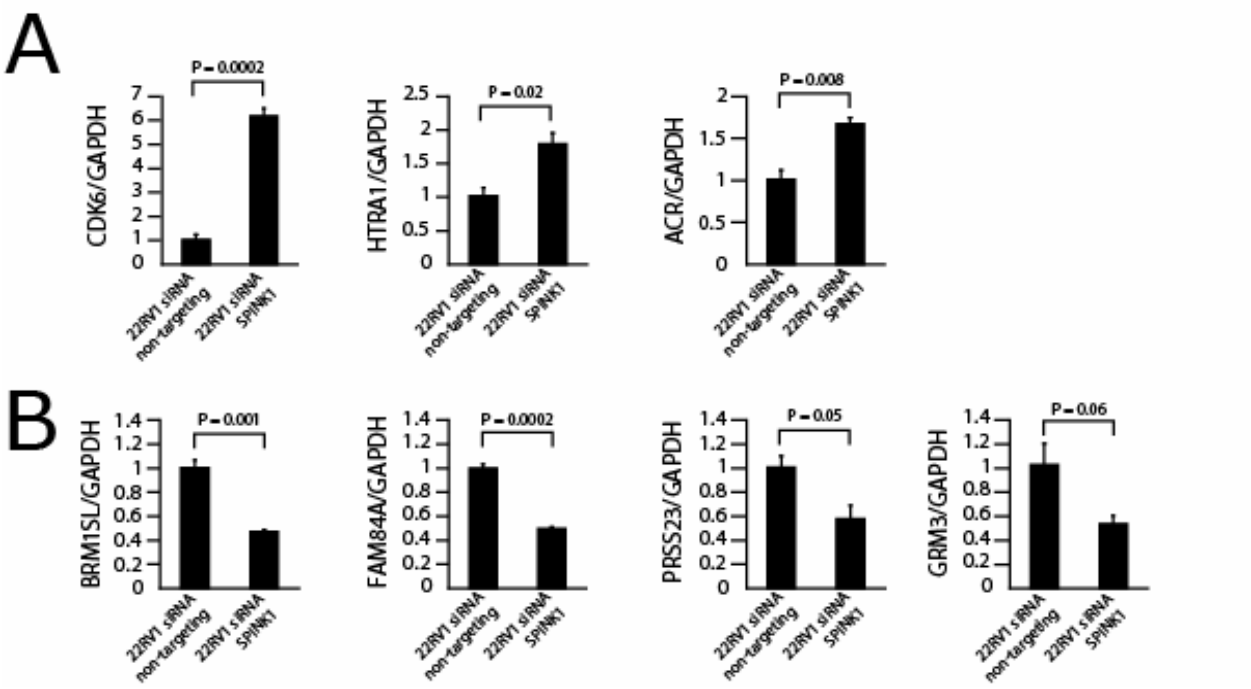
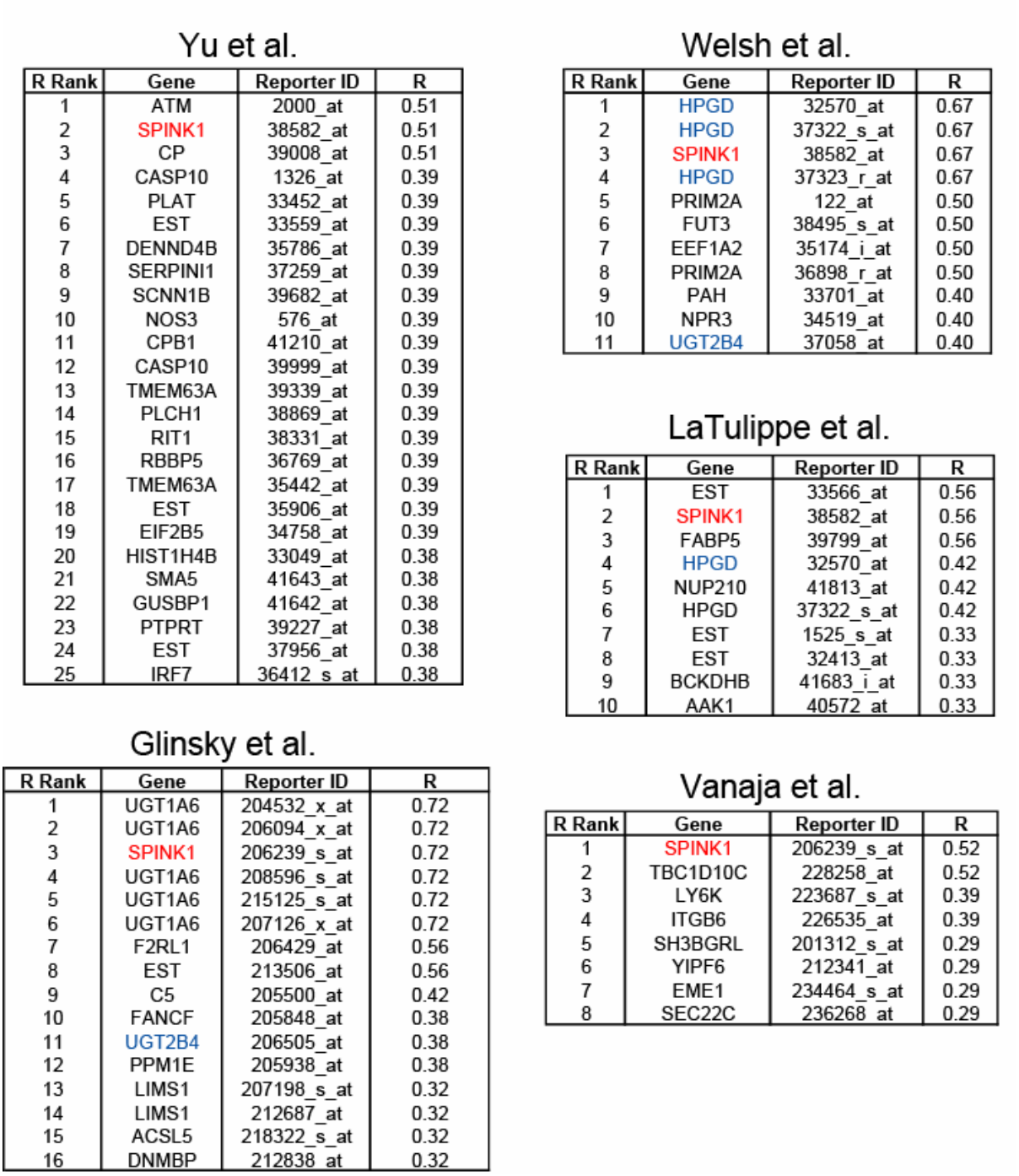


Figure S5. qPCR Confirmation of Genes Differentially Expressed upon *SPINK1* Knockdown in 22RV1 Cells

22RV1 cells were transfected with siRNA against SPINK1 (siRNA siSPINK1) or non-targeting control siRNA (NT siRNA). Total RNA was isolated and expression profiling was performed using the Agilent Whole Human Genome Oligo Microarray (GPL4133). Differentially expressed features are indicated in Table S4. Selected overexpressed (A) and underexpressed (B) genes in 22RV1 siSPINK1 cells were assessed by quantitative PCR as shown. The amount of target gene in each sample was normalized to the amount of *GAPDH*. Mean (n = 3) + SEM are shown, and p values are indicated.



Yu et al.

R Rank	Gene	Reporter ID	R
1	ATM	2000_at	0.51
2	SPINK1	38582_at	0.51
3	CP	39008_at	0.51
4	CASP10	1326_at	0.39
5	PLAT	33452_at	0.39
6	EST	33559_at	0.39
7	DENND4B	35786_at	0.39
8	SERPINI1	37259_at	0.39
9	SCNN1B	39682_at	0.39
10	NOS3	576_at	0.39
11	CPB1	41210_at	0.39
12	CASP10	39999_at	0.39
13	TMEM63A	39339_at	0.39
14	PLCH1	38869_at	0.39
15	RIT1	38331_at	0.39
16	RBBP5	36769_at	0.39
17	TMEM63A	35442_at	0.39
18	EST	35906_at	0.39
19	EIF2B5	34758_at	0.39
20	HIST1H4B	33049_at	0.38
21	SMA5	41643_at	0.38
22	GUSBP1	41642_at	0.38
23	PTPRT	39227_at	0.38
24	EST	37956_at	0.38
25	IRF7	36412_s_at	0.38

Welsh et al.

R Rank	Gene	Reporter ID	R
1	HPGD	32570_at	0.67
2	HPGD	37322_s_at	0.67
3	SPINK1	38582_at	0.67
4	HPGD	37323_r_at	0.67
5	PRIM2A	122_at	0.50
6	FUT3	38495_s_at	0.50
7	EEF1A2	35174_i_at	0.50
8	PRIM2A	36898_r_at	0.50
9	PAH	33701_at	0.40
10	NPR3	34519_at	0.40
11	UGT2B4	37058_at	0.40

LaTulippe et al.

R Rank	Gene	Reporter ID	R
1	EST	33566_at	0.56
2	SPINK1	38582_at	0.56
3	FABP5	39799_at	0.56
4	HPGD	32570_at	0.42
5	NUP210	41813_at	0.42
6	HPGD	37322_s_at	0.42
7	EST	1525_s_at	0.33
8	EST	32413_at	0.33
9	BCKDHB	41683_i_at	0.33
10	AAK1	40572_at	0.33

Glinsky et al.

R Rank	Gene	Reporter ID	R
1	UGT1A6	204532_x_at	0.72
2	UGT1A6	206094_x_at	0.72
3	SPINK1	206239_s_at	0.72
4	UGT1A6	208596_s_at	0.72
5	UGT1A6	215125_s_at	0.72
6	UGT1A6	207126_x_at	0.72
7	F2RL1	206429_at	0.56
8	EST	213506_at	0.56
9	C5	205500_at	0.42
10	FANCF	205848_at	0.38
11	UGT2B4	206505_at	0.38
12	PPM1E	205938_at	0.38
13	LIMS1	207198_s_at	0.32
14	LIMS1	212687_at	0.32
15	ACSL5	218322_s_at	0.32
16	DNMBP	212838_at	0.32

Vanaja et al.

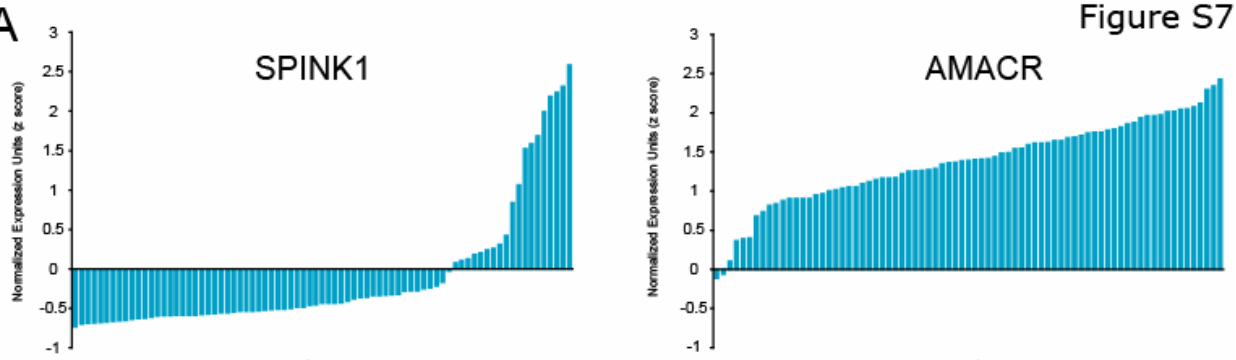
R Rank	Gene	Reporter ID	R
1	SPINK1	206239_s_at	0.52
2	TBC1D10C	228258_at	0.52
3	LY6K	223687_s_at	0.39
4	ITGB6	226535_at	0.39
5	SH3BGRL	201312_s_at	0.29
6	YIPF6	212341_at	0.29
7	EME1	234464_s_at	0.29
8	SEC22C	236268_at	0.29

Figure S6. Identification of Genes Showing Coexpression with *SPINK1* across Multiple Prostate Cancer Profiling Studies

Genes showing coexpression with *SPINK1* (R > 0.5) from the prostate cancer profiling studies included in the Meta-COPA analysis. *SPINK1* was queried in the Oncomine database using the coexpression module. For each study, all genes showing R > 0.3 are listed, along with the corresponding feature identification. *SPINK1* is indicated in red. Genes showing R > 0.3 in multiple studies are indicated in blue.

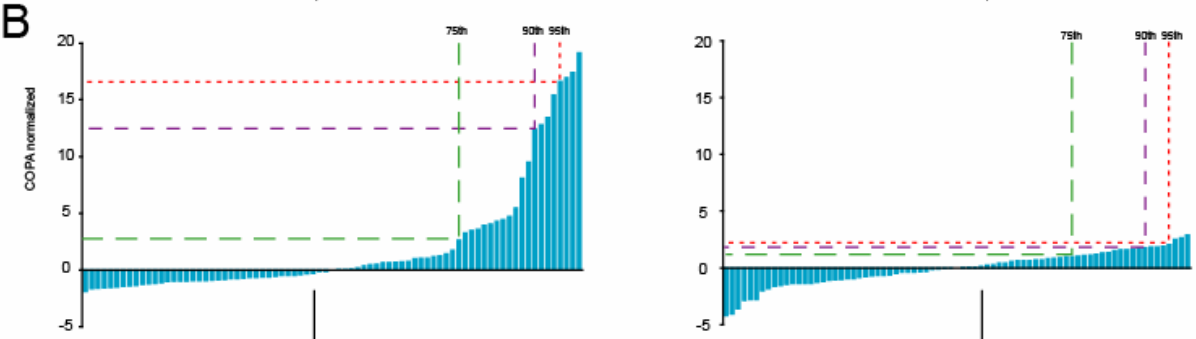
Figure S7

A



COPA Transformation
1. center median to zero
2. scale median absolute deviation to one
3. calculate COPA score at 75th, 90th and 95th percentile

B



4. Rank all genes in data set by COPA score at 75th, 90th and 95th percentile

C

75th

Rank	Gene	Reporter	COPA score
1	HLA-DQA1	203290_at	4.693
2	ORM1	205041_s_at	3.986
3	GP2	214324_at	3.437
4	TGM4	217566_s_at	3.296
5	FAM118A	219629_at	3.176
6	ERG	222079_at	3.131
7	TGM4	206260_at	3.046
8	EST	205048_s_at	2.97
9	CST4	206224_at	2.775
10	COL2A1	217404_s_at	2.77

90th

Rank	Gene	Reporter	COPA score
1	TDO2	205943_at	16.028
2	TGM4	217566_s_at	15.06
3	TTN	208195_at	14.016
4	SLC3A1	205799_s_at	13.113
5	SPINK1	206239_s_at	12.4
6	ORM1	205040_at	11.115
7	CST4	206224_at	10.352
8	SI	206664_at	9.612
9	COL2A1	217404_s_at	9.356
10	GP2	214324_at	9.147

95th

Rank	Gene	Reporter	COPA score
1	TTN	208195_at	27.898
2	SLC3A1	205799_s_at	23.472
3	TNNC1	209904_at	21.907
4	MYL2	209742_s_at	18.681
5	EST	205430_at	18.298
6	ORM1	205040_at	18.078
7	TDO2	205943_at	17.897
8	TGM4	217566_s_at	17.387
9	ETV1	217053_x_at	17.17
10	ETV1	221911_at	16.908

5. Define outlier genes as those that rank in the top 100 COPA scores at the 75th, 90th, or 95th percentile
6. Rank genes by the number of data sets in which they rank in the top 100 COPA scores at the 75th, 90th or 95th percentile

D

Meta COPA Rank	Gene	# of Studies	Avg. Rank
1	ERG	7	19.3
2	SPINK1	5	29.8
3	GPR116	5	46
4	ORM1	4	10
5	ETV1	4	23
6	MYL2	4	26.8
7	NEB	4	27
8	TGM4	4	30.8
9	NELL2	4	33.5
10	KRT13	4	49
11	SLC26A4	4	63.3

Figure S7. Meta-Outlier Analysis Summary

COPA analysis was performed on 7 prostate cancer gene expression data sets in Oncomine 3.0 (www.oncomine.org).

(A) First, for each data set considering all samples, gene expression values (in Oncomine normalized expression units) are median-centered per gene, setting each gene's median expression value to zero. Each bar in the figure represents an individual sample. Second, the median absolute deviation (MAD) is calculated per gene and scaled to 1 by dividing each gene expression value by its MAD. Of note, median and MAD are used for transformation as opposed to mean and standard deviation so that outlier expression values do not unduly influence the distribution estimates, and are thus preserved post-normalization.

(B) Third, for each gene in each data set, COPA scores are computed as the 75^{th}, 90^{th} and 95^{th} percentile of the transformed gene expression values. Thus, each gene in each data set has 3 COPA scores, one at each percentile value, representing the degree of overexpression in decreasing subsets of cases. The expression of *SPINK1* (outlier profile) and *AMACR* (typical biomarker profile) from the Glinsky *et al*. study are shown in the left and right panels, respectively.

(C) Fourth, in each data set, all genes are rank-ordered by the 3 COPA scores, generating 3 rank-ordered lists of genes per data set.

(D) Fifth, for each data set, we defined outlier genes as those that ranked in the top 100 COPA scores in any one of the 3 rank-ordered lists. Sixth, to identify "meta-outlier" genes, we ranked genes by the number of data sets where the gene was identified as an outlier gene. Genes identified as outliers in the same number of studies were further ranked by their average outlier rank across those studies.

Figure S8

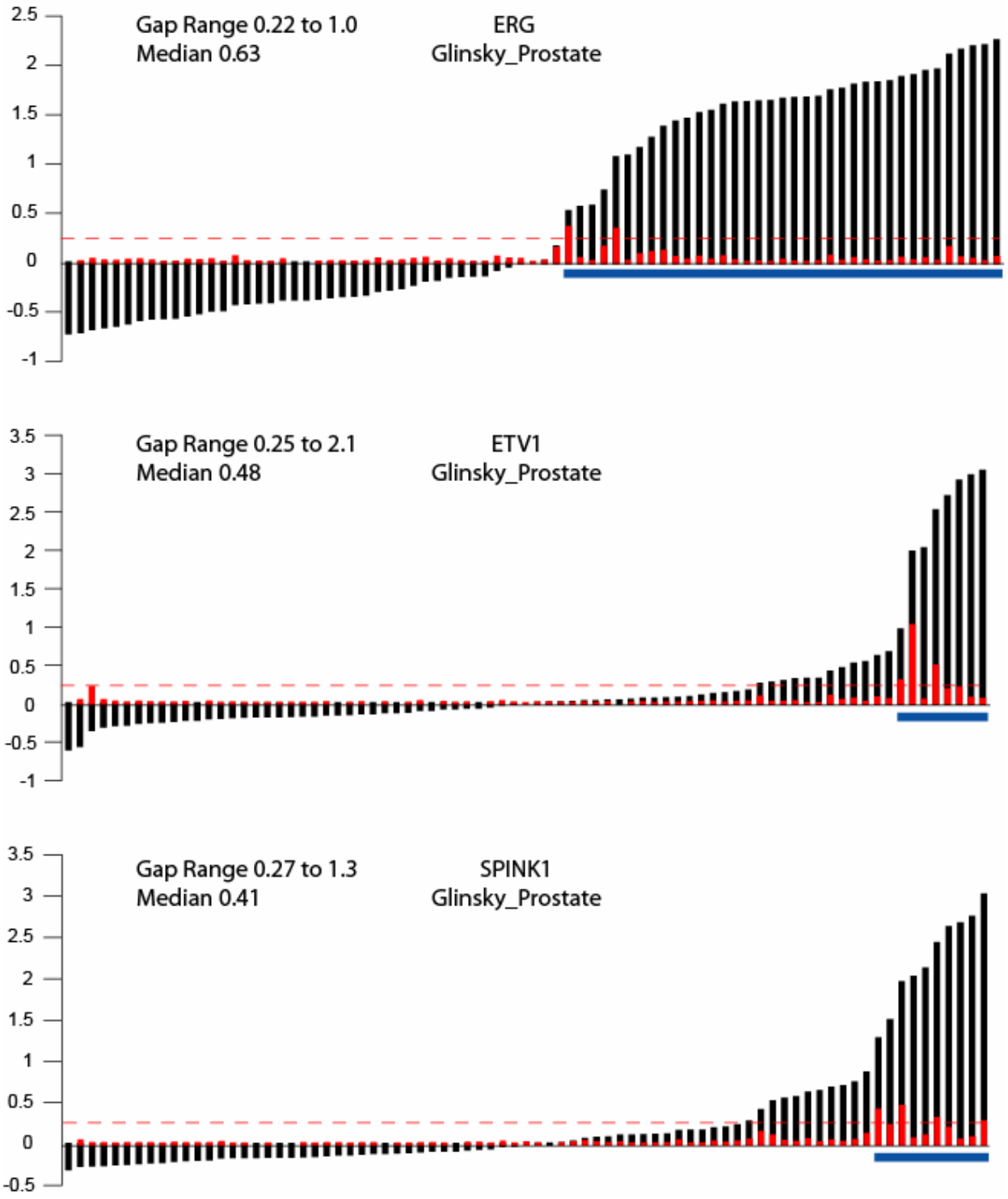

Gap Range 0.22 to 1.0
Median 0.63
ERG
Glinsky_Prostate
2.5
2
1.5
1
0.5
0
-0.5
-1
Gap Range 0.25 to 2.1
Median 0.48
ETV1
Glinsky_Prostate
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
Gap Range 0.27 to 1.3
Median 0.41
SPINK1
Glinsky_Prostate
3.5
3
2.5
2
1.5
1
0.5
0
-0.5

Figure S8. Identification of Samples Showing Outlier Expression

To identify individual samples showing outlier-expression in each microarray data set, we used a two step process to recreate the visual identification of the natural “gap” between non-outlier and outlier samples. First, gene expression values for all prostate samples in the data set (excluding metastatic prostate cancer) were median centered. Next, all samples were rank ordered in ascending order and the difference between each rank ordered sample and the preceding sample was calculated. The difference, or gap, between visually identified non-outlier and outlier populations in all studies for *ERG* ranged from 0.22 to 1.0 (median 0.63). Hence, we defined the first sample with a positive median centered value and greater than 0.22 normalized expression unit gap compared to the preceding sample as the transition to samples with outlier expression. Similar gaps were observed for *ETV1* and *SPINK1* gene expression values and the same criteria (positive median centered value and greater than 0.22 unit gap) was used to define the outlier population. Median centered expression values (in normalized expression units, z-scores) for *SPINK1*, *ETV1* and *ERG* from the “Glinsky_prostate” data set are plotted in black. The difference between each sample and the preceding sample are plotted in red. The 0.22 gap threshold is shown in the dashed red line. Outlier populations defined by this method are indicated in blue. Range and Median gap values for each gene in all data sets is given. No gap > 0.22 units was identified for *ETV1* in the GSE2109 data set.