Supplementary information

Common germline risk variants impact somatic alterations and clinical features across cancers

S Namba and Y Saito et al.

Corresponding to Keisuke Kataoka (kkataoka-tky@umin.ac.jp) and Yukinori Okada (yokada@sg.med.osaka-u.ac.jp)

Table of contents

■ **Supplementary Figures 4**

Supplementary Figure 1: The comparison of the best PRS and C+T PRS with genome-wide significant variants in UKB regarding metrics other than R^2. 4

Supplementary Figure 2: Receiver operating characteristic curves of the PRSs for individual cancer types in UKB. 5

Supplementary Figure 3: Odds ratios stratified by PRS percentiles for individual cancer types. 7

Supplementary Figure 4: Genetically inferred population and sex of TCGA samples. 9

Supplementary Figure 5: Evaluation of PRSs using TCGA resources. 11

Supplementary Figure 6: Meta-analyses of germline–somatic associations across cancer types in TCGA. 13

Supplementary Figure 7: Associations between PRS values and SCNAs in TCGA. 15

Supplementary Figure 8: Associations between PRS values and driver mutations in TCGA. 16

Supplementary Figure 9: Meta-analyses of associations between PRS values and hallmark signatures in TCGA. 18

Supplementary Figure 10: Associations between PRS values and clinical and other features in TCGA. 19
Supplementary Figure 11: Meta-analyses of germline–somatic associations across four PCAWG PRAD cohorts. 21

■ **Supplementary Tables 22**

Supplementary Table 1: GWAS summary statistics used for PRS. 22
Supplementary Table 2: Definition of cases and controls in UKB. 24
Supplementary Table 3: Parameters and numbers of variants used for PRS. 26
Supplementary Table 4: Evaluation of PRSs using UKB resources. 27
Supplementary Table 5: Number of patients without sex aneuploidy in TCGA. 28
Supplementary Table 6: The gene-level associations for UCEC and BRCA in the cell cycle pathway. 29

■ **Supplementary Materials and Methods 30**

UK Biobank (UKB) 30
Heritability estimates 30
Pathway enrichment analysis 30
The Cancer Genome Atlas (TCGA) 31
Somatic mutation analysis 34
Somatic copy number analysis 34
Immune status analysis 35
Single-sample Gene Set Enrichment analysis 35
Clinical data analysis 35
Validation using the Pan-Cancer Analysis of Whole Genomes (PCAWG) PRAD cohorts 36

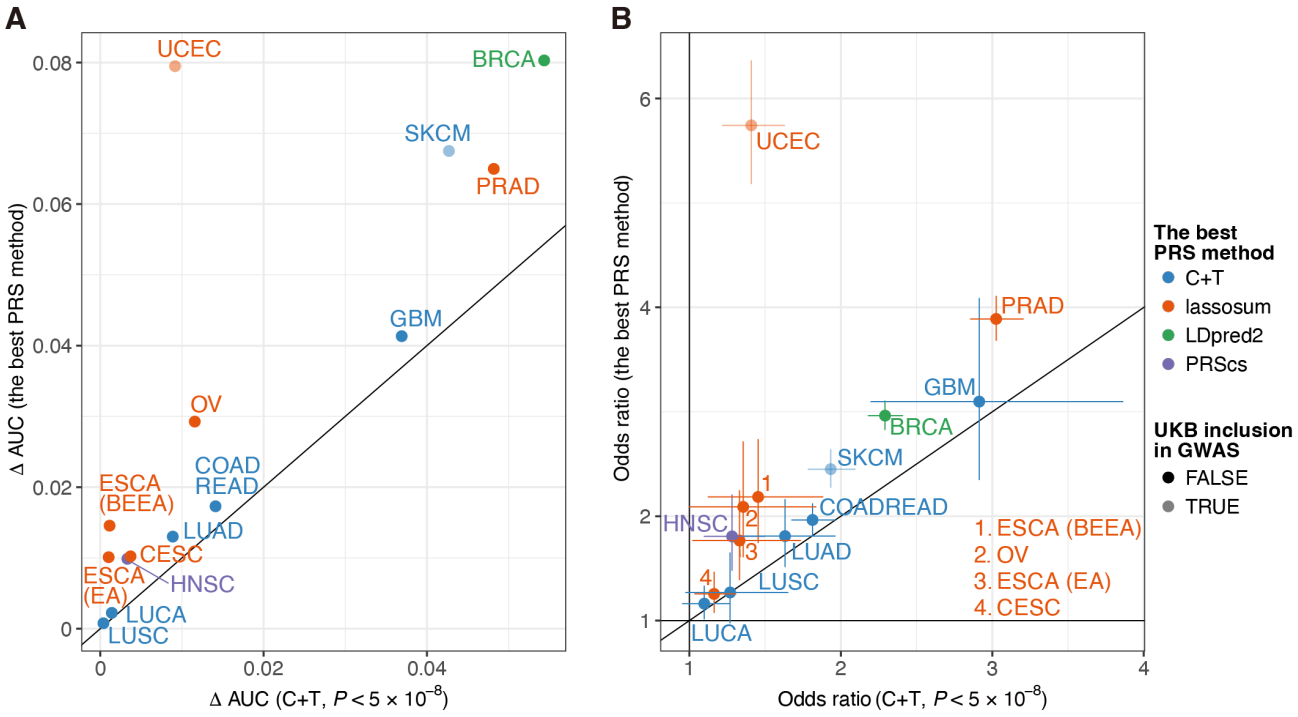
■ **Supplementary Notes 37**

Supplementary Note 1: Acknowledgments. 37

Supplementary Note 2: Contributing authors participating in the Glioma International Case Control Study, the UCSF Adult Glioma Study, and the GliomaScan consortia..... 45

■ **Supplementary References 49**

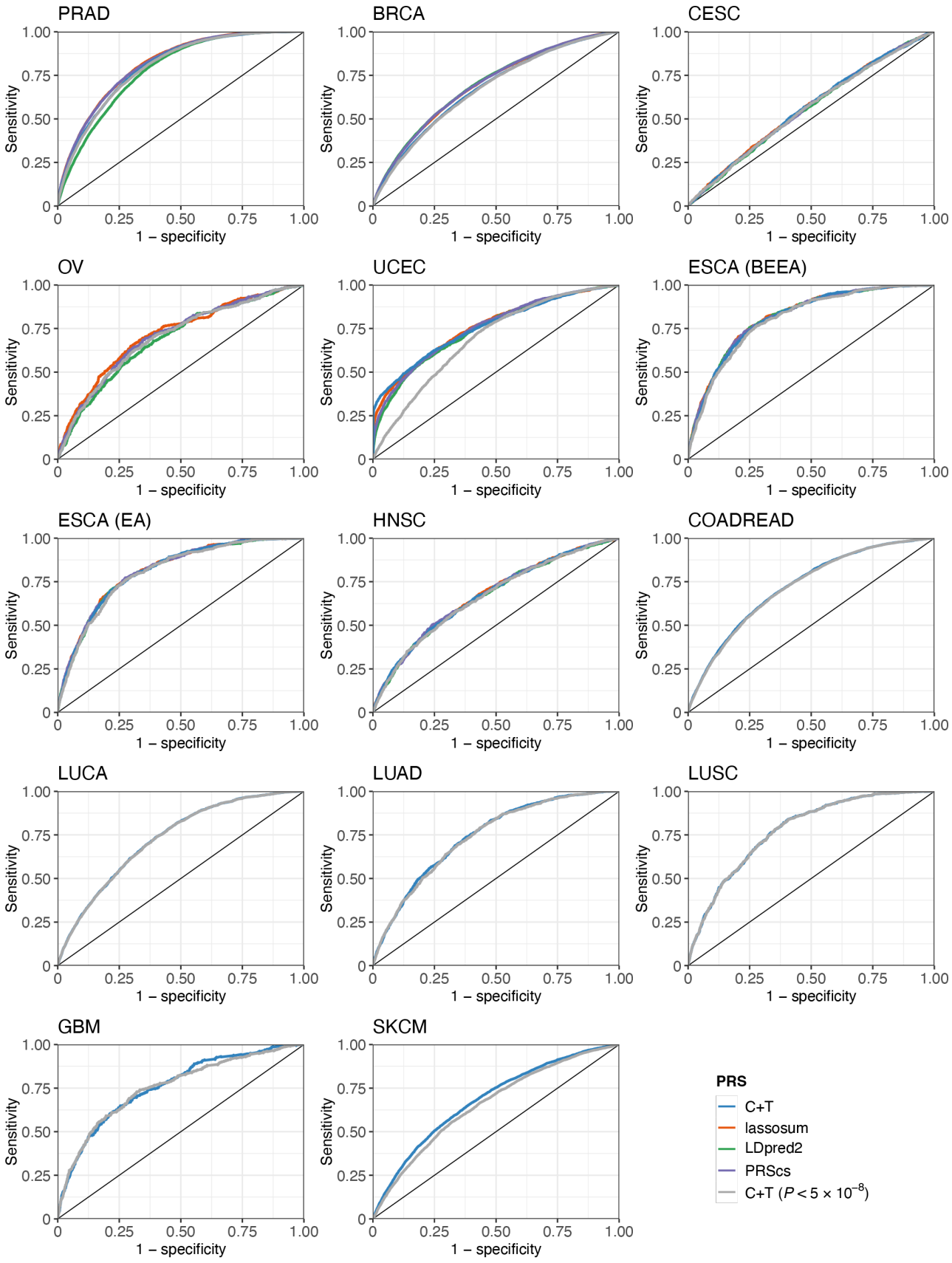
Supplementary Figures



Supplementary Figure 1: The comparison of the best PRS and C+T PRS with genome-wide significant variants in UKB regarding metrics other than R^2.

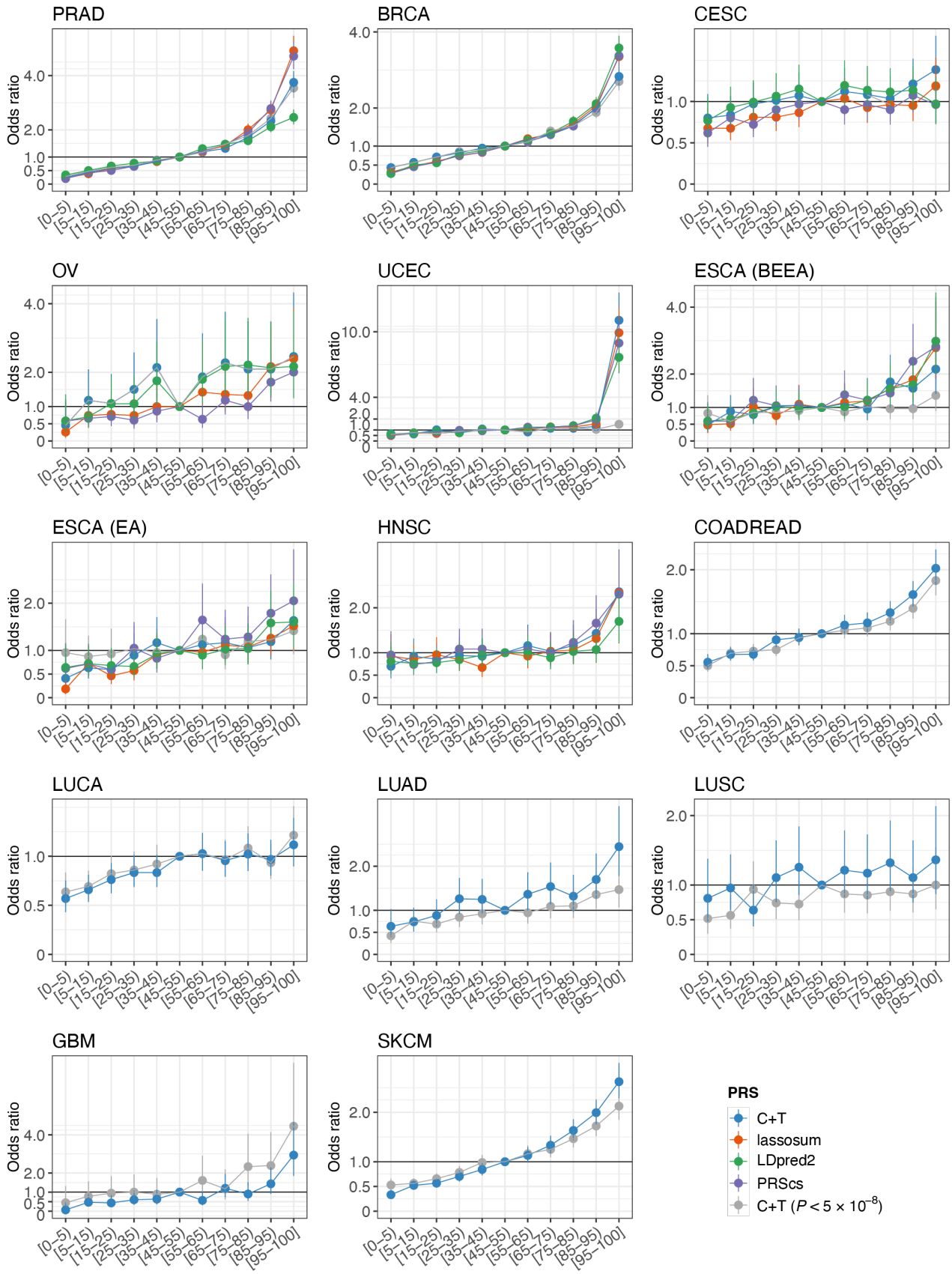
(A) Scatter plots showing the difference of area under the receiver operating characteristic curve (ΔAUC). For each of the best and C+T PRS, we calculated the difference between the full model and the reduced model including all covariates but PRS (i.e., age, sex, and top 20 genetic principal components [PCs]).

(B) Scatter plots showing the odds ratio of the largest PRS decile against the rest of the samples. Error bars represent standard error.


PRAD
BRCA
CESC
OV
UCEC
ESCA (BEEA)
ESCA (EA)
HNSC
COADREAD
LUCA
LUAD
LUSC
GBM
SKCM
Sensitivity
1 – specificity
0
0.25
0.50
0.75
1.00
PRS
C+T
lassosum
LDpred2
PRScs
C+T ($P < 5 \times 10^{-8}$)

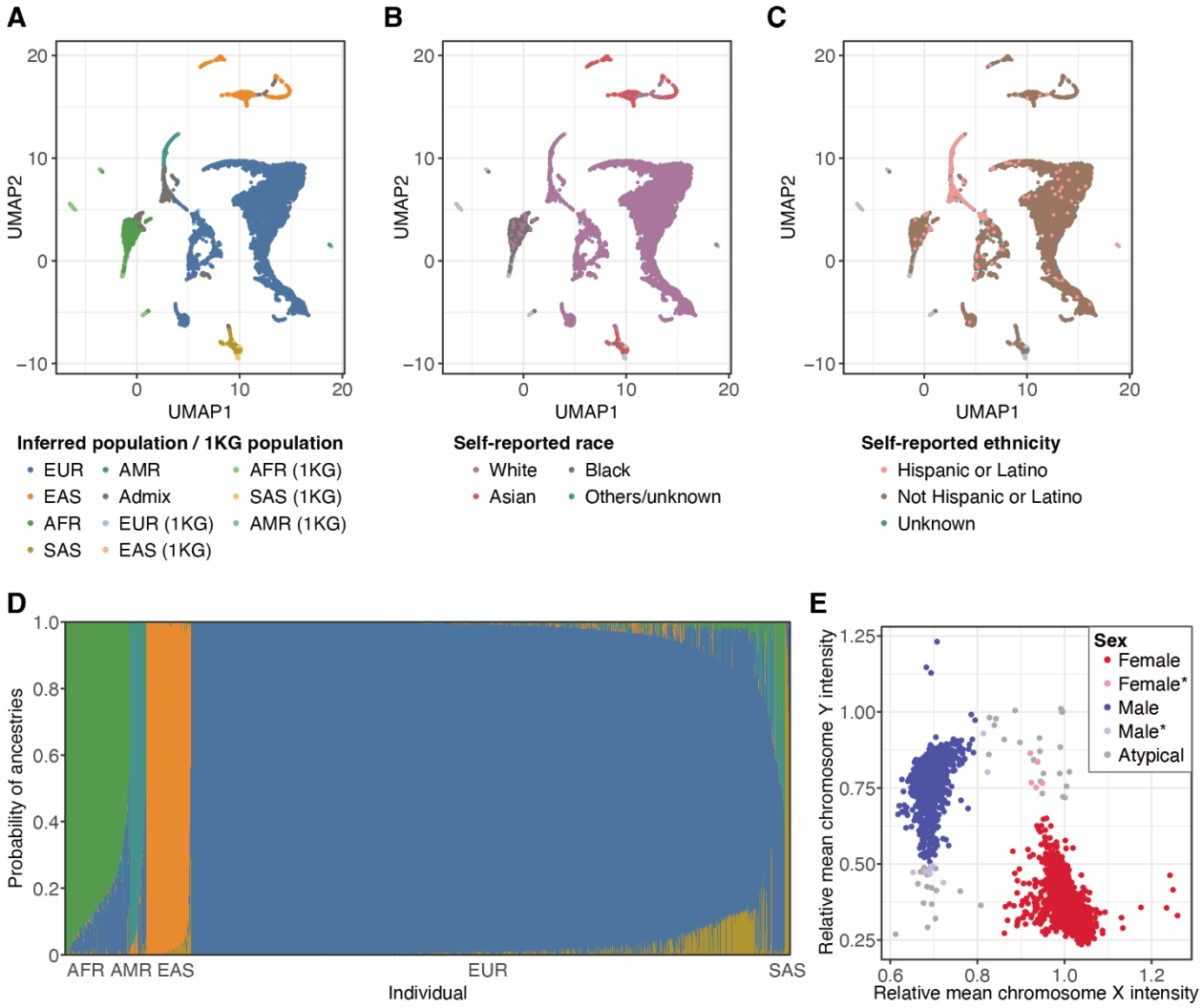
Supplementary Figure 2: Receiver operating characteristic curves of the PRSs for individual cancer types in UKB.

The PRSs were evaluated in UKB data with age, sex, and top 20 genetic PCs as covariates.


PRAD
BRCA
CESC
OV
UCEC
ESCA (BEEA)
ESCA (EA)
HNSC
COADREAD
LUCA
LUAD
LUSC
GBM
SKCM
Odds ratio
[0–5)
[5–15)
[15–25)
[25–35)
[35–45)
[45–55)
[55–65)
[65–75)
[75–85)
[85–95)
[95–100]
PRS
C+T
lassosum
LDpred2
PRScs
C+T ($P < 5 \times 10^{-8}$)

Supplementary Figure 3: Odds ratios stratified by PRS percentiles for individual cancer types.

Odds ratios were calculated against the 45–55 percentile. The odds ratios of C+T with the threshold of 5×10^{-8} were not shown for HNSC because there was no individual in the 45–55 percentile due to the sparse distribution of the PRS values. Error bars represent standard error.



Supplementary Figure 4: Genetically inferred population and sex of TCGA samples.

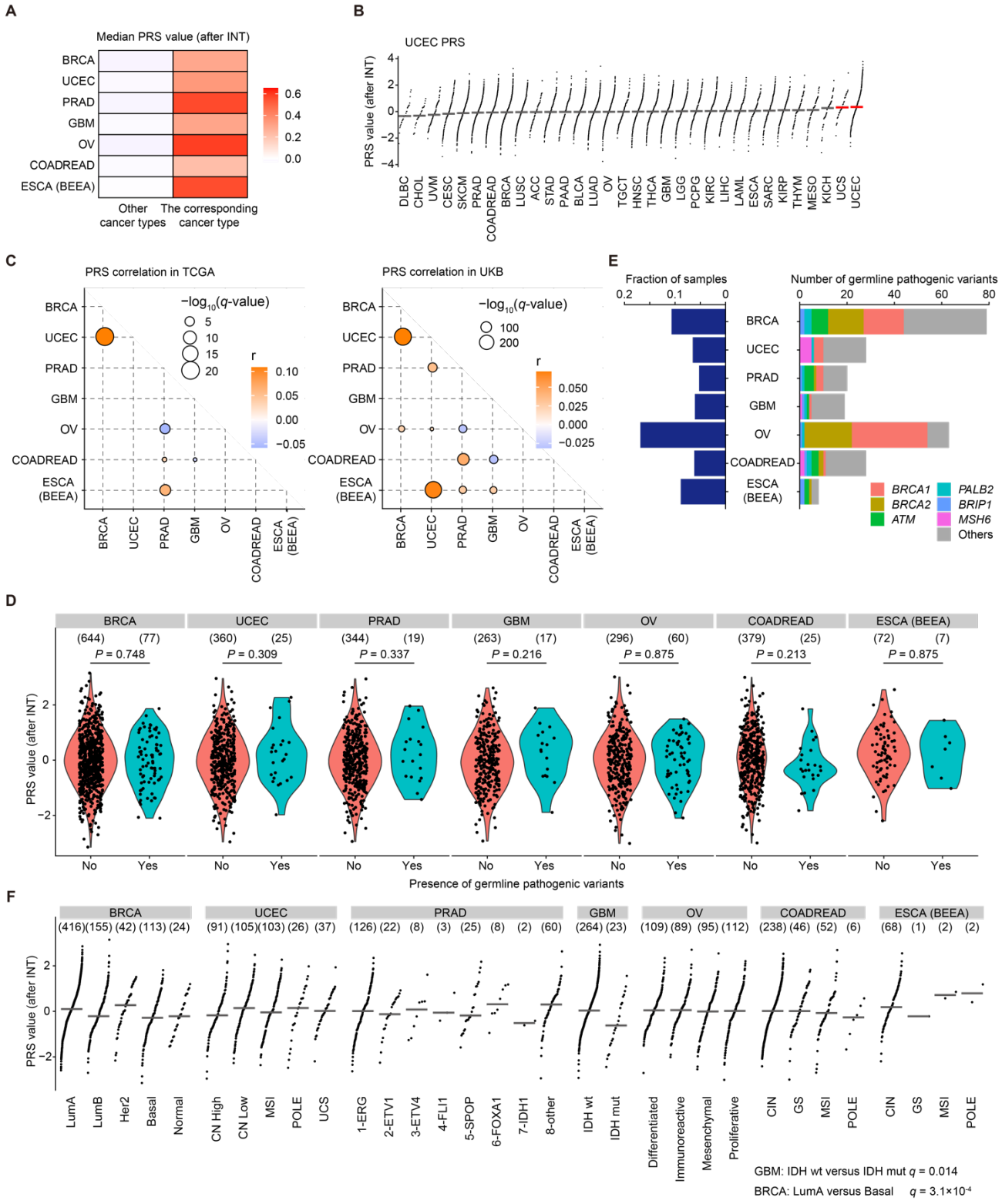
(A) A UMAP 2D plot derived from the first 20 PCs of genotype data of TCGA samples merged with the 1000 Genomes Project (1KG) samples. EUR, European; EAS, East Asian; AFR, African; SAS, South Asian; AMR, American.

(B, C) The same scatter plots as **(A)**, colored by self-reported race **(B)** and self-reported ethnicity **(C)**, showing high concordance between genetically inferred population and the self-reported data. 1KG samples were colored in gray.

(D) Probability of five ancestries estimated by ADMIXTURE (1).

(E) Mean intensity of signals from the probe on the genotype array for sex chromosomes. The intensity was divided by the mean intensity of autosomes. Asterisks indicate that the

sample itself was classified as atypical, but another sample derived from the same individual was classified as either female or male.



Supplementary Figure 5: Evaluation of PRSs using TCGA resources.

(A) Heatmap showing the median PRS values (after rank-based inverse normal transformation [INT]) in the corresponding cancer type (right; e.g., BRCA PRS values of BRCA samples) and other cancer types (left; e.g., BRCA PRS values of non-BRCA samples).

(B) UCEC PRS value distribution (after INT) across different cancer types for 7,965 TCGA samples of European ancestry. Each dot represents a sample and horizontal lines indicate the median PRS value in each cancer type.

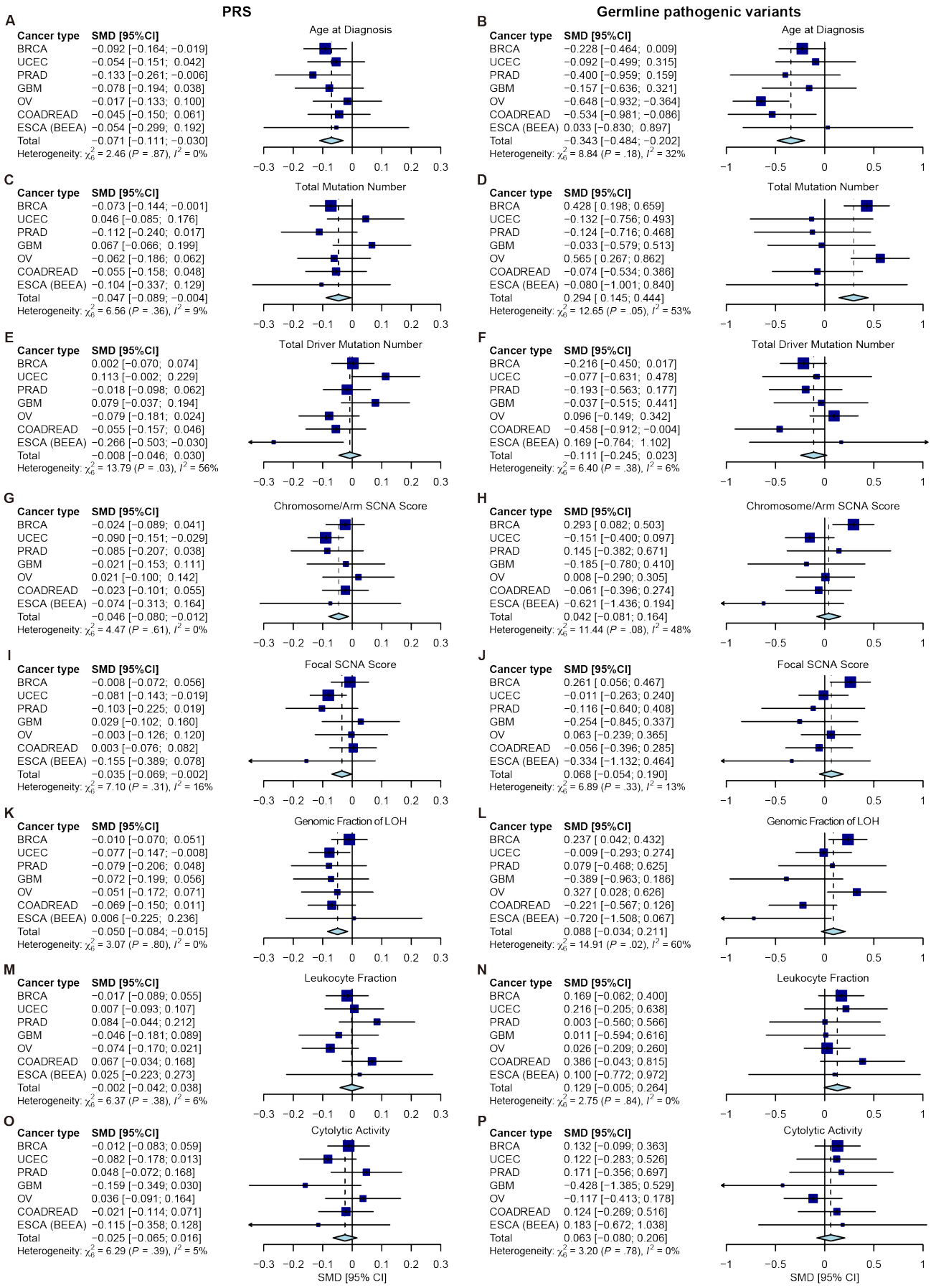
(C) Pair-wise correlations of PRS values for different cancer types in TCGA (left) and UKB (right) cohorts. The correlations in UKB were calculated for the individuals without any cancer. Orange and blue colors depict positive and negative associations. Two-sided Pearson's correlation test.

(D) Violin plots showing the distribution of PRS values (after INT) stratified by the presence of germline pathogenic variants in TCGA. Each dot represents a sample. Two-sided Welch's t-test.

(E) Fraction of TCGA samples harboring the germline pathogenic variants in each cancer type (left) and number of detected germline pathogenic variants in the TCGA cohort across different cancer types (right).

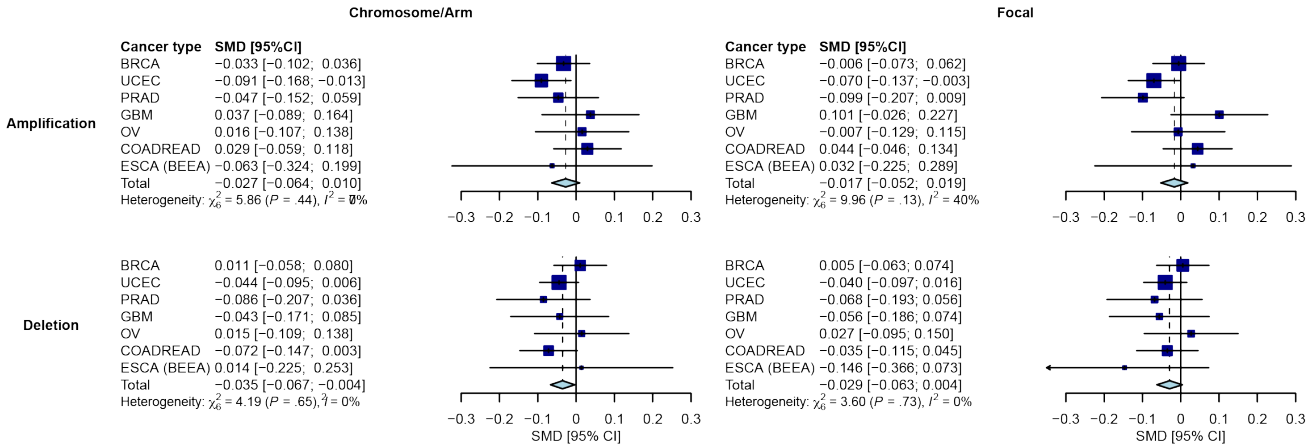
(F) PRS value distribution (after INT) across different subtypes for TCGA samples of European ancestry. Each dot represents a sample and horizontal lines indicate the median PRS value in each subtype. LumA, luminal A; Lum B, luminal B; Her2, Her2-enriched; Basal, basal-like; Normal, normal-like; CN, copy number; MSI, microsatellite instability; UCS, uterine carcinosarcoma; wt, wild-type; mut, mutated; CIN, chromosomal instability; GS, genomically stable. Two-sided Welch's t-test with Benjamini–Hochberg correction.

(D, F) Numbers in parentheses indicate numbers of samples examined.


PRS
Germline pathogenic variants
A
Age at Diagnosis
Cancer type SMD [95%CI]
BRCA −0.092 [−0.164; −0.019]
UCEC −0.054 [−0.151; 0.042]
PRAD −0.133 [−0.261; −0.006]
GBM −0.078 [−0.194; 0.038]
OV −0.017 [−0.133; 0.100]
COADREAD −0.045 [−0.150; 0.061]
ESCA (BEEA) −0.054 [−0.299; 0.192]
Total −0.071 [−0.111; −0.030]
Heterogeneity: χ^2_6 = 2.46 (P = .87), I^2 = 0%
B
Age at Diagnosis
Cancer type SMD [95%CI]
BRCA −0.228 [−0.464; 0.009]
UCEC −0.092 [−0.499; 0.315]
PRAD −0.400 [−0.959; 0.159]
GBM −0.157 [−0.636; 0.321]
OV −0.648 [−0.932; −0.364]
COADREAD −0.534 [−0.981; −0.086]
ESCA (BEEA) 0.033 [−0.830; 0.897]
Total −0.343 [−0.484; −0.202]
Heterogeneity: χ^2_6 = 8.84 (P = .18), I^2 = 32%
C
Total Mutation Number
Cancer type SMD [95%CI]
BRCA −0.073 [−0.144; −0.001]
UCEC 0.046 [−0.085; 0.176]
PRAD −0.112 [−0.240; 0.017]
GBM 0.067 [−0.066; 0.199]
OV −0.062 [−0.186; 0.062]
COADREAD −0.055 [−0.158; 0.048]
ESCA (BEEA) −0.104 [−0.337; 0.129]
Total −0.047 [−0.089; −0.004]
Heterogeneity: χ^2_6 = 6.56 (P = .36), I^2 = 9%
D
Total Mutation Number
Cancer type SMD [95%CI]
BRCA 0.428 [0.198; 0.659]
UCEC −0.132 [−0.756; 0.493]
PRAD −0.124 [−0.716; 0.468]
GBM −0.033 [−0.579; 0.513]
OV 0.565 [0.267; 0.862]
COADREAD −0.074 [−0.534; 0.386]
ESCA (BEEA) −0.080 [−1.001; 0.840]
Total 0.294 [0.145; 0.444]
Heterogeneity: χ^2_6 = 12.65 (P = .05), I^2 = 53%
E
Total Driver Mutation Number
Cancer type SMD [95%CI]
BRCA 0.002 [−0.070; 0.074]
UCEC 0.113 [−0.002; 0.229]
PRAD −0.018 [−0.098; 0.062]
GBM 0.079 [−0.037; 0.194]
OV −0.079 [−0.181; 0.024]
COADREAD −0.055 [−0.157; 0.046]
ESCA (BEEA) −0.266 [−0.503; −0.030]
Total −0.008 [−0.046; 0.030]
Heterogeneity: χ^2_6 = 13.79 (P = .03), I^2 = 56%
F
Total Driver Mutation Number
Cancer type SMD [95%CI]
BRCA −0.216 [−0.450; 0.017]
UCEC −0.077 [−0.631; 0.478]
PRAD −0.193 [−0.563; 0.177]
GBM −0.037 [−0.515; 0.441]
OV 0.096 [−0.149; 0.342]
COADREAD −0.458 [−0.912; −0.004]
ESCA (BEEA) 0.169 [−0.764; 1.102]
Total −0.111 [−0.245; 0.023]
Heterogeneity: χ^2_6 = 6.40 (P = .38), I^2 = 6%
G
Chromosome/Arm SCNA Score
Cancer type SMD [95%CI]
BRCA −0.024 [−0.089; 0.041]
UCEC −0.090 [−0.151; −0.029]
PRAD −0.085 [−0.207; 0.038]
GBM −0.021 [−0.153; 0.111]
OV 0.021 [−0.100; 0.142]
COADREAD −0.023 [−0.101; 0.055]
ESCA (BEEA) −0.074 [−0.313; 0.164]
Total −0.046 [−0.080; −0.012]
Heterogeneity: χ^2_6 = 4.47 (P = .61), I^2 = 0%
H
Chromosome/Arm SCNA Score
Cancer type SMD [95%CI]
BRCA 0.293 [0.082; 0.503]
UCEC −0.151 [−0.400; 0.097]
PRAD 0.145 [−0.382; 0.671]
GBM −0.185 [−0.780; 0.410]
OV 0.008 [−0.290; 0.305]
COADREAD −0.061 [−0.396; 0.274]
ESCA (BEEA) −0.621 [−1.436; 0.194]
Total 0.042 [−0.081; 0.164]
Heterogeneity: χ^2_6 = 11.44 (P = .08), I^2 = 48%
I
Focal SCNA Score
Cancer type SMD [95%CI]
BRCA −0.008 [−0.072; 0.056]
UCEC −0.081 [−0.143; −0.019]
PRAD −0.103 [−0.225; 0.019]
GBM 0.029 [−0.102; 0.160]
OV −0.003 [−0.126; 0.120]
COADREAD 0.003 [−0.076; 0.082]
ESCA (BEEA) −0.155 [−0.389; 0.078]
Total −0.035 [−0.069; −0.002]
Heterogeneity: χ^2_6 = 7.10 (P = .31), I^2 = 16%
J
Focal SCNA Score
Cancer type SMD [95%CI]
BRCA 0.261 [0.056; 0.467]
UCEC −0.011 [−0.263; 0.240]
PRAD −0.116 [−0.640; 0.408]
GBM −0.254 [−0.845; 0.337]
OV 0.063 [−0.239; 0.365]
COADREAD −0.056 [−0.396; 0.285]
ESCA (BEEA) −0.334 [−1.132; 0.464]
Total 0.068 [−0.054; 0.190]
Heterogeneity: χ^2_6 = 6.89 (P = .33), I^2 = 13%
K
Genomic Fraction of LOH
Cancer type SMD [95%CI]
BRCA −0.010 [−0.070; 0.051]
UCEC −0.077 [−0.147; −0.008]
PRAD −0.079 [−0.206; 0.048]
GBM −0.072 [−0.199; 0.056]
OV −0.051 [−0.172; 0.071]
COADREAD −0.069 [−0.150; 0.011]
ESCA (BEEA) 0.006 [−0.225; 0.236]
Total −0.050 [−0.084; −0.015]
Heterogeneity: χ^2_6 = 3.07 (P = .80), I^2 = 0%
L
Genomic Fraction of LOH
Cancer type SMD [95%CI]
BRCA 0.237 [0.042; 0.432]
UCEC −0.009 [−0.293; 0.274]
PRAD 0.079 [−0.468; 0.625]
GBM −0.389 [−0.963; 0.186]
OV 0.327 [0.028; 0.626]
COADREAD −0.221 [−0.567; 0.126]
ESCA (BEEA) −0.720 [−1.508; 0.067]
Total 0.088 [−0.034; 0.211]
Heterogeneity: χ^2_6 = 14.91 (P = .02), I^2 = 60%
M
Leukocyte Fraction
Cancer type SMD [95%CI]
BRCA −0.017 [−0.089; 0.055]
UCEC 0.007 [−0.093; 0.107]
PRAD 0.084 [−0.044; 0.212]
GBM −0.046 [−0.181; 0.089]
OV −0.074 [−0.170; 0.021]
COADREAD 0.067 [−0.034; 0.168]
ESCA (BEEA) 0.025 [−0.223; 0.273]
Total −0.002 [−0.042; 0.038]
Heterogeneity: χ^2_6 = 6.37 (P = .38), I^2 = 6%
N
Leukocyte Fraction
Cancer type SMD [95%CI]
BRCA 0.169 [−0.062; 0.400]
UCEC 0.216 [−0.205; 0.638]
PRAD 0.003 [−0.560; 0.566]
GBM 0.011 [−0.594; 0.616]
OV 0.026 [−0.209; 0.260]
COADREAD 0.386 [−0.043; 0.815]
ESCA (BEEA) 0.100 [−0.772; 0.972]
Total 0.129 [−0.005; 0.264]
Heterogeneity: χ^2_6 = 2.75 (P = .84), I^2 = 0%
O
Cytolytic Activity
Cancer type SMD [95%CI]
BRCA −0.012 [−0.083; 0.059]
UCEC −0.082 [−0.178; 0.013]
PRAD 0.048 [−0.072; 0.168]
GBM −0.159 [−0.349; 0.030]
OV 0.036 [−0.091; 0.164]
COADREAD −0.021 [−0.114; 0.071]
ESCA (BEEA) −0.115 [−0.358; 0.128]
Total −0.025 [−0.065; 0.016]
Heterogeneity: χ^2_6 = 6.29 (P = .39), I^2 = 5%
P
Cytolytic Activity
Cancer type SMD [95%CI]
BRCA 0.132 [−0.099; 0.363]
UCEC 0.122 [−0.283; 0.526]
PRAD 0.171 [−0.356; 0.697]
GBM −0.428 [−1.385; 0.529]
OV −0.117 [−0.413; 0.178]
COADREAD 0.124 [−0.269; 0.516]
ESCA (BEEA) 0.183 [−0.672; 1.038]
Total 0.063 [−0.080; 0.206]
Heterogeneity: χ^2_6 = 3.20 (P = .78), I^2 = 0%
−0.3 −0.2 −0.1 0 0.1 0.2 0.3
−1 −0.5 0 0.5 1
SMD [95% CI]

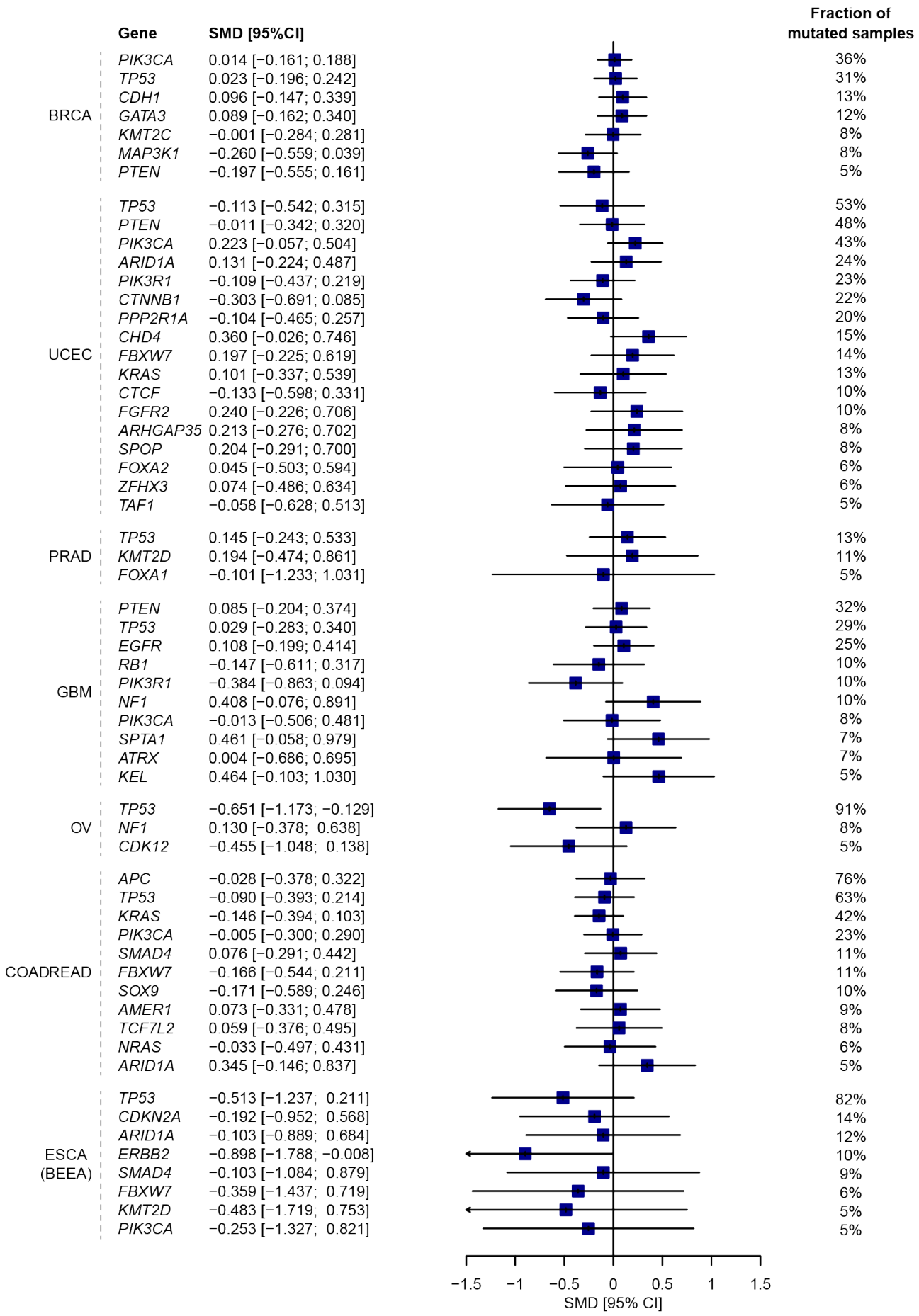
Supplementary Figure 6: Meta-analyses of germline–somatic associations across cancer types in TCGA.

Forest plots of cancer type-specific and pooled standardized mean difference (SMD) of PRS values **(A, C, E, G, I, K, M, and O)** and germline pathogenic variants **(B, D, F, H, J, L, N, and P)** for age at diagnosis **(A, B)**, total mutation number **(C, D)**, total driver mutation number **(E, F)**, chromosome/arm SCNA score **(G, H)**, focal SCNA score **(I, J)**, genomic fraction of LOH **(K, L)**, leukocyte fraction **(M, N)**, and cytolytic activity **(O, P)**. CI, Confidence interval.



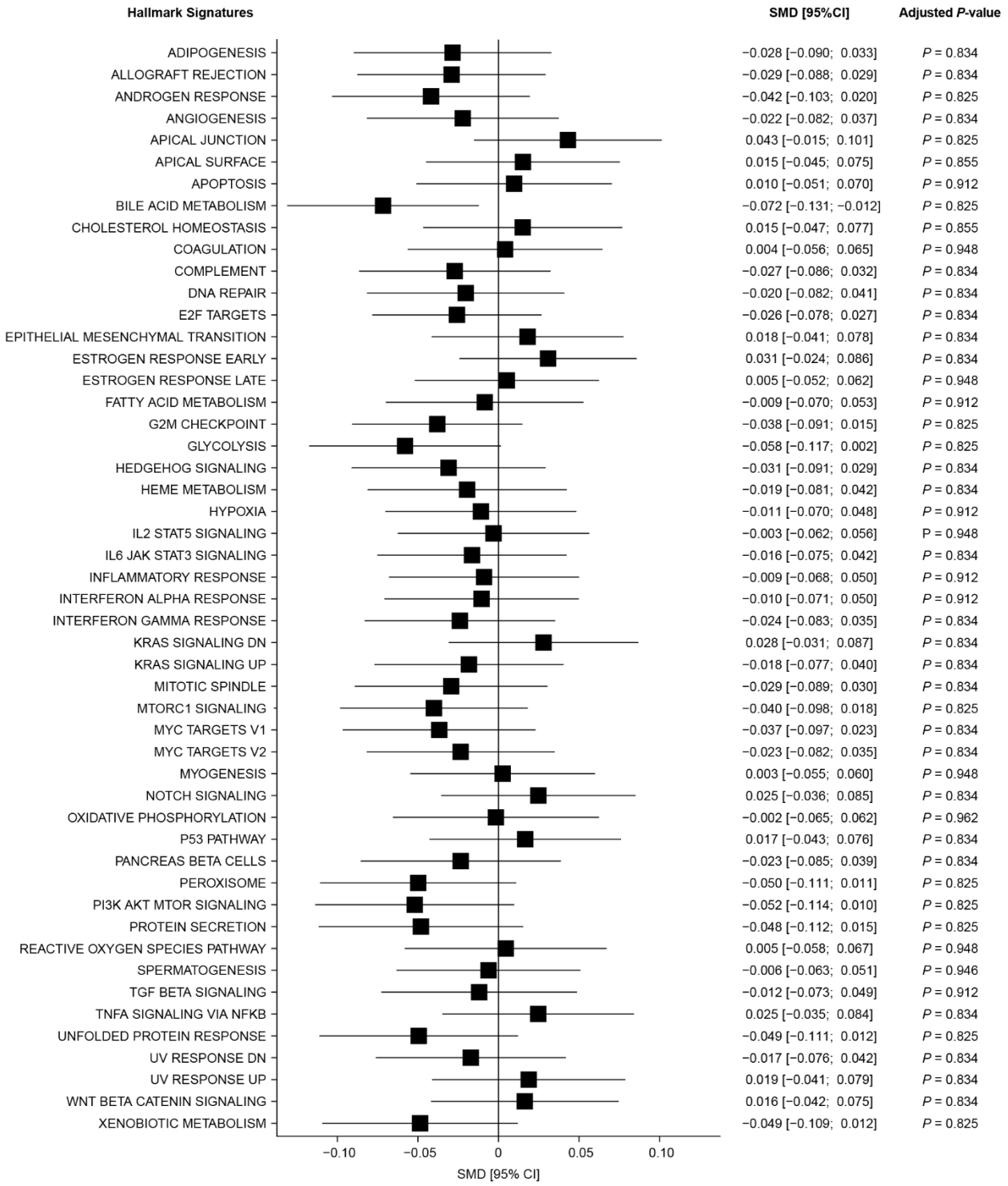
Supplementary Figure 7: Associations between PRS values and SCNAs in TCGA.

Forest plots of cancer type-specific and pooled SMD of PRS values for chromosome/arm amplification (top left), deletion (bottom left), focal amplification (top right), and deletion (bottom right) SCNA scores.


Gene
SMD [95%CI]
Fraction of mutated samples
BRCA
PIK3CA 0.014 [−0.161; 0.188] 36%
TP53 0.023 [−0.196; 0.242] 31%
CDH1 0.096 [−0.147; 0.339] 13%
GATA3 0.089 [−0.162; 0.340] 12%
KMT2C −0.001 [−0.284; 0.281] 8%
MAP3K1 −0.260 [−0.559; 0.039] 8%
PTEN −0.197 [−0.555; 0.161] 5%
UCEC
TP53 −0.113 [−0.542; 0.315] 53%
PTEN −0.011 [−0.342; 0.320] 48%
PIK3CA 0.223 [−0.057; 0.504] 43%
ARID1A 0.131 [−0.224; 0.487] 24%
PIK3R1 −0.109 [−0.437; 0.219] 23%
CTNNB1 −0.303 [−0.691; 0.085] 22%
PPP2R1A −0.104 [−0.465; 0.257] 20%
CHD4 0.360 [−0.026; 0.746] 15%
FBXW7 0.197 [−0.225; 0.619] 14%
KRAS 0.101 [−0.337; 0.539] 13%
CTCF −0.133 [−0.598; 0.331] 10%
FGFR2 0.240 [−0.226; 0.706] 10%
ARHGAP35 0.213 [−0.276; 0.702] 8%
SPOP 0.204 [−0.291; 0.700] 8%
FOXA2 0.045 [−0.503; 0.594] 6%
ZFHX3 0.074 [−0.486; 0.634] 6%
TAF1 −0.058 [−0.628; 0.513] 5%
PRAD
TP53 0.145 [−0.243; 0.533] 13%
KMT2D 0.194 [−0.474; 0.861] 11%
FOXA1 −0.101 [−1.233; 1.031] 5%
GBM
PTEN 0.085 [−0.204; 0.374] 32%
TP53 0.029 [−0.283; 0.340] 29%
EGFR 0.108 [−0.199; 0.414] 25%
RB1 −0.147 [−0.611; 0.317] 10%
PIK3R1 −0.384 [−0.863; 0.094] 10%
NF1 0.408 [−0.076; 0.891] 10%
PIK3CA −0.013 [−0.506; 0.481] 8%
SPTA1 0.461 [−0.058; 0.979] 7%
ATRX 0.004 [−0.686; 0.695] 7%
KEL 0.464 [−0.103; 1.030] 5%
OV
TP53 −0.651 [−1.173; −0.129] 91%
NF1 0.130 [−0.378; 0.638] 8%
CDK12 −0.455 [−1.048; 0.138] 5%
COADREAD
APC −0.028 [−0.378; 0.322] 76%
TP53 −0.090 [−0.393; 0.214] 63%
KRAS −0.146 [−0.394; 0.103] 42%
PIK3CA −0.005 [−0.300; 0.290] 23%
SMAD4 0.076 [−0.291; 0.442] 11%
FBXW7 −0.166 [−0.544; 0.211] 11%
SOX9 −0.171 [−0.589; 0.246] 10%
AMER1 0.073 [−0.331; 0.478] 9%
TCF7L2 0.059 [−0.376; 0.495] 8%
NRAS −0.033 [−0.497; 0.431] 6%
ARID1A 0.345 [−0.146; 0.837] 5%
ESCA (BEEA)
TP53 −0.513 [−1.237; 0.211] 82%
CDKN2A −0.192 [−0.952; 0.568] 14%
ARID1A −0.103 [−0.889; 0.684] 12%
ERBB2 −0.898 [−1.788; −0.008] 10%
SMAD4 −0.103 [−1.084; 0.879] 9%
FBXW7 −0.359 [−1.437; 0.719] 6%
KMT2D −0.483 [−1.719; 0.753] 5%
PIK3CA −0.253 [−1.327; 0.821] 5%
−1.5 −1 −0.5 0 0.5 1 1.5
SMD [95% CI]

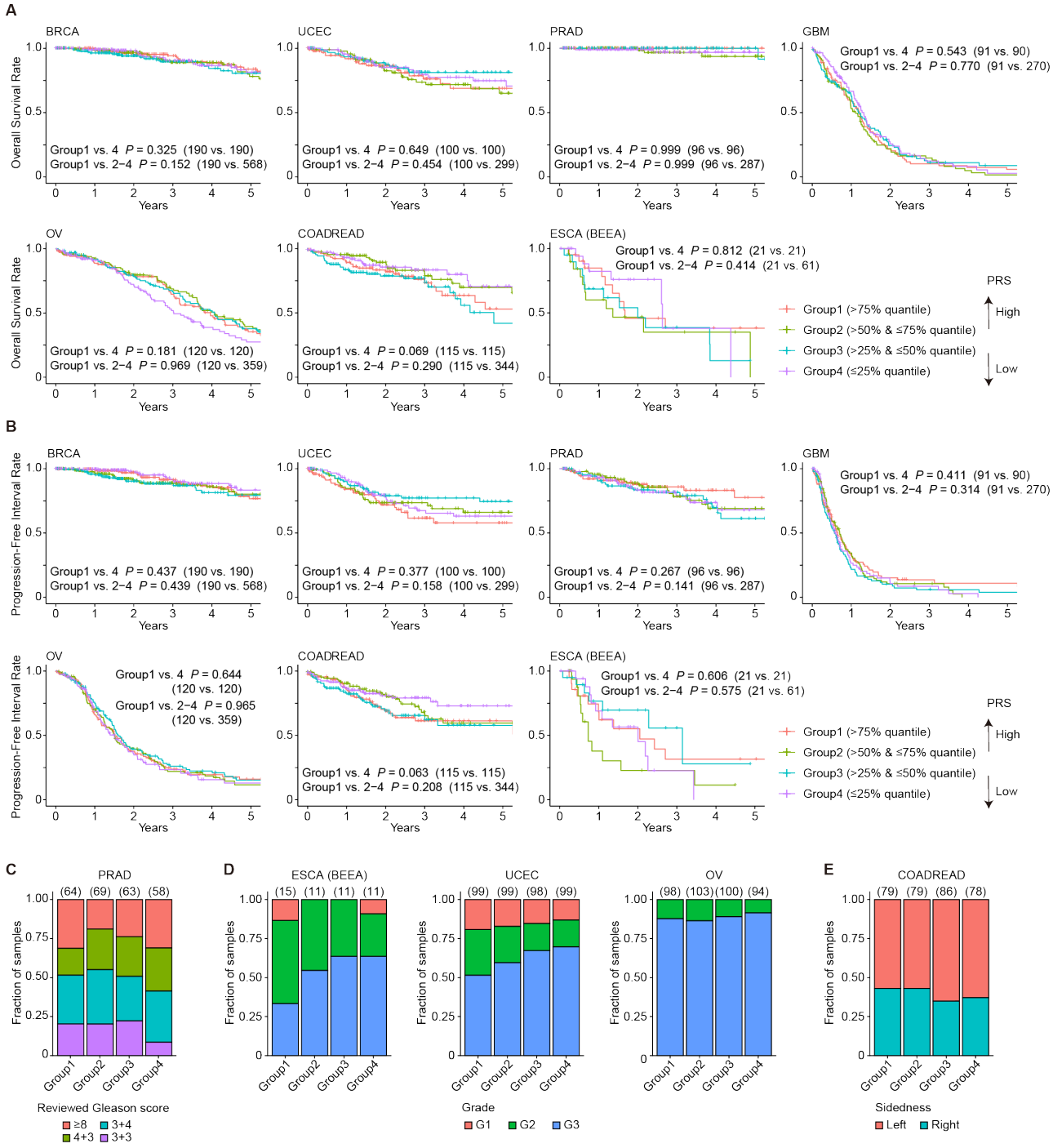
Supplementary Figure 8: Associations between PRS values and driver mutations in TCGA.

The association between each driver mutation and PRS values in individual cancer types. Generalized linear model (GLM) using subtypes and germline pathogenic variants as covariates. Driver genes detected in less than 5% of the samples were excluded from visualization. GLM for *SPOP* in PRAD and *IDH1* in GBM did not converge because they were subtype-defining drivers; hence, they were excluded from this figure.



Supplementary Figure 9: Meta-analyses of associations between PRS values and hallmark signatures in TCGA.

The associations between PRS values and 50 hallmark signatures from meta-analyses. Numbers on the right represent pooled SMDs with 95% CIs and *P*-values (after Benjamini-Hochberg adjustment).



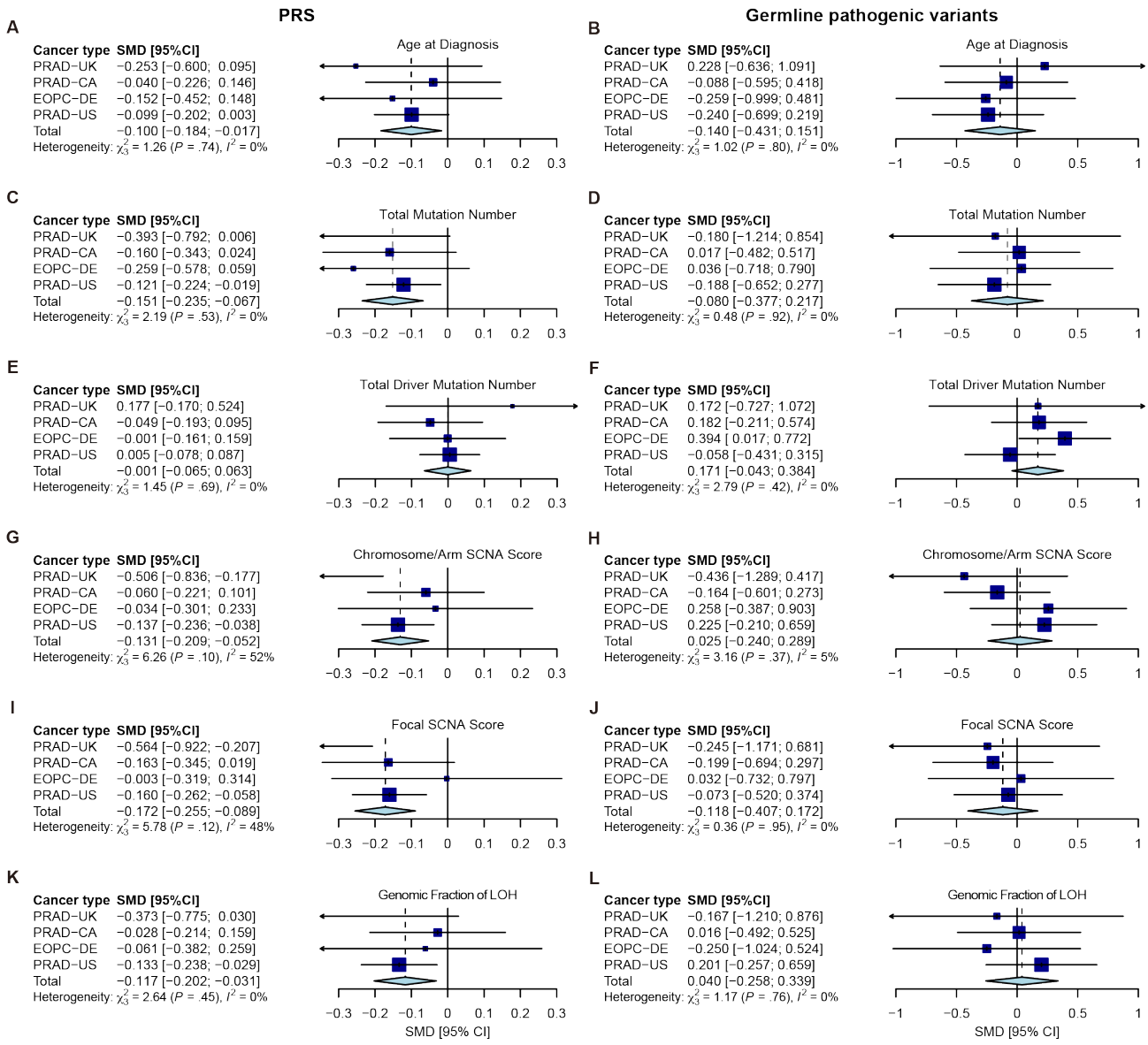
Supplementary Figure 10: Associations between PRS values and clinical and other features in TCGA.

(A, B) Kaplan-Meier survival curves of overall survival **(A)** and progression-free interval **(B)** for BRCA, UCEC, PRAD, GBM, OV, COADREAD, and ESCA patients grouped by PRS value quartiles. *P*-values were calculated using the Cox proportional hazards model.

(C) Reviewed Gleason scores for PRAD patients grouped by PRAD PRS value quartiles.

(D) Histological grade for ESCA (left), UCEC (middle), and OV (right) patients grouped by PRS value quartiles. GX (unknown grade) samples were excluded from the analysis.

(E) Colorectal cancer sidedness for COADREAD patients grouped by PRS value quartiles. Numbers in parentheses indicate numbers of samples examined.



Supplementary Figure 11: Meta-analyses of germline–somatic associations across four PCAWG PRAD cohorts.

Forest plots of cancer type-specific and pooled SMD of PRS values **(A, C, E, G, I, and K)** and germline pathogenic variants **(B, D, F, H, J, and L)** for age at diagnosis **(A, B)**, total mutation number **(C, D)**, total driver mutation number **(E, F)**, chromosome/arm SCNA score **(G, H)**, focal SCNA score **(I, J)**, and genomic fraction of LOH **(K, L)**. PRAD-UK, Prostate Adenocarcinoma – United Kingdom; PRAD-CA, Prostate Adenocarcinoma – Canada; EOPC-DE, Early Onset Prostate Cancer – Germany; PRAD-US, Prostate Adenocarcinoma – United States. We note that PRAD-US is the same dataset as TCGA PRAD cohort.

Supplementary Tables

Supplementary Table 1: GWAS summary statistics used for PRS.

Study code	Cancer type	GWAS	Number of samples		Available variants	Heritability reported in the original articles (SE)		LDSC heritability (SE)
			Case	Control		LDSC[a]	GCTA-GREML-LDMS[a]	
BRCA	Breast cancer	Ref. [2]	133,384	113,789	genome-wide variants	-	-	0.11 (0.0067)
PRAD	Prostate cancer	Ref. [3]	79,194	61,112	genome-wide variants	-	-	0.11 (0.013)
COADREAD	Colorectal cancer	Ref. [4]	58,131	67,347	lead variants[b]	-	0.118	-
SKCM	Skin cutaneous melanoma	Ref. [5]	30,134	81,415	variants with $P < 1 \times 10^{-4}$	0.098 (0.02)	-	-
LUCA	Lung cancer	Ref. [6]	29,266	56,450	variants with $P < 1 \times 10^{-5}$	0.089 (0.0131)	-	-
LUAD	Lung adenocarcinoma	Ref. [6]	11,273	55,483	variants with $P < 1 \times 10^{-5}$	0.067 (0.0094)	-	-
LUSC	Lung squamous cell carcinoma	Ref. [6]	7,426	55,627	variants with $P < 1 \times 10^{-5}$	0.062 (0.013)	-	-
UCEC	Uterine endometrial carcinoma	Ref. [7]	12,906	108,979	genome-wide variants	-	-	0.042 (0.012)
OV	Ovarian serous carcinoma	Ref. [8]	14,049	40,941	genome-wide variants	-	-	0.043 (0.0093)
ESCA (BEEA)[c]	Esophageal adenocarcinoma and Barrett's esophagus	Ref. [9]	10,279	17,159	genome-wide variants	-	-	0.16 (0.024)
ESCA (EA)[c]	Esophageal adenocarcinoma	Ref. [9]	4,112	17,159	genome-wide variants	-	-	0.16 (0.025)
GBM	Glioblastoma multiforme	Ref. [10]	4,512	10,582	selected variants[d]	-	-	-
HNSC	Head and neck squamous cell carcinoma	Ref. [11]	6,034	6,585	genome-wide variants	-	-	0.057 (0.023)
CESC	cervical cancer	Ref. [12]	2,866	6,481	genome-wide variants	-	-	0.050 (0.087)

[a] Methods used to estimate heritability

[b] Independent lead variants with $P < 1 \times 10^{-5}$

[c] "ESCA (BEEA)" is a GWAS for Barrett's esophagus and esophageal adenocarcinoma, which considered both diseases as a single entity because of a very high genetic correlation ($r_g \sim 1$) between these two diseases (9). On the other hand, "ESCA (EA)" is a GWAS only for esophageal adenocarcinoma.

[d] Variants with the nominal significance in a previous GWAS meta-analysis (13). We used GWAS summary statistics that was not stratified by age.

SE, standard error; LDSC, linkage disequilibrium score regression (14); GCTA-GREML-LDMS, linkage disequilibrium- and minor allele frequency-stratified multicomponent genomic restricted maximum likelihood analysis, implemented in genome-wide complex trait analysis (15).

Supplementary Table 2: Definition of cases and controls in UKB.

Study code	Number of samples		Inclusion criteria of case samples				Inclusion criteria of control samples[a]
	Case	Control	Self-reported diseases	ICD-10 codes	Histology	Sex	
BRCA	11,469	142,892	breast cancer	C50	-	female	female
PRAD	7,100	126,652	prostate cancer	C61	-	male	male
COADREAD	4,525	269,544	colon cancer/sigmoid cancer, large bowel cancer/colorectal cancer, rectal cancer	C18, C19, C20	-	both	both
SKCM	4,242	269,544	malignant melanoma	C43	-	both	both
LUCA	1,996	269,544	lung cancer, non-small cell lung cancer, small cell lung cancer	C34	-	both	both
LUAD	868	269,544	-	C34	Adenocarcinoma, intestinal type; Adenocarcinoma in tubulovillous adenoma; Adenocarcinoma in villous adenoma; Adenocarcinoma (NOS)	both	both
LUSC	510	269,544	-	C34	Squamous cell carcinoma; Squamous cell carcinoma, keratinizing; Squamous cell carcinoma, large cell, non-keratinizing; Squamous cell carcinoma, micro-invasive; Squamous cell carcinoma, small cell, non-keratinizing	both	both
UCEC	1,587	142,892	uterine/endometrial cancer	C54	-	female	female

OV	367	142,892	-	C56	Papillary serous cystadenocarcinoma; Serous cystadenocarcinoma; Serous cystadenoma, borderline malignancy; Serous papillary cystic tumor of borderline malignancy; Serous surface papillary carcinoma	female	female
ESCA (BEEA)	482	269,544	-	C15	Adenocarcinoma, NOS; Adenocarcinoma, intestinal type; Adenocarcinoma in tubulovillous adenoma	both	both
ESCA (EA)	482	269,544	-	C15	Adenocarcinoma, NOS; Adenocarcinoma, intestinal type; Adenocarcinoma in tubulovillous adenoma	both	both
GBM	262	269,544	-	C71	Giant cell glioblastoma; Glioblastoma (NOS)	both	both
HNSC	689	269,544	oropharynx / oropharyngeal cancer, tongue cancer	C01, C02, C03, C04, C05, C06, C09, C10, C13, C14	-	both	both
CESC	1,482	142,892	cervical cancer	C53	-	female	female

[a] Control samples were restricted to the samples without any cancer diagnosis or self-reported cancer.

ICD, International Classification of Diseases; NOS, not otherwise specified.

Supplementary Table 3: Parameters and numbers of variants used for PRS.

Study code	The best PRS method	The best C+T threshold	The best lassosum parameters		Number of variants used										
			s	lambda	C+T (5×10^{-8})	C+T (1×10^{-7})	C+T (1×10^{-6})	C+T (1×10^{-5})	C+T (1×10^{-4})	C+T (1×10^{-3})	C+T (1×10^{-2})	C+T (1×10^{-1})	lassosum[a]	PRScs	LDpred2
BRCA	LDpred2	1×10^{-4}	0.5	0.00428133	364	395	555	876	1,700	4,679	17,474	80,031	112,710	1,081,107	1,142,637
PRAD	lassosum	1×10^{-4}	0.5	0.00545559	374	398	525	794	1,545	5,012	26,005	170,009	168,700	1,082,596	1,204,849
COADREAD	C+T (1×10^{-5})	1×10^{-5}	-	-	35	39	51	61	-	-	-	-	-	-	-
SKCM	C+T (1×10^{-4})	1×10^{-4}	-	-	108	120	156	280	672	-	-	-	-	-	-
LUCA	C+T (1×10^{-5})	1×10^{-5}	-	-	48	50	82	133	-	-	-	-	-	-	-
LUAD	C+T (1×10^{-5})	1×10^{-5}	-	-	19	23	33	74	-	-	-	-	-	-	-
LUSC	C+T (1×10^{-5})	1×10^{-5}	-	-	29	30	45	90	-	-	-	-	-	-	-
UCEC	lassosum	1×10^{-1}	0.9	0.00695193	17	19	32	75	329	1,871	11,415	64,118	529,365	1,075,588	1,126,937
OV	lassosum	5×10^{-8}	0.9	0.01128838	35	39	50	97	395	2,757	21,477	161,855	144,810	1,082,662	1,207,877
ESCA (BEEA)	lassosum	1×10^{-1}	0.9	0.00885867	15	19	44	94	295	1,700	11,964	83,657	601,980	1,074,317	1,076,634
ESCA (EA)	lassosum	1×10^{-1}	0.5	0.01128838	5	6	10	42	211	1,451	11,352	84,768	356,743	1,077,063	1,082,942
GBM	C+T (1×10^{-1})	1×10^{-1}	-	-	20	20	25	48	144	393	778	910	-	-	-
HNSC	PRScs	1×10^{-1}	0.9	0.00162378	2	2	10	41	210	1,426	9,814	58,334	4,141,292	931,477	994,382
CESC	lassosum	1×10^{-3}	0.9	0.02976351	2	3	5	13	50	346	2,236	14,463	2,814	229,973	223,583

[a] 80 PRSs were calculated for each study according to the hyperparameters. This table shows the data of the most predictable PRS.

Supplementary Table 4: Evaluation of PRSs using UKB resources.

Study code	PRS	ΔAUC (full model - reduced model)	AUC (full model)	AUC (reduced model)	AUC evaluated without covariates (Phenotype ~ PRS)	Nagelkerke's R^2	*P* (Delong's test)	Adjusted *P* (Benjamini-Hochberg)
BRCA	LDpred2	8.03×10^{-2}	0.696	0.616	0.659	5.55×10^{-2}	2.41×10^{-239}	1.68×10^{-238}
UCEC	lassosum	7.95×10^{-2}	0.761	0.681	0.697	1.10×10^{-1}	1.23×10^{-29}	4.30×10^{-29}
SKCM	best C+T ($P < 1 \times 10^{-4}$)	6.75×10^{-2}	0.682	0.615	0.645	2.61×10^{-2}	6.01×10^{-80}	2.81×10^{-79}
PRAD	lassosum	6.50×10^{-2}	0.813	0.748	0.701	7.98×10^{-2}	1.62×10^{-242}	2.27×10^{-241}
GBM	best C+T ($P < 1 \times 10^{-1}$)	4.13×10^{-2}	0.758	0.717	0.651	2.16×10^{-2}	2.42×10^{-5}	4.23×10^{-5}
OV	lassosum	2.93×10^{-2}	0.717	0.687	0.620	1.42×10^{-2}	4.91×10^{-5}	7.64×10^{-5}
COADREAD	best C+T ($P < 1 \times 10^{-5}$)	1.73×10^{-2}	0.724	0.707	0.594	1.15×10^{-2}	1.02×10^{-27}	2.87×10^{-27}
ESCA (BEEA)	lassosum	1.46×10^{-2}	0.819	0.804	0.614	1.23×10^{-2}	1.40×10^{-5}	2.80×10^{-5}
ESCA (EA)	lassosum	1.01×10^{-2}	0.814	0.804	0.594	8.75×10^{-3}	3.66×10^{-4}	5.13×10^{-4}
CESC	lassosum	1.02×10^{-2}	0.563	0.553	0.537	1.58×10^{-3}	2.33×10^{-2}	2.51×10^{-2}
HNSC	PRScs	9.88×10^{-3}	0.672	0.662	0.572	5.61×10^{-3}	2.13×10^{-2}	2.48×10^{-2}
LUCA	best C+T ($P < 1 \times 10^{-5}$)	2.24×10^{-3}	0.729	0.727	0.543	1.35×10^{-3}	4.83×10^{-3}	6.15×10^{-3}
LUSC	best C+T ($P < 1 \times 10^{-5}$)	7.71×10^{-4}	0.778	0.777	0.539	7.46×10^{-4}	4.18×10^{-1}	4.18×10^{-1}
LUAD	best C+T ($P < 1 \times 10^{-5}$)	1.30×10^{-2}	0.743	0.730	0.580	6.49×10^{-3}	5.01×10^{-6}	1.17×10^{-5}

Reduced model included all covariates but PRS (i.e., age, sex, and top 20 genetic principal components [PCs]). AUC, area under the curve.

Supplementary Table 5: Number of patients without sex aneuploidy in TCGA.

TCGA Cancer code	Corresponding study code	EUR	EAS	AFR	AMR	SAS	Admix
ACC	-	64	2	0	0	0	12
BLCA	-	283	38	12	3	1	29
BRCA	BRCA	759	52	116	6	6	82
CESC	-	151	19	17	24	0	62
CHOL	-	27	2	2	2	0	3
COAD	COADREAD	329	12	46	0	0	24
DLBC	-	27	14	1	1	1	4
ESCA[a]	ESCA (BEEA)	115	42	3	1	0	17
GBM	GBM	362	5	25	3	2	45
HNSC	-	405	5	36	5	6	47
KICH	-	55	1	4	1	1	4
KIRC	-	402	6	30	13	0	39
KIRP	-	194	5	38	4	1	28
LAML	-	162	1	9	1	0	6
LGG	-	403	8	9	10	2	43
LIHC	-	152	153	14	4	0	22
LUAD	-	426	9	33	2	0	26
LUSC	-	368	8	16	0	0	23
MESO	-	79	0	0	0	1	5
OV	OV	479	13	23	4	4	38
PAAD	-	152	11	6	2	0	7
PCPG	-	140	3	13	2	4	11
PRAD	PRAD	383	8	40	2	0	37
READ	COADREAD	131	1	2	0	0	8
SARC	-	205	6	12	3	0	18
SKCM	-	437	12	1	5	0	7
STAD	-	254	72	8	5	0	58
TGCT	-	116	4	1	11	0	16
THCA	-	335	40	16	21	15	50
THYM	-	91	12	5	2	1	11
UCEC	UCEC	361	30	77	8	2	48
UCS	UCEC	38	3	4	0	0	4
UVM	-	80	0	0	0	0	0

[a] We used only esophagus adenocarcinoma (*n* = 82 for EUR samples) for the germline–somatic association study to match the histology with the GWAS cohort.

EUR, European; EAS; East Asian; AFR, African; AMR, American; SAS, South Asian.

Supplementary Table 6: The gene-level associations for UCEC and BRCA in the cell cycle pathway.

	UCEC		BRCA	
Gene	***Z*-score**	***P*-value**	***Z*-score**	***P*-value**
CDKN2B	2.71	2.4×10^{-3}	6.54	6.9×10^{-13}
CDKN2A	1.52	4.6×10^{-2}	6.08	1.1×10^{-11}
CCND1	2.49	3.8×10^{-3}	5.55	3.9×10^{-10}
CDKN1A	1.28	7.3×10^{-2}	3.92	3.1×10^{-6}
CDKN2C	-0.68	7.1×10^{-1}	3.29	7.4×10^{-5}
CCNE1	0.18	3.6×10^{-1}	2.35	1.4×10^{-3}
E2F3	0.31	2.9×10^{-1}	2.13	2.3×10^{-3}
CDK6	-0.31	5.3×10^{-1}	1.61	1.1×10^{-2}
E2F1	1.35	6.7×10^{-2}	1.71	1.1×10^{-2}
CDKN1B	0.43	2.8×10^{-1}	1.60	1.3×10^{-2}
CDK2	1.31	7.8×10^{-2}	0.67	1.2×10^{-1}
CDK4	-0.29	5.6×10^{-1}	0.26	2.1×10^{-1}
CCND3	0.40	2.5×10^{-1}	-0.48	3.9×10^{-1}
CCND2	2.67	2.0×10^{-3}	-1.08	6.5×10^{-1}
RB1	1.42	5.3×10^{-2}	-1.26	7.5×10^{-1}

Supplementary Materials and Methods

UK Biobank (UKB)

UKB is a population-based prospective cohort with approximately 500,000 participants. The detailed characteristics of UKB were described elsewhere (16). The UKB analysis was conducted via the application (47821). We used the imputed genotype data (release version 3) and excluded variants with an INFO score < 0.5. To minimize the variability due to population stratification, we restricted our analyses to white British individuals who met all of the following criteria: (i) the self-reported ethnicity was white British; (ii) the genetic ethnicity defined in the UKB data-field 22006 was “Caucasian”; (iii) the genetic sex was matched to the self-reported sex and was not inferred as sex chromosome aneuploidy; (iv) the individuals were not outliers in heterozygosity or missing rate; (v) the individuals were not excluded from kinship inference process. Subsequently, for each pair of relatives inferred to be third degree or closer, individuals with a higher missing rate were removed, resulting in 335,048 unrelated individuals.

Heritability estimates

We evaluated common SNP heritability and genetic correlation using LD score regression (14). We used LD scores calculated for HapMap3 variants using European samples in the 1000 Genomes Project (14), and the baseline LD model v2.2 (17) as the annotations used for the heritability analysis. The estimated heritability and genetic correlation were converted to the liability scale, assuming that the population prevalence matched the prevalence in the UKB analysis data.

Pathway enrichment analysis

We ran MAGMA (18) to summarize GWAS summary statistics of UCEC and BRCA into the

gene-level statistics and to perform a pathway enrichment analysis. We used the 1000 Genomes Project (phase 3) (19) European samples as the reference panel. Pathway enrichment was evaluated for the canonical cancer pathways defined in TCGA (20), and the enrichment with false discovery rate < 0.05 was considered to be significant. Among genes in pathways significantly enriched in both UCEC and BRCA, genes with MAGMA gene-level *P*-value < 0.05 were considered as those associated with both cancer types. To calculate the subset of the PRS values, we chose three genes that were associated with both cancer types, and used the variants overlapping with the gene body of these genes among the variants used for the best PRSs.

The Cancer Genome Atlas (TCGA)

Data acquisition

Germline genotype data for common variants were obtained using Affymetrix Genome-Wide Human SNP Array 6.0. Samples that passed quality control were chosen based on TCGA PanCan Atlas Project. Birdseed files representing 905,600 variants for 11,520 normal tissues were downloaded from GDC Portal (database of Genotypes and Phenotypes (dbGaP) accession phs000178.v10.p8) and annotated with the corresponding annotation file (release 35). Somatic and germline mutational data in Mutation Annotation Format (compiled by TCGA PanCanAtlas MC3 Working Group and Germline Working Group, respectively), RNA-seq expression data, In Silico Admixture Removal (ISAR)-corrected copy number segment data, and clinical information were downloaded from GDC (https://gdc.cancer.gov/about-data/publications/pancanatlas and https://gdc.cancer.gov/about-data/publications/PanCanAtlas-Germline-AWG). TCGA subtype information was obtained from original cancer type-specific studies or PanCan Atlas studies (21–24). Evaluated subtypes were shown in **Supplementary Figure 5F**. Germline variants classified as 'pathogenic' or 'likely pathogenic' by TCGA PanCan Atlas Germline Working Group (25) were

considered as germline pathogenic variants.

Quality control of genotype data

We first excluded low confidence genotypes (the Birdseed [26] metric, confidence, larger than 0.05) per sample. As a result, genotypes were retained in at least one sample for 868,179 variants. Samples were excluded if the genotype call rate was less than 0.95 or heterozygosity was outside of three standard deviations from the mean value. Duplicates and closely related individuals were detected with King (27). For a pair within the second degree, one individual with a lower call rate was excluded. For multiple samples from a single individual, we discarded all of the samples unless they were all inferred as duplicates of the same individual. Subsequently, we chose one sample per individual prioritizing peripheral blood samples and samples with the highest call rate. Variants were excluded if (i) the call rate was less than 0.97, (ii) minor allele frequency was less than 0.005, or (iii) *P*-value of Hardy-Weinberg Equilibrium (HWE) with the mid-*P* adjustment was less than 1×10^{-6} in at least one population. Among variants with low HWE *P*-values, one variant, rs10052657, was rescued because it was associated with esophageal squamous cell carcinoma in GWAS catalog (28). After quality control, we retained 751,839 variants for 10,228 individuals for imputation.

Imputation

We first matched the strand of TCGA genotype data with those of 1000 Genomes Project (phase 3) (19). For palindrome variants, we utilized the LD structure of the European population by using Genotype Harmonizer (29) (v1.4.23, https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer) with its default settings, and additionally, variants were retained if the difference in allele frequencies between the two datasets were less than 0.35. Statistical phasing by Shapeit4 (30) (v4.1.2,

https://odelaneau.github.io/shapeit4/) were conducted with the option "--mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10m" for more accurate phasing than the default settings. Subsequently, we conducted imputation with Minimac4 (31) (v1.0.1, http://genome.sph.umich.edu/wiki/Minimac4), and kept variants with (i) R^2 larger than 0.7 and (ii) minor allele frequency larger than 0.005, resulting in 22,044,972 variants.

Ancestry analysis

We merged non-palindrome autosome variants of TCGA genotype data with those of 1000 Genomes Project. After LD pruning and excluding high LD regions (32), we performed principal component analysis (PCA) with smartpca (33,34) (EIGENSOFT v6.1.4, https://www.hsph.harvard.edu/alkes-price/software/) and subsequently conducted dimension reduction for the first 20 PCs by UMAP (35,36). Because clusters were clearly separated in a two-dimensional plot, we manually inspected the clusters and found that they matched well with self-reported race and ethnicity (**Supplementary Figure 4**). We also used a likelihood model implemented in ADMIXTURE (1) (v1.3.0, http://dalexander.github.io/admixture/) with the number of ancestries of five. Samples were considered as admixed if the sum of the probabilities of all ancestries except for the best-inferred ancestry was greater than 0.2.

Sex inference

We obtained the probe intensity of the genotyping array by running Birdseed (26) (v1.4, https://www.broadinstitute.org/birdsuite/birdsuite-install) for non-tumor samples. We used the same batches as used in the original genotyping procedure by TCGA. We calculated the mean intensity in sex chromosomes in proportion to that in autosomal chromosomes, and classified samples as male if the relative mean intensity of chromosome X was less than 0.8 and that of chromosome Y was greater than 0.5. Similarly, we classified samples as female if the relative mean intensity of chromosome X was greater than 0.85 and that of chromosome

Y was less than 0.7. Other samples were considered to have atypical sex chromosomes and were not used to analyze germline–somatic association.

Somatic mutation analysis

Somatic mutations with low variant allele frequency (less than 0.05) were excluded, and total somatic mutation numbers were calculated as all mutations (both synonymous and non-synonymous) in coding regions.

Tumors were classified as hypermutated if they met the criteria below: (i) samples harboring pathogenic somatic *POLE*/*POLD1* mutations (37), (ii) microsatellite instable samples, determined using MSIsensor as previously described (38), (iii) samples with total somatic mutation number above 75th percentile + 3 * interquartile range in each cancer type, after excluding (i) and (ii) samples. Hypermutator samples were excluded when evaluating somatic mutations to ensure accurate and representative results for the majority of tumor samples.

Driver gene list for each cancer type was obtained from TCGA PanCan Atlas Project (39). Non-synonymous somatic mutations in the driver genes of the corresponding cancer type were considered as somatic driver mutations. Driver mutation numbers were calculated as a sum of somatic driver mutation numbers for each sample. Meta-analysis was performed if the gene was considered as a driver in three or more cancer types.

Somatic copy number analysis

To quantify the genomic fraction of LOH for each sample, we used publicly available ISAR-corrected ABSOLUTE copy number calls. X and Y chromosome regions were discarded from the analysis. The genomic fraction of LOH was defined as 100 times the total length of LOH regions/length of the genome, as previously described (40).

To distinguish focal and chromosome/arm SCNAs, we used GISTIC 2.0 (v2.0.23, https://github.com/broadinstitute/gistic2) applied to the ISAR-corrected ABSOLUTE

segmented file. Default parameters were used, except that a broad length cutoff was 0.7 chromosome arms. Focal and chromosome/arm SCNA scores were calculated using GISTIC2.0 output as previously described (40,41). First, for each sample, each focal-event log2 copy number ratio from GISTIC2.0 was classified into the following score: 2 if the log2 ratio ≥ 1, 1 if the log2 ratio < 1 and ≥ 0.25, 0 if the log2 ratio < 0.25 and ≥ –0.25, –1 if the log2 ratio < –0.25 and ≥ –1, and –2 if the log2 ratio < –1. The absolute values of the score from each focal and chromosome/arm event were summed into focal and chromosome/arm scores, respectively. For amplification and deletion scores, only amplification and deletion events were considered, respectively.

Immune status analysis

Estimation of leukocyte fraction was generated using a mixture model of DNA methylation in pure leukocytes versus normal tissue. More details and all calls can be found elsewhere (42). Cytolytic activity metric was obtained by calculating the geometric mean of *GZMA* and *PRF1* expression.

Single-sample Gene Set Enrichment analysis

RNA-seq non-transformed gene expression values were input into single-sample gene set enrichment analysis (ssGSEA2.0, https://github.com/broadinstitute/ssGSEA2.0) (43) with default parameters. Fifty hallmark signature gene sets from the Molecular Signature Database (MSigDB v7.0) (44) were evaluated, and the signature enrichment score for each signature was generated for each sample.

Clinical data analysis

For each cancer type, samples were split into PRS value quartile groups. For survival analysis, Kaplan-Meier plots were visualized using survival and survminer R packages. Cox

Proportional Hazards model was used to compare PRS value quartiles in each cancer type. Gleason score (PRAD), histological grade (ESCA, UCEC, and OV), and sidedness (COADREAD) information were obtained from original cancer type-specific studies or TCGA PanCan Atlas studies (21,22,45–47). These clinical features were compared between four PRS value quartiles.

Validation using the Pan-Cancer Analysis of Whole Genomes (PCAWG) PRAD cohorts

For non-TCGA PCAWG cohorts, files of consensus somatic variant calls, germline variant calls, copy number calls, donor clinical information, and ancestry calls were obtained from ICGC Data Portal (https://dcc.icgc.org/). The consensus germline single-nucleotide variant call was used for PRS calculation. To identify germline pathogenic variants, we selected coding variants of 152 cancer predisposition genes compiled by TCGA (25). We then filtered for rare variants with ≤ 0.05% allele frequency in 1000 Genomes and Exome Aggregation Consortium (release r0.3.1). Among them, 'pathogenic' or 'likely pathogenic' variants annotated by Cancer Predisposition Sequencing Reporter (48) were considered as germline pathogenic variants. We excluded individuals of non-European ancestry from our analysis. Germline–somatic associations were evaluated in the same manner as TCGA analysis, except that subtypes were not considered in generalized linear model due to lack of subtype information for non-TCGA samples.

Supplementary Notes

Supplementary Note 1: Acknowledgments.

S.N. was supported by Takeda Science Foundation. Y.S. was supported by JSPS KAKENHI (22K20808). S.M., D.C.W. and P.G. are supported by Australian National Health and Medical Research Council Fellowships/Investigator grants. Y.O. was supported by JSPS KAKENHI (22H00476), and AMED (JP21gm4010006, JP22km0405211, JP22ek0410075, JP22km0405217, JP22ek0109594), JST Moonshot R&D (JPMJMS2021, JPMJMS2024), Takeda Science Foundation, and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University. K.K. was supported by AMED (JP21cm0106575, JP22ama221510), JST Moonshot R&D (JPMJMS2022), the Uehara Memorial Foundation, and Keio University Academic Development Funds. The supercomputing resources were provided by the Human Genome Center, the Institute of Medical Science, the University of Tokyo. The authors would like to acknowledge the following organizations and groups for data sharing.

TCGA

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Head and neck squamous cell carcinoma

We acknowledge Drs. Valérie Gaborieau and Paul Brennan for the GWAS summary statistics of head and neck squamous cell carcinoma.

Glioblastoma

The GICC was supported by grants from the US National Institutes of Health (NIH)

(R01CA139020 (M.L.B. and B.S.M.), R01CA52689 (M.R.W.), R01CA52689 (M.L.B.) and P30CA125123 (M. Scheurer). Additional support was provided by the McNair Medical Institute (M. Scheurer) and the Population Sciences Biorepository at Baylor College of Medicine (M. Scheurer). Work was additionally supported by Acta Oncologica through the Royal Swedish Academy of Science (B.S.M.'s salary) and by the Swedish Research Council (B.S.M.) and the Swedish Cancer Foundation (B.S.M.). We are grateful to the National Clinical Brain Tumor Group and to all of the clinicians and research nurses throughout Sweden who identified all of the cases.

Esophageal cancer

We acknowledge Drs. Paul Pharoah, Carlos Caldas, and Rebecca C Fitzgerald for the GWAS summary statistics of esophageal cancer and Barrett’s esophagus. The Barrett's and Esophageal Adenocarcinoma Genetic Susceptibility Study (BEAGESS) was funded by grant R01 CA136725 from the National Cancer Institute. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

Bonn (Germany): this work was supported by funding from the Else Kröner Fresenius Stiftung (EKFS) (grant number 2013_A118 awarded to I.G. and J.S.). M.M.N. is a member of the DFG funded Excellence Cluster ImmunoSensation. The Heinz Nixdorf Recall cohort was established with the generous support of the Heinz Nixdorf Foundation, Germany.

Cambridge (UK): the UK Barrett’s oesophagus gene study was funded by a Medical Research Council Programme grant. The UK SOCS study was funded by CRUK as well as funding from the Cambridge NIHR biomedical research centre and the Cambridge Experimental Cancer Medicine Centre. Genotyping of Cambridge samples was supported by funding from the US National Cancer Institute at the National Institutes of Health (grant number R01CA136725 awarded to T.L.V and D.C.W). This study made use of data generated

by the Wellcome Trust Case Control Consortium: Funding for the project was provided by the Wellcome Trust under award 076113; a full list of the investigators who contributed to the generation of the data is available from the website (http://www.wtccc.org.uk/).

Oxford (UK): this work was supported by the Esophageal Adenocarcinoma GenE Consortia incorporating the ChOPIN project (grant C548/A5675), the Inherited Predisposition of neoplasia analysis of genomic DNA (IPOD) from AspECT and BOSS clinical trials project (grant MGAG1G7R), Cancer Research UK (AspECT, grants C548/A4584 and D9612L00090), the Histological AssessmeNt Determining EpitheliaL Response (HANDEL) (grant C548/A9085), the AstraZeneca UK educational grant, the University Hospitals of Leicester R and D grant, and AspECT (T91 5211 University of Oxford grant HDRMJQ0).

A subset of the controls used with the Barrett's and Oesophageal Adenocarcinoma data were obtained from dbGaP. The MD Anderson controls were drawn from study accession phs000187.v1.p1. Genotyping of these controls were done through the University of Texas MD Anderson Cancer Center (UTMDACC) and the Johns Hopkins University Center for Inherited Disease Research (CIDR). We acknowledge the principal investigators of this study: Christopher Amos, Qingyi Wei, and Jeffrey E Lee. Controls from the Genome-Wide Association Study of Parkinson Disease were obtained from dbGaP (study accession: phs000196.v2.p1). This work, in part, used data from the National Institute of Neurological Disorders and Stroke (NINDS) dbGaP database from the CIDR: NeuroGenetics Research Consortium Parkinson's disease study. We acknowledge the principal investigators and coinvestigators of this study: Haydeh Payami, John Nutt, Cyrus Zabetian, Stewart Factor, Eric Molho, and Donald Higgins. Controls from the Chronic Renal Insufficiency Cohort (CRIC) were drawn from dbGaP (study accession: phs000524.v1.p1). The CRIC study was done by the CRIC investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Data and samples from CRIC reported here were supplied by NIDDK Central Repositories. This report was not prepared in collaboration with investigators

of the CRIC study and does not necessarily reflect the opinions or views of the CRIC study, the NIDDK Central Repositories, or the NIDDK. We acknowledge the principal investigators and the project officer of this study: Harold I Feldman, Raymond R Townsend, Lawrence J Appel, Mahboob Rahman, Akinlolu Ojo, James P Lash, Jiang He, Alan S Go, and John W Kusek.

Lung cancer

We acknowledge the Transdisciplinary Research In Cancer of the Lung (TRICL) Research Team, which is a part of the Genetic Associations and MEchanisms in ONcology (GAME-ON) consortium, and associated with the International Lung Cancer Consortium (ILCCO), for the GWAS summary statistics of lung cancer.

Prostate cancer

We acknowledge The PRACTICAL consortium, CRUK, BPC3, CAPS, and PEGASUS for the GWAS summary statistics of prostate cancer.

The Prostate cancer genome-wide association analyses are supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative).

We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now PCUK), The Orchid Cancer Appeal, Rosetrees Trust,

The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust.

The Prostate Cancer Program of Cancer Council Victoria also acknowledges grant support from The National Health and Medical Research Council, Australia (126402, 209057, 251533, 396414, 450104, 504700, 504702, 504715, 623204, 940394, 614296,), VicHealth, Cancer Council Victoria, The Prostate Cancer Foundation of Australia, The Whitten Foundation, PricewaterhouseCoopers, and Tattersall's. EAO, DMK, and EMK acknowledge the Intramural Program of the National Human Genome Research Institute for their support.

Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I] and by Cancer Research UK grant A8197/A16565. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher).

Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund.

The BPC3 was supported by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 to D.J.H., U01-CA98710 to S.M.G., U01-CA98216 to E.R., and U01-CA98758 to B.E.H., and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics).

CAPS GWAS study was supported by the Swedish Cancer Foundation (grant no 09-0677, 11-484, 12-823), the Cancer Risk Prediction Center (CRisP; www.crispcenter.org), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council, Swedish Research Council (grant no K2010-70X-20430-04-3, 2014-2269)

PEGASUS was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.

Breast cancer

The breast cancer genome-wide association analyses for BCAC and CIMBA were supported by Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), The National Institutes of Health (CA128978, X01HG007492- the DRIVE consortium), the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, the Quebec Breast Cancer Foundation, the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), the European Union's Horizon 2020 Research and Innovation Programme (634935 and 633784), the Post-Cancer GWAS initiative (U19 CA148537, CA148065 and

CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. All studies and funders are listed in Zhang *et al.* (Nat Genet, 2020).

Colorectal cancer

The colorectal cancer GWAS meta-analysis was supported by the following:

CPS-II: Grants from the American Cancer Society (to P.T. Campbell and S.M. Gapstur).

DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, and CH 117/1-1), and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814).

HPFS: National Institutes of Health (P01 CA055075, UM1 CA167552, R01 CA137178, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003).

NHS: National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003).

PLCO: National Institutes of Health, to 10 PLCO screening centers, a coordinating center, a data management center and sample processing laboratories: University of Colorado Denver (N01-CN-25514); Georgetown University (N01-CN-25522); Pacific Health Research Institute (N01-CN-25515); Henry Ford Health System (N01-CN-25512); University of Minnesota (N01-CN-25513); Washington University (N01-CN-25516); University of Pittsburgh (N01-CN-25511); University of Utah (N01-CN-25524); Marshfield Clinic Research Foundation (N01-CN-25518); University of Alabama at Birmingham (N01-CN-75022); Westat CC (N01-CN-25476); IMS, Inc. (HHSN261201300008I); and Leidos, Inc. (HHSN261200800001E).

WHI: National Institutes of Health (contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C). Please refer

to the Women's Health Initiative Clinical Trial and Observational Study page on dbGaP (phs000200).

Supplementary Note 2: Contributing authors participating in the Glioma International Case Control Study, the UCSF Adult Glioma Study, and the GliomaScan consortia.

The Glioma International Case Control Study comprised: Elizabeth B. Claus (School of Public Health, Yale University, New Haven, CT 06510, United States and Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA 02115, United States); Dora Il'yasova (Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, GA 30303, United States; Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, United States and Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, NC 27710, United States); Joellen Schildkraut (Department of Public Health Sciences, School of Medicine, University of Virginia, Charlottesville, VA 22903, United States); Jill S. Barnholtz-Sloan (Department of Population and Quantitative Health Sciences and the Cleveland Center for Health Outcomes Research, Case Western Reserve University School of Medicine, Cleveland, OH 44106, United States); Sara H. Olson (Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10017, United States); Jonine L. Bernstein (Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10017, United States); Christoffer Johansen (Danish Cancer Society Research Center, Survivorship, Danish Cancer Society, Copenhagen 2100, Denmark; 15Oncology Clinic, Finsen Centre, Rigshospitalet, University of Copenhagen, Copenhagen 2100, Denmark); Robert B. Jenkins (Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, MN 55905, United States); Beatrice S. Melin (Department of Radiation Sciences, Umeå University, Umeå 901 87, Sweden); Margaret R. Wrensch (Department of Neurological Surgery, School of Medicine, University of California, San Francisco, CA 94143, United States); Richard S. Houlston (Division of Molecular Pathology, The Institute of Cancer Research, London SW7 3RP, United Kingdom); Melissa L. Bondy (Department of

Epidemiology and Population Health, Stanford Cancer Institute, Stanford University, Stanford, CA 94305, United States).

The UCSF Adult Glioma Study comprised: M.R. Wrensch, T. Rice, J.K. Wiencke, L.S. McCoy, H.M. Hansen, M. Berger, P. Bracci, S. Chang, J. Clarke, A. Molinaro, A. Perry, M. Pezmecki, M. Prados, I. Smirnov, T. Tihan, K. Walsh, J. Wiemels and S. Zheng. The UCSF Adult Glioma Study was supported by the NIH (grant numbers R01CA52689 (M.R.W. and J.K.W.), P50CA097257 (M.R.W. and J.K.W.), R01CA126831 (J.K.W.) and R01CA139020 (M.R.W.)), the Loglio Collective (M.R.W. and J.K.W.), the National Brain Tumor Foundation (M.R.W.), the Stanley D. Lewis and Virginia S. Lewis Endowed Chair in Brain Tumor Research (M.R.W.), the Robert Magnin Newman Endowed Chair in Neuro-oncology (J.K.W.) and by donations from the families and friends of J. Berardi, H. Glaser, E. Olsen, R.E. Cooper and W. Martinusen. This project also was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through UCSF–CTSI grant UL1 RR024131 (UCSF CTSI). The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code section 103885, the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C (awarded to the Cancer Prevention Institute of California), contract HHSN261201000035C (awarded to the University of Southern California) and contract HHSN261201000034C (awarded to the Public Health Institute), and the Centers for Disease Control and Prevention's National Program of Cancer Registries under agreement # U58DP003862-01 (awarded to the California Department of Public Health). The ideas and opinions expressed herein are those of the author(s), and endorsement by the State of California Department of Public Health, the

National Cancer Institute and the Centers for Disease Control and Prevention, or their contractors and subcontractors, is not intended nor should be inferred.

The GliomaScan Consortium comprised: U. Andersson (Department of Radiation Sciences, Umeå University, Umeå, Sweden), P. Rajaraman (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA), S.J. Chanock (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA), M.S. Linet (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA), Z. Wang (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA), M. Yeager (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA), L.E.B. Freeman (Division of Cancer Epidemiology and Genetics, National Cancer Institute), S. Koutros (Division of Cancer Epidemiology and Genetics, National Cancer Institute), D. Albanes (Division of Cancer Epidemiology and Genetics, National Cancer Institute), K. Visvanathan (Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA and Johns Hopkins Sidney Kimmel Comprehensive Cancer Center), V.L. Stevens (Epidemiology Research Program, American Cancer Society), R. Henriksson (Department of Radiation Sciences, Oncology, Umea University), D.S. Michaud (Department of Public Health and Community Medicine, Tufts University Medical School), M. Feychting (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet), A. Ahlbom (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet), G.G. Giles (Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Victoria, Australia and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne), R. Milne (Cancer Epidemiology Centre, Cancer Council of Victoria and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne), R. McKean-Cowdin (Department of Preventive Medicine, Keck School of Medicine, University of

Southern California, Los Angeles, California, USA), L. Le Marchand (Cancer Research Center, University of Hawaii, Honolulu, Hawaii, USA), M. Stampfer (Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA and Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA), A.M. Ruder (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), T. Carreon (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), G. Hallmans (Department of Public Health and Clinical Medicine/Nutritional Research, Umea University, Umea, Sweden), A. Zeleniuch-Jacquotte (Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA), J.M. Gaziano (Massachusetts Veteran's Epidemiology, Research and Information Center, Geriatric Research Education and Clinical Center, VA Boston Healthcare System, Boston, Massachusetts, USA), H.D. Sesso (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA), M.P. Purdue (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), E. White (Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington, Seattle, Washington, USA) and J. Buring (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA).

Supplementary References

1. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **2011**;12:246.
2. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet* **2020**;52:572–81.
3. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **2018**;50:928–36.
4. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* **2019**;51:76–87.
5. Landi MT, Bishop DT, MacGregor S, Machiela MJ, Stratigos AJ, Ghiorzo P, *et al.* Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nat Genet* **2020**;52:494–504.
6. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **2017**;49:1126–32.
7. O'Mara TA, Glubb DM, Amant F, Annibali D, Ashton K, Attia J, *et al.* Identification of nine new susceptibility loci for endometrial cancer. *Nat Commun* **2018**;9:3166.
8. Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, Winham SJ, *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **2017**;49:680–91.
9. Gharahkhani P, Fitzgerald RC, Vaughan TL, Palles C, Gockel I, Tomlinson I, *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol* **2016**;17:1363–73.
10. Ostrom QT, Kinnersley B, Armstrong G, Rice T, Chen Y, Wiencke JK, *et al.* Age-specific genome-wide association study in glioblastoma identifies increased proportion of 'lower grade glioma'-like features associated with younger age. *Int J Cancer* **2018**;143:2359–66.
11. Lesseur C, Diergaarde B, Olshan AF, Wünsch-Filho V, Ness AR, Liu G, *et al.* Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* **2016**;48:1544–50.
12. Leo PJ, Madeleine MM, Wang S, Schwartz SM, Newell F, Pettersson-Kymmer U, *et al.* Defining the genetic susceptibility to cervical neoplasia—A genome-wide

association study. *PLoS Genet* **2017**;13:1–20.
13. Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, *et al.* Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* **2017**;49:789–94.
14. Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **2015**;47:291–5.
15. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **2015**;47:1114.
16. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018**;562:203–9.
17. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **2017**;49:1421–7.
18. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput Biol* **2015**;11:e1004219.
19. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, *et al.* A global reference for human genetic variation. *Nature* **2015**;526:68–74.
20. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **2018**;173:321-337.e10.
21. Liu Y, Sethi NS, Thorsson S, Bass AJ, Laird PW. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* **2018**;33:721–35.
22. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **2015**;163:1011–25.
23. Cuzzubbo S, Fujimoto J, Worrell R, Carter C, Zuna R, Wong CK, *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **2018**;33:690-705.e9.
24. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**;455:1061–8.
25. Huang K-L, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **2018**;173:355-370.e14.
26. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **2008**;40:1253.
27. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust

relationship inference in genome-wide association studies. *Bioinformatics* **2010**;26:2867–73.
28. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **2019**;47:D1005–12.
29. Deelen P, Bonder M, van der Velde K, Westra H-J, Winder E, Hendriksen D, *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res Notes* **2014**;7:901.
30. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **2019**;10:5436.
31. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* **2015**;31:782–4.
32. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna K V., *et al.* Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am J Hum Genet* **2008**;83:132.
33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **2006**;38:904–9.
34. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLOS Genet* **2006**;2:e190.
35. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **2018**;3:861.
36. Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, *et al.* Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nat Commun* **2020**;11:1–11.
37. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **2017**;171:1042-1056.e10.
38. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **2014**;30:1015–6.
39. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018**;173:371-385.e18.
40. Chatsirisupachai K, Lesluyes T, Paraoan L, Loo P Van, Magalhães JP de. An integrative analysis of the age-associated multi-omic landscape across cancers. *Nat Commun 2021 121* **2021**;12:1–17.
41. Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, *et al.* Integrated Analysis of

Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **2018**;34:549-560.e9.

42. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, *et al.* The Immune Landscape of Cancer. *Immunity* **2018**;48:812-830.e14.
43. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**;462:108–12.
44. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **2015**;1:417–25.
45. Levine DA. Integrated genomic characterization of endometrial carcinoma. *Nature* **2013**;497:67–73.
46. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **2011**;474:609–15.
47. Kim J, Bowlby R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **2017**;541:169–75.
48. Nakken S, Saveliev V, Hofmann O, Møller P, Myklebost O, Hovig E. Cancer Predisposition Sequencing Reporter (CPSR): A flexible variant report engine for high-throughput germline screening in cancer. *Int J cancer* **2021**;149:1955–60.