









Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts

Sara R. Rashkin ^{1,8}, Rebecca E. Graff ^{1,2,8}, Linda Kachuri¹, Khanh K. Thai², Stacey E. Alexeeff², Maruta A. Blatchins², Taylor B. Cavazos ^{1,3}, Douglas A. Corley², Nima C. Emami^{1,3}, Joshua D. Hoffman¹, Eric Jorgenson ², Lawrence H. Kushi ², Travis J. Meyers¹, Stephen K. Van Den Eeden ^{2,4}, Elad Ziv^{5,6,7}, Laurel A. Habel², Thomas J. Hoffmann ^{1,2,5}, Lori C. Sakoda ^{2,9}✉ & John S. Witte^{1,4,5,7,9}✉

Deciphering the shared genetic basis of distinct cancers has the potential to elucidate carcinogenic mechanisms and inform broadly applicable risk assessment efforts. Here, we undertake genome-wide association studies (GWAS) and comprehensive evaluations of heritability and pleiotropy across 18 cancer types in two large, population-based cohorts: the UK Biobank (408,786 European ancestry individuals; 48,961 cancer cases) and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging cohorts (66,526 European ancestry individuals; 16,001 cancer cases). The GWAS detect 21 genome-wide significant associations independent of previously reported results. Investigations of pleiotropy identify 12 cancer pairs exhibiting either positive or negative genetic correlations; 25 pleiotropic loci; and 100 independent pleiotropic variants, many of which are regulatory elements and/or influence cross-tissue gene expression. Our findings demonstrate widespread pleiotropy and offer further insight into the complex genetic architecture of cross-cancer susceptibility.

¹ Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. ² Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. ³ Program in Biological and Medical Informatics, University of California, San Francisco, San Francisco, CA, USA. ⁴ Department of Urology, University of California, San Francisco, San Francisco, CA, USA. ⁵ Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. ⁶ Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. ⁷ Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA. ⁸ These authors contributed equally: Sara R. Rashkin, Rebecca E. Graff. ⁹ These authors jointly supervised this work: Lori C. Sakoda, John S. Witte. ✉email: lori.sakoda@kp.org; jwitte@ucsf.edu

The global burden of cancer is substantial, with an estimated 18.1 million individuals diagnosed each year and approximately 9.6 million deaths attributed to the disease¹. Efforts toward cancer prevention, screening, and treatment are thus imperative, but they require a more comprehensive understanding of the underpinnings of carcinogenesis than we currently possess. While studies of twins², families³, and unrelated populations^{4–6} have demonstrated substantial heritability and familial clustering for many cancers, the extent to which genetic variation is unique versus shared across different types of cancer remains unclear.

Genome-wide association studies (GWAS) of individual cancers have identified loci associated with multiple cancer types, including 1q32 (*MDM4*)^{7,8}; 2q33 (*CASP8-ALS2CR12*)^{9,10}; 3q28 (*TP63*)^{11,12}; 4q24 (*TET2*)^{13,14}; 5p15 (*TERT-CLPTMIL*)^{9,12}; 6p21 (HLA complex)^{15,16}; 7p15¹⁷; 8q24^{12,18}; 11q13^{18,19}; 17q12 (*HNF1B*)^{18,20}; and 19q13 (*MERIT40*)²¹. In addition, recent studies have tested single-nucleotide polymorphisms (SNPs) previously associated with one cancer to discover pleiotropic associations with other cancer types^{22–25}. Consortia, such as the Genetic Associations and Mechanisms in Oncology, have looked for variants and pathways shared by breast, colorectal, lung, ovarian, and prostate cancers^{26–30}. Comparable studies for other cancers—including those that are less common—have yet to be reported.

In addition to individual variants, recent studies have evaluated genome-wide genetic correlations between pairs of cancer types^{4–6}. One evaluated 13 cancer types and found shared heritability between kidney and testicular cancers, diffuse large B-cell lymphoma (DLBCL) and osteosarcoma, DLBCL and chronic lymphocytic leukemia (CLL), and bladder and lung cancers⁴. Another study of six cancer types found correlations between colorectal cancer and both lung and pancreatic cancers⁵. In an updated analysis with increased sample size, the same group identified correlations of breast cancer with colorectal, lung, and ovarian cancers and of lung cancer with colorectal and head/neck cancers⁶. While these studies provide compelling evidence for shared heritability across cancers, they lack data on several cancer types (e.g., cervix, melanoma, and thyroid).

Here, we present analyses of genome-wide SNP data on 18 cancer types, examining 408,786 individuals of European ancestry from two large, independent, and contemporary cohorts unselected for phenotype—the UK Biobank (UKB) and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging (GERA) cohorts. We seek to detect risk SNPs and pleiotropic loci and variants and to estimate the heritability of and genetic correlations between cancer types. We then conduct in silico functional analyses of pleiotropic variants to catalog biological mechanisms potentially shared across cancers. Leveraging the wealth of individual-level genetic and phenotypic data from both cohorts allows us to extensively interrogate the shared genetic basis of susceptibility to different cancer types, with the ultimate goal of better understanding common genetic mechanisms of carcinogenesis and improving risk assessment. We find widespread pleiotropy that offers further insights into the complex genetic architecture of cross-cancer susceptibility.

Results

Genome-wide association analyses of individual cancers. We found 21 previously unreported genome-wide significant associations between variants and cancers at $P < 5 \times 10^{-8}$ upon meta-analysis of the UKB and GERA results (Table 1). These included 20 unique variants, with 1 variant that was associated with two cancers (rs78378222). Nine of these 21 associations were in known susceptibility regions for the cancer of interest but

independent of previously reported variants ($r^2 < 0.1$; see “Methods”). The remaining 12 were in regions not previously associated with the cancer of interest in individuals of European ancestry. Fourteen of these 21 associations indicated pleiotropy in that the relevant variants were in regions previously associated with at least one of the other cancer types evaluated in this study (Table 1). The effect estimates for these 21 associations were not materially changed when stratified by age at diagnosis, Surveillance, Epidemiology, and End Results Program (SEER) grade, or SEER stage (heterogeneity $P > 0.05$ /[number of strata and variants]; see “Methods”).

In addition, there were nine previously unreported variants associated with cancers at $P < 5 \times 10^{-8}$ that were only genotyped or imputed in one cohort (Supplementary Data 1; yellow rows). For the sake of completeness and future efforts, Supplementary Data 1 also includes the 21 associations from Table 1 (green rows) and an additional 113 suggestive associations ($P < 1 \times 10^{-6}$) independent of previously reported results. Finally, we replicated 308 independent cancer risk variants identified as GWAS significant by previous studies (Supplementary Data 2; $P < 1 \times 10^{-6}$).

In genome-wide sensitivity analyses in the UKB cohort restricted to incident cases (i.e., excluding prevalent cases), our findings for significant and suggestive associations were essentially unchanged (heterogeneity $P > 0.05$ /[number of variants per cancer]; see “Methods”; Supplementary Fig. 1). Similarly, genome-wide sensitivity analysis results in the UKB cohort for esophageal and stomach cancers separately were comparable to those for the two phenotypes combined (heterogeneity $P > 0.05$ /6; see “Methods”; Supplementary Fig. 2).

Genome-wide heritability and genetic correlation. Array-based heritability estimates across cancers ranged from $h^2 = 0.04$ (95% CI: 0.00–0.13) for oral cavity/pharyngeal cancer to $h^2 = 0.26$ (95% CI: 0.15–0.38) for testicular cancer (Table 2). For some of the cancers, our array-based heritability estimates were comparable to twin- or family-based heritability estimates^{2,3} but were more precise. Several were also similar to array-based heritability estimates from consortia comprised of multiple studies^{4–6}. One of our highest heritability estimates was observed for thyroid cancer ($h^2 = 0.21$; 95% CI: 0.09–0.33), a cancer that has not been evaluated in other array-based studies.

Among pairs of cancers, only colon and rectal cancers ($r_g = 0.85$, $P = 5.33 \times 10^{-7}$) were genetically correlated at a Bonferroni-corrected significance threshold of $P = 0.05/153 = 3.27 \times 10^{-4}$ or using a false discovery rate (FDR) threshold of $q < 0.1$ (Fig. 1a, Table 3 and Supplementary Data 3). However, at a nominal threshold of $P = 0.05$, we observed suggestive relationships between 11 other pairs. Seven pairs showed positive correlations: esophageal/stomach cancer was correlated with Non-Hodgkin’s lymphoma (NHL; $r_g = 0.40$, $P = 0.0089$), breast ($r_g = 0.26$, $P = 0.0069$), lung ($r_g = 0.44$, $P = 0.0035$), and rectal ($r_g = 0.32$, $P = 0.024$) cancers; bladder and breast cancers ($r_g = 0.22$, $P = 0.017$); melanoma and testicular cancer ($r_g = 0.23$, $P = 0.028$); and prostate and thyroid cancers ($r_g = 0.23$, $P = 0.013$). The remaining four pairs showed negative correlations: endometrial and testicular cancers ($r_g = -0.41$, $P = 0.0064$); esophageal/stomach cancer and melanoma ($r_g = -0.27$, $P = 0.038$); lung cancer and melanoma ($r_g = -0.28$, $P = 0.0048$); and NHL and prostate cancer ($r_g = -0.21$, $P = 0.012$).

Locus-specific pleiotropy. We detected 25 pleiotropic regions associated with more than one cancer ($P < 5 \times 10^{-8}$ for each cancer; independent regions were defined using our linkage disequilibrium [LD] clumping procedure; see “Methods”; Fig. 1b and

Table 1 Previously unreported genome-wide significant loci from meta-analysis of UKB and GERA SNPs for each cancer site.

Cancer site	SNP	Chromosome	Position	Gene	REF/ ALT ^a	MAF UKB	MAF GERA	OR UKB	OR GERA	OR Meta	Meta P
Bladder	rs76088467 ^{b,c}	6	21795787	CASC15	G/A	0.025	0.030	0.67	0.59	0.64	2.34 × 10 ⁻⁸
Breast	rs6752414 ^{b,c}	2	121425339	Intergenic	T/C	0.077	0.083	0.89	0.87	0.88	1.81 × 10 ⁻⁹
Breast	rs8027730 ^{c,d}	15	49872585	FAM227B	A/C	0.48	0.48	1.06	1.08	1.06	2.68 × 10 ⁻⁸
Cervix	rs10175462 ^{c,d}	2	113988492	PAX8	A/G	0.36	0.37	1.16	1.08	1.15	7.71 × 10 ⁻¹⁴
Cervix	rs2856437 ^{b,c}	6	32157364	PBX2	A/G	0.063	0.047	0.76	0.88	0.77	1.24 × 10 ⁻¹⁵
Colon	rs7151887 ^{c,d}	8	103561978	Upstream of ODF1	G/C	0.015	0.017	0.65	0.61	0.64	1.27 × 10 ⁻⁸
Colon	rs8114643 ^d	20	7833046	Intergenic	G/A	0.14	0.15	0.83	0.84	0.83	2.10 × 10 ⁻⁹
Esophagus/ Stomach	rs75960256 ^{b,c}	2	106687838	C2orf40	G/A	0.024	0.022	0.52	0.67	0.53	1.04 × 10 ⁻⁸
Kidney	rs112248293 ^{c,d}	15	61500352	RORA	A/G	0.024	0.025	0.53	0.62	0.55	3.36 × 10 ⁻⁹
Lung	rs10863899 ^d	1	211666218	5'UTR of RD3	G/A	0.42	0.42	1.23	1.09	1.18	1.91 × 10 ⁻⁸
Lung	rs146099759 ^d	5	12883592	Intergenic	A/G	0.024	0.028	0.69	0.57	0.64	3.50 × 10 ⁻⁸
Lung	rs12543486 ^{b,c}	8	13012376	DLG1	C/T	0.17	0.16	1.30	1.18	1.26	3.51 × 10 ⁻⁸
Lymphocytic Leukemia	rs114490818 ^d	3	126099101	Intergenic	A/G	0.022	0.011	0.48	0.53	0.48	2.86 × 10 ⁻⁸
Lymphocytic Leukemia	rs61965473 ^{c,d}	13	95571786	Intergenic	T/C	0.023	0.023	0.52	0.44	0.49	3.95 × 10 ⁻⁸
Lymphocytic Leukemia	rs78878222 ^{b,c}	17	7571752	3'UTR of TP53	G/T	0.012	0.014	0.44	0.34	0.40	1.89 × 10 ⁻⁹
Melanoma	rs9818780 ^{c,d}	3	156492758	Intergenic	C/T	0.49	0.48	0.92	0.89	0.91	3.16 × 10 ⁻⁸
Melanoma	rs12186662 ^d	5	90356197	ADGRV1	G/A	0.32	0.36	0.90	0.89	0.90	1.09 × 10 ⁻⁸
Melanoma	rs55797833 ^{b,c}	9	21995044	5'UTR OF CDKN2A	G/T	0.023	0.021	1.71	1.72	1.71	6.71 × 10 ⁻¹²
Melanoma	rs78378222 ^{b,c}	17	7571752	3'UTR of TP53	G/T	0.012	0.014	0.70	0.63	0.67	1.18 × 10 ⁻⁸
Rectum	rs145503185 ^d	9	23455764	Intergenic	C/T	0.013	0.018	0.57	0.50	0.55	4.36 × 10 ⁻⁸
Thyroid	2:173859846_TA_Td	2	173859846	RAPGEF4	T/TA	0.25	0.26	1.45	1.15	1.36	3.49 × 10 ⁻⁸

MAF minor allele frequency calculated in all controls, OR odds ratio.

^a REF is reference allele and ALT allele is effect allele; bold allele is minor allele

^b Indicates SNPs in known susceptibility loci for cancer of interest in European ancestry but independent of previously reported variants (LD $r^2 < 0.1$ in Europeans).

^c Indicates SNPs in loci previously associated with at least one of the other cancers evaluated in this study in European ancestry.

^d Indicates SNPs in loci not previously associated with the cancer of interest in European ancestry.

Table 2 Heritability estimates (h^2) and 95% confidence intervals (CIs) for each cancer based on the union set of UKB and GERA SNPs and previous estimates.

Cancer site	Current study (array based)	Jiang et al. ^a (array based)	Sampson et al. ^b (array based)	Mucci et al. ^c (twin/family based)
Bladder	0.08 (0.04–0.12)		0.12 (0.09–0.16)	0.07 (0.02–0.11) ^d
Breast	0.10 (0.08–0.13)	0.14 (0.12–0.16)	0.10 (0.00–0.20) ^e	0.31 (0.11–0.51)
Cervix	0.07 (0.02–0.12)			0.13 (0.06–0.15) ^{d,f}
Colon	0.07 (0.04–0.10)	0.09 (0.07–0.11) ^g		0.15 (0.00–0.45)
Endometrium	0.13 (0.07–0.18)		0.18 (0.09–0.27)	0.27 (0.11–0.43)
Esophagus/stomach	0.14 (0.07–0.21)		0.38 (0.17–0.59) ^h	0.22 (0.00–0.55) ⁱ
Kidney	0.09 (0.04–0.15)		0.15 (0.02–0.27)	0.38 (0.21–0.55)
Lung	0.15 (0.10–0.20)	0.08 (0.05–0.10)	0.21 (0.14–0.27)	0.18 (0.00–0.42)
Lymphocytic leukemia	0.14 (0.05–0.23)		0.22 (0.16–0.28) ^j	0.09 (0.09–0.16) ^{d,k}
Melanoma	0.08 (0.04–0.11)			0.58 (0.43–0.73)
Non-Hodgkin's lymphoma	0.13 (0.03–0.23)		0.09 (0.04–0.15) ^l	0.10 (0.08–0.10) ^d
Oral cavity/pharynx	0.04 (0.00–0.13)	0.10 (0.05–0.14)		0.09 (0.00–0.60)
Ovary	0.07 (0.01–0.13)	0.03 (0.02–0.05)		0.39 (0.23–0.55)
Pancreas	0.06 (0.00–0.18)	0.05 (0.00–0.10) ^m	0.10 (0.04–0.16)	
Prostate	0.16 (0.13–0.20)	0.18 (0.14–0.22)	0.38 (0.24–0.51)	0.57 (0.51–0.63)
Rectum	0.11 (0.07–0.16)			0.14 (0.00–0.50)
Testis	0.26 (0.15–0.38)		0.30 (0.08–0.51)	0.25 (0.15–0.37) ^d
Thyroid	0.21 (0.09–0.33)			0.53 (0.52–0.53) ^d

^aTaken from ref. 6, 95% CI calculated from provided standard error.
^bTaken from ref. 4.
^cTaken from ref. 2, except where not included in analysis or 95% CI range was >0.60; remaining taken from ref. 3, as marked.
^dTaken from ref. 3, family-based not twin.
^eEstrogen receptor negative (ER-).
^fFor in situ (invasive: $h^2 = 0.22$ [0.14–0.27]).
^gColorectal.
^hFor esophageal in Asian population (stomach in Asian population: $h^2 = 0.25$ [0.00–0.52]).
ⁱStomach.
^jFor chronic lymphocytic leukemia.
^kAge >15 years.
^lFor diffuse large B cell lymphoma.
^mTaken from ref. 5.

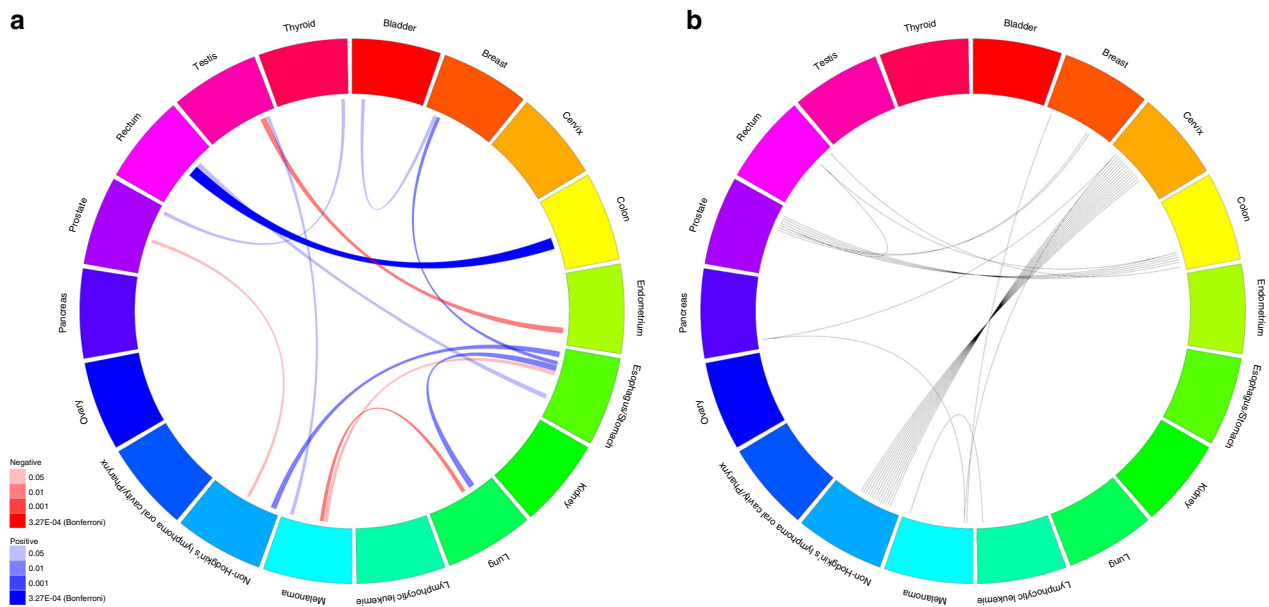


Fig. 1 Cross-cancer genetic correlations (r_g) calculated via LD-score regression (LDSC) and associated cancers from the locus-specific pleiotropy analysis. **a Cancer pairs are connected if the genetic correlation had $P < 0.05$, width of the line is proportional to magnitude of r_g , color of the line indicates direction of correlation (red is negative and blue is positive), and shading is proportional to strength of association according to P , where the Bonferroni-corrected threshold is $0.05/153 = 3.27 \times 10^{-4}$; **b** cancer pairs are connected by a line (each line represents one region) if a region contains any SNPs associated with either cancer, where regions are formed around index SNPs with $P < 5 \times 10^{-8}$ for any cancer in the cancer-specific meta-analyses and SNPs are added if they have $P < 5 \times 10^{-8}$ for any cancer, are within 500 kb of the index SNP, and have LD $r^2 > 0.5$ with the index SNP.**

Table 3 Cross-cancer genetic correlations (r_g) calculated via LD-score regression (LDSC) for all cancer pairs with $P < 0.05$.

Cancer site 1	Cancer site 2	r_g (95% CI)	P
Bladder	Breast	0.22 (0.04–0.41)	0.017
Breast	Esophagus/stomach	0.26 (0.07–0.44)	0.0069
Colon	Rectum	0.85 (0.52–1.00)	5.33×10^{-7}
Endometrium	Testis	−0.41 (−0.70 to −0.11)	0.0064
Esophagus/stomach	Lung	0.44 (0.15–0.74)	0.0035
Esophagus/stomach	Melanoma	−0.27 (−0.53 to −0.01)	0.038
Esophagus/stomach	Non-Hodgkin's lymphoma	0.40 (0.10–0.70)	0.0089
Esophagus/stomach	Rectum	0.32 (0.04–0.60)	0.024
Lung	Melanoma	−0.28 (−0.47 to −0.08)	0.0048
Melanoma	Testis	0.23 (0.03–0.44)	0.028
Non-Hodgkin's lymphoma	Prostate	−0.21 (−0.37 to −0.05)	0.012
Prostate	Thyroid	0.23 (0.05–0.41)	0.013

CI confidence interval.

Supplementary Table 1). Most were at known cancer pleiotropic loci: HLA (14 regions), 8q24 (7 regions), *TERT-CLPTMIL* (2 regions), and *TP53* (1 region). All of the HLA regions were associated with both cervical cancer and NHL. Five regions in 8q24 were associated with prostate and colon cancers (one also associated with rectal cancer), and two were associated with prostate and breast cancers. Of the regions in *TERT-CLPTMIL*, one was associated with breast cancer and melanoma, and the other was associated with melanoma and cervical and pancreatic cancers. The *TP53* region, indexed by rs78378222, was associated with melanoma and lymphocytic leukemia. The remaining pleiotropic region, indexed by rs6507874, was in *SMAD7*, which has been previously linked to colorectal cancer³¹, and we confirmed its association with colon and rectal cancers separately.

Genome-wide variant-specific pleiotropy. We assessed variant-specific pleiotropy by testing all variants genome-wide using the summary statistics for each cancer using ASSET. We found 85 independent (LD $r^2 < 0.1$) one-directional pleiotropic variants with at least two associated cancers, the same direction of effect for all associated cancers, and an overall pleiotropic $P < 5 \times 10^{-8}$ (Supplementary Data 4). Of these one-directional pleiotropic variants, there were 17 for which the overall pleiotropic P was smaller than the P for each of the associated cancers (Fig. 2 and Table 4). While 84 of the 85 one-directional pleiotropic variants were in regions that have previously been associated with any cancer, 68 were associated with at least one cancer not previously reported. The variant in a region not previously associated with any cancer is rs150260898, intronic of *RABIF5*, which was associated with melanoma and oral cavity/pharyngeal cancer.

We also considered bidirectional pleiotropic associations, wherein the same allele for a given variant was associated with an increased risk for some cancers but a decreased risk for others. We found 15 such variants with $P < 5 \times 10^{-8}$, all of which were independent from one another and from the one-directional pleiotropic variants (LD $r^2 < 0.1$; Fig. 3, Table 5 and Supplementary Data 5). There were eight variants where the overall pleiotropic P was smaller than the P for the associated cancers. While all of the bidirectional pleiotropic variants were in regions that have previously been associated with cancer, six were independent of known risk variants, and all 15 were associated with at least one cancer not previously reported.

For any pair of cancers associated with the same variant, the type of association falls in one of three categories: (1) SNPs identified in the one-directional analysis, where all associations are in the same direction; (2) SNPs identified in the bidirectional

analysis, where both cancers in the pair are associated in the same direction (both risk increasing or both risk decreasing), even though at least one other cancer is associated in the opposite direction; and (3) SNPs identified in the bidirectional analysis, where the pair of cancers are associated in opposite directions (one risk increasing and one risk decreasing). For each of the possible 153 pairs of cancers, we tabulated how many of the 100 pleiotropic SNPs fall into each category (Fig. 4a and Supplementary Data 6). The number of one- and bidirectional SNPs shared by cancer pairs ranged from one (bladder and breast) to 13 (lymphocytic leukemia and testis) (Fig. 4a and Supplementary Data 6). For 30 cancer pairs, the shared associations had exclusively the same direction of effect (i.e., tabulating across the first two categories of pleiotropic SNPs). For three cancer pairs, at least 50% of the shared variants were associated in opposite directions.

For each of the 100 independent SNPs showing either one- or bidirectional pleiotropy (Supplementary Data 4–5), we assessed whether the results differed according to age at diagnosis, SEER grade, or SEER stage for any of the associated cancers. After correcting for the number of SNPs and strata tested, only a single one-directional pleiotropic SNP showed heterogeneity across case subtypes. rs111362352-C was significantly positively associated with the risk of low grade prostate cancer in GERA, while it was not associated with high-grade disease. These results are consistent with previous findings for this SNP (or SNPs in strong LD): the C allele has been associated with lower Gleason score, and it is located at *KLK3*, the prostate-specific antigen gene, which may reflect its previous association with lower grade prostate cancer^{32,33}.

Functional characterization of pleiotropic variants. The biological significance of these 100 independent pleiotropic variants (Supplementary Data 4–5) was evaluated using in silico annotation tools (Supplementary Data 7)^{34–36}. Pleiotropic variants were enriched in intergenic ($P = 0.043$) and non-coding RNA transcripts ($P = 0.015$) compared to all variants in the reference panel of UKB European descent individuals (Fig. 4b). The distribution of DeepSea functional significance scores was skewed toward 0 ($P = 7.3 \times 10^{-4}$), indicating a higher likelihood of regulatory effects compared to a reference distribution of 1000 Genomes variants (Fig. 4d). Suggestively functional variants ($n = 26$, DeepSEA score < 0.05) were also predicted to be pathogenic by Combined Annotation-Dependent Depletion³⁶ (CADD; mean score of 10.66, corresponding to the top 10% of deleterious substitutions). Twenty-two of the 100 pleiotropic variants were characterized by active chromatin states, 33 were classified as enhancers, and 64 had

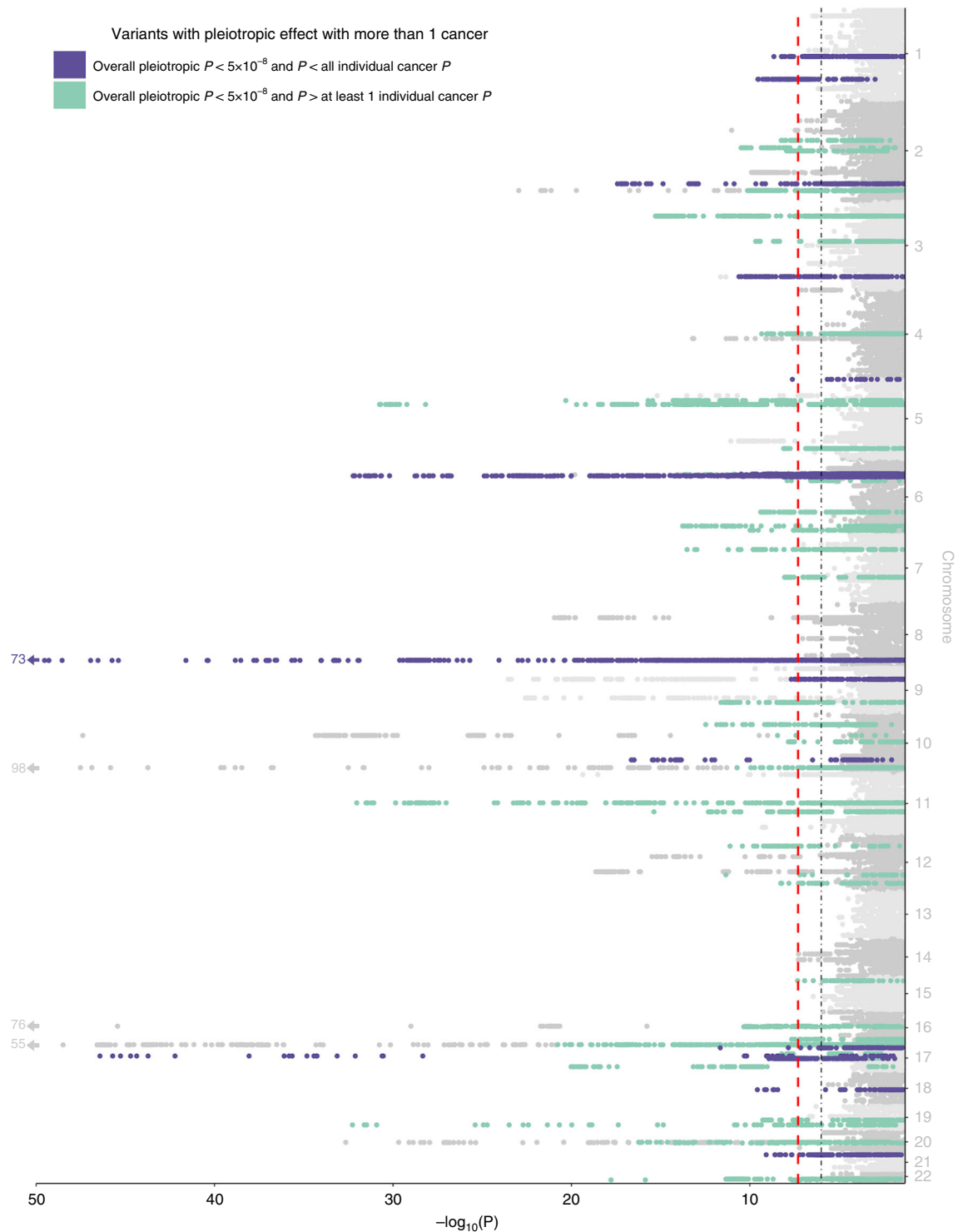


Fig. 2 Manhattan plot displaying one-directional variant-specific pleiotropy from ASSET. The red dashed line represents the genome-wide significance threshold ($P < 5 \times 10^{-8}$), and the black dotted line represents a suggestive threshold ($P < 1 \times 10^{-6}$). Highlighted in purple are genome-wide significant loci where the overall pleiotropic P is less than all individual P for the selected cancers. Highlighted in green are the genome-wide significant loci where the overall pleiotropic P is greater than at least one of the individual P for the selected cancers. All highlighted loci are independent of bidirectional SNPs with smaller overall P .

significant ($FDR < 0.05$) effects on gene expression (Fig. 4c). Five variants belonged to all three classes (Fig. 4c).

Consistent with hypothesized pleiotropy, 78.1% of the 64 expression quantitative trait loci (eQTLs) identified among the pleiotropic variants had more than one target tissue, and 78.1% influenced the expression of more than one gene (Supplementary

Fig. 3), for a total of 596 significant SNP-gene pairs. The most common expression tissues for eQTLs among pleiotropic variants were whole blood (49.2%), followed by adipose (14.8%) and esophageal (4.7%) tissues. Regulatory effects mediated by chromatin looping were observed for 28 variants, including 3 enhancer-promoter links in 6p21.23 (rs535777, rs73728618) and

Table 4 Top independent variants from the one-directional variant-specific pleiotropic analysis.

SNP	Locus	P	OR	Associated cancer sites
rs6587551	1q21.3	2.23×10^{-9}	1.07	Bl, Ki, Lu, Me, Pa, Pr
rs1398148	1q32.1	2.82×10^{-10}	0.92	Bl, Ki, Me, Pr, Th
rs2349073	2q33.1	3.78×10^{-18}	1.09	Br, En, Es, Le, Me, NHL, Ov, Pa, Te
rs2293607	3q26.2	2.46×10^{-11}	1.08	Bl, Co, En, Es, Ki, Le, Me, Pr, Th
rs148297846	5p15.33	2.42×10^{-8}	0.86	Bl, En, Ki, Lu, Pr
rs130071	6p21.33	3.26×10^{-11}	0.89	Co, Es, NHL, Or, Re, Te
rs2395191	6p21.32	5.93×10^{-10}	0.85	Le, NHL, Or, Th
rs73728618	6p21.32	5.83×10^{-33}	1.48	Ce, NHL
rs113661590	8q24.21	1.06×10^{-8}	1.10	Br, Es, Or, Pr, Re, Th
rs6983267	8q24.21	7.52×10^{-74}	1.24	Co, Pr, Re
9:21964331_CA_C	9p21.3	2.08×10^{-8}	0.92	Br, En, Ki, Lu, Me, Or, Ov, Pa, Pr, Te, Th
rs11813268	10q24.338	2.27×10^{-17}	0.89	En, Ki, Lu, Me, Or, Ov, Pr, Th
rs78378222	17p13.1	2.20×10^{-12}	0.72	Es, Ki, Le, Me, Pa, Pr, Re
rs11263763	17q12	3.78×10^{-47}	0.82	En, Pr, Te
rs2532389	17q21.31	1.17×10^{-9}	1.07	Br, Ce, Lu, Me, Te
rs4939827	18q21.1	2.65×10^{-10}	1.15	Co, Re
rs34978822	20q13.33	8.21×10^{-10}	1.34	Bl, Le, Lu, Me, Pr, Th

OR odds ratio, Bl bladder, Br breast, Ce cervix, Co colon, En endometrium, Es esophagus/stomach, Ki kidney, Le lymphocytic leukemia, Lu lung, Me melanoma, NHL non-Hodgkin's lymphoma, Or oral cavity/pharynx, Ov ovary, Pa pancreas, Pr prostate, Re rectum, Te testis, Th thyroid.

22q13.2 (rs5759167, *PACSN2* promoter; Supplementary Fig. 4). Notably, rs5759167 is an eQTL for *PACSN2* in whole blood (BIOS QTL: $P = 9.89 \times 10^{-14}$; GTE_x v8: $P = 3.39 \times 10^{-7}$).

The functional profile of the 100 pleiotropic variants was significantly different across multiple features when compared to a randomly selected set of 100 independent variants. Pleiotropic variants had a significantly higher proportion of enhancers ($P = 3.38 \times 10^{-4}$), eQTLs ($P = 3.38 \times 10^{-4}$; >1 tissue: $P = 2.33 \times 10^{-3}$; >1 gene: $P = 1.34 \times 10^{-4}$), and chromatin interactions ($P = 3.48 \times 10^{-4}$). Pleiotropic variants did not have a significantly higher proportion in active chromatin states ($P = 0.48$).

Genes represented by pleiotropic variants were significantly enriched for 36 KEGG pathways that formed two clusters broadly characterized by immune-related functions and cancer-specific genes (Supplementary Table 2 and Supplementary Fig. 5). Top-ranking pathways in the first cluster included antigen processing and presentation ($P = 4.29 \times 10^{-6}$), cell adhesion molecules ($P = 4.29 \times 10^{-6}$), allograft rejection ($P = 4.29 \times 10^{-6}$), cancer-related infections (human T cell leukemia virus 1: $P = 4.35 \times 10^{-6}$; Epstein-Barr virus: $P = 3.49 \times 10^{-5}$), and autoimmune diseases (type I diabetes: $P = 9.84 \times 10^{-6}$; inflammatory bowel disease: $P = 1.16 \times 10^{-3}$). The second cluster was enriched for genes related to multiple cancers (gastric: $P = 9.94 \times 10^{-5}$; small cell lung cancer: $P = 3.14 \times 10^{-3}$; prostate: $P = 3.65 \times 10^{-3}$), drug resistance (endocrine resistance: $P = 2.55 \times 10^{-4}$), and cellular senescence ($P = 0.014$).

Discussion

In this study of cancer pleiotropy in two large cohorts, we found multiple lines of evidence for a shared genetic basis of several cancer types. By characterizing pleiotropy at the genome-wide, locus-specific, and variant-specific levels for a large number of cancer sites, we generated several insights into cancer susceptibility. Specifically, we detected 21 previously unreported genome-wide significant variant associations across 11 of the 18 individual cancers examined. We also detected 100 independent variants displaying one- or bidirectional pleiotropy that were enriched for a number of regulatory functions that reflect hallmarks of carcinogenesis.

One notable finding from our cervical cancer GWAS was rs10175462 in *PAX8* on 2q13, which, to our knowledge, is the first genome-wide significant cervical cancer risk SNP identified

outside of the HLA region in a European ancestry population¹⁵. In a candidate SNP study of *PAX8* eQTLs in a Han Chinese population, two variants in LD with rs10175462 in Europeans (rs1110839, $r^2 = 0.33$; rs4848320, $r^2 = 0.34$) were suggestively associated with cervical cancer risk in the same direction³⁷. Several GWAS findings also provided evidence of pleiotropy, in that previously unreported risk variants for one cancer had known associations with one or more other cancers. For instance, rs9818780 was associated with melanoma and has been implicated in sunburn risk³⁸. This intergenic variant is an eQTL for *LINC00886* and *METTL15P1* in skin tissue. The former gene has previously been linked to breast cancer³⁹, and both genes have been implicated in ovarian cancer⁴⁰. Beyond the previously unreported associations, our GWAS detected 308 independent associations with $P < 1 \times 10^{-6}$ that confirmed signals identified in previous GWAS with $P < 5 \times 10^{-8}$. This finding strengthened our confidence in using our genome-wide summary statistics for subsequent analyses of cancer pleiotropy.

In evaluating pairwise genetic correlations between the 18 cancer types, we observed the strongest signal for colon and rectal cancers—an expected relationship consistent with findings from a twin study⁴¹. We also identified several cancer pairs for which the genetic correlations were nominally significant. One pair supported by previous evidence is melanoma and testicular cancer; some studies have found that individuals with a family history of the former are at an increased risk for the latter^{42,43}. Esophageal/stomach cancer was a component of five correlated pairs—with melanoma, NHL, and breast, lung, and rectal cancers. Despite some similarities between esophageal and stomach cancers, testing them as a combined phenotype may have inflated the number of correlated cancers.

Our genetic correlation results contrast with some previous findings^{4–6}; we did not find several correlations that they did and found others that they did not. The differences may be partly due to a smaller number of cases in our cohorts for some sites. Further studies with larger sample sizes are necessary to validate our correlations, as those that did not attain Bonferroni-corrected significance may have been due to chance. However, we achieved comparable or higher cancer-specific heritability estimates for breast, colon, and lung cancers, which suggests that differences in study design may also play a role. Previous analyses aggregated case-control studies recruited during different time periods.

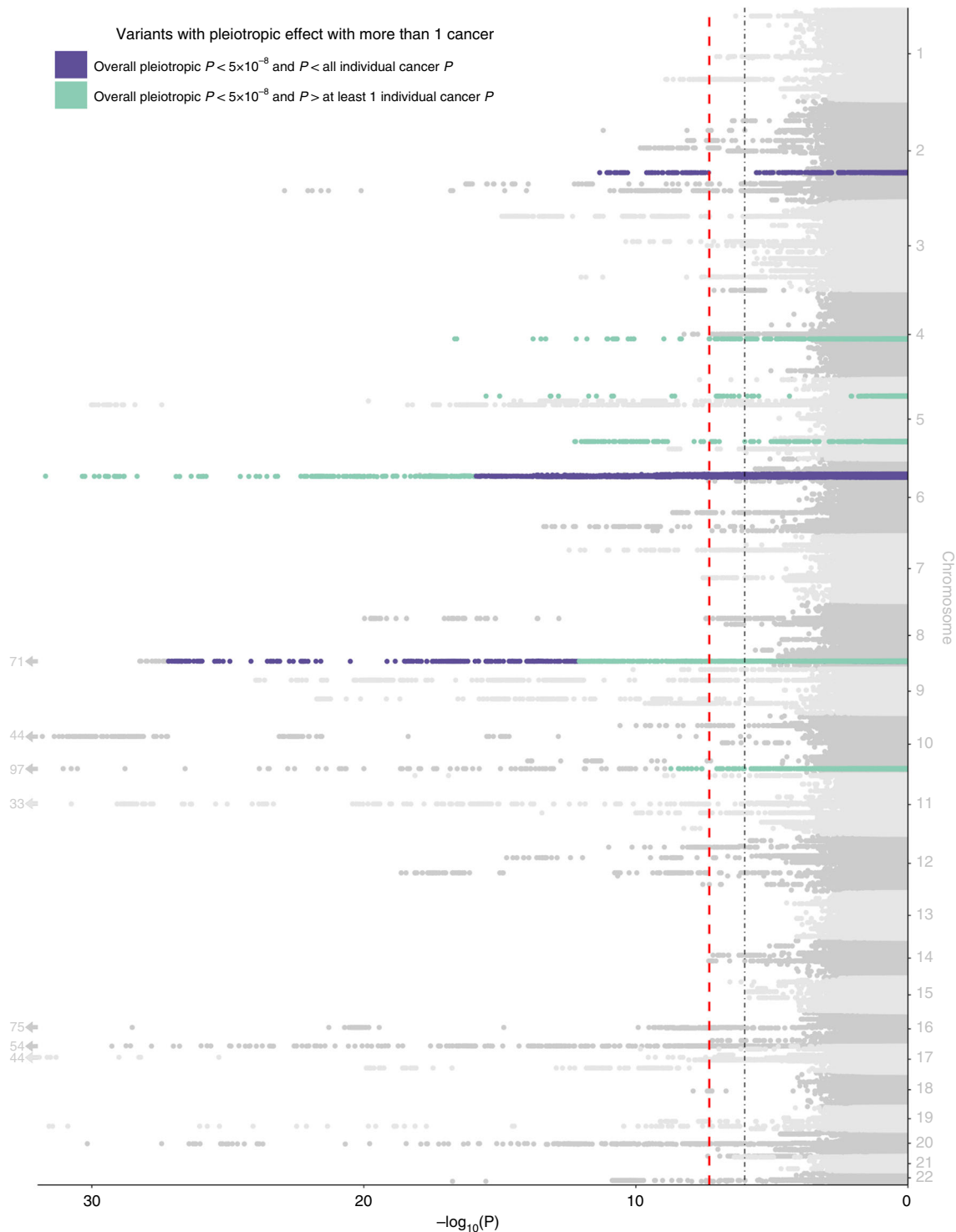


Fig. 3 Manhattan plot displaying bidirectional variant-specific pleiotropy from ASSET. The red dashed line represents the genome-wide significance threshold ($P < 5 \times 10^{-8}$), and the black dotted line represents a suggestive threshold ($P < 1 \times 10^{-6}$). Highlighted are loci with overall pleiotropic $P < 5 \times 10^{-8}$, the two directional $P < 0.05$, and not in LD with a one-directional SNP with smaller P . Loci in purple are genome-wide significant loci where the overall pleiotropic P is less than all individual P for the selected cancers, and loci in green are genome-wide significant loci where the overall pleiotropic P is greater than at least one of the individual P for the selected cancers.

While such meta-analyses can be effective at reducing residual population stratification, our extensive quality control processes also seemingly mitigated population stratification; the mean λ_{GC} across the 18 cancers was 1.02 (standard deviation = 0.027). Moreover, our design allowed for the assessment of cross-cancer relationships in the same set of individuals and the

examination of several cancers that have yet to be studied in large consortia.

The assessment of pleiotropy at the locus level confirmed previously reported associations at 5p15.33, HLA, and 8q24 (refs. 9,12,15,16,18). Out of the 25 pleiotropic loci that we identified, most were at these known cancer pleiotropic loci. Over half, all in

Table 5 Top independent variants from the bidirectional variant-specific pleiotropic analysis.

SNP	Locus	P Overall	P Increasing	P Decreasing	OR Increasing	OR Decreasing	Associated cancer sites increasing	Associated cancer sites decreasing
rs7282844	2q31.1	4.68 × 10 ⁻¹²	1.79 × 10 ⁻¹⁰	8.57 × 10 ⁻⁴	1.17	0.85	Bl, Co, Ki, Pr, Th	En, Le, Pa, Re, Te
rs10007915	4q24	2.19 × 10 ⁻¹⁷	9.12 × 10 ⁻¹⁶	5.58 × 10 ⁻⁴	1.13	0.96	Pr, Re	Bl, Br, Ce, Ki, Le, Me, NHL, Or, Ov, Te, Th
rs35407	5p13.2	3.11 × 10 ⁻¹⁶	2.91 × 10 ⁻²	2.64 × 10 ⁻¹⁶	1.32	0.55	Ki, Le, Pa, Te	Me, Or, Ov
rs717417	5q31.3	5.70 × 10 ⁻¹³	1.03 × 10 ⁻²	1.69 × 10 ⁻¹²	1.04	0.64	Br, Lu, NHL, Or, Pa, Th	Te
rs17190106	6p21.33	1.61 × 10 ⁻¹⁰	5.37 × 10 ⁻¹⁰	1.11 × 10 ⁻²	1.24	0.92	Ce, Lu, NHL, Re	Bl, Le, Me, Or, Pa, Pr, Te, Th
rs9266766	6p21.33	2.22 × 10 ⁻⁹	1.29 × 10 ⁻⁸	7.14 × 10 ⁻³	1.20	0.95	Ce, NHL, Th	Br, Co, Le, Pa, Pr, Re, Te
rs114060326	6p21.33	3.00 × 10 ⁻¹⁰	4.53 × 10 ⁻⁵	2.53 × 10 ⁻⁷	1.30	0.65	Bl, Ce, Le, Lu, Pa, Re, Te	Es, NHL
rs2763979	6p21.33	1.73 × 10 ⁻¹⁴	8.60 × 10 ⁻¹⁰	5.56 × 10 ⁻⁷	1.08	0.89	Ce, Ki, Pr	En, NHL, Or, Re
rs34563311	6p21.32	1.36 × 10 ⁻¹⁶	8.76 × 10 ⁻⁹	3.76 × 10 ⁻¹⁰	1.19	0.80	Ce, Le, Lu, Pa	En, NHL, Th
rs9270747	6p21.32	1.99 × 10 ⁻³²	1.05 × 10 ⁻³²	2.41 × 10 ⁻²	1.26	0.92	Ce	Le, NHL, Re, Te
rs555777	6p21.32	3.65 × 10 ⁻¹²	7.97 × 10 ⁻⁵	1.49 × 10 ⁻⁹	1.08	0.86	Ce, Ki, Pr, Th	Es, Le, NHL, Or, Ov, Re, Te
rs78809737	8q24.21	8.35 × 10 ⁻¹³	8.57 × 10 ⁻¹³	3.02 × 10 ⁻²	1.29	0.80	Es, Ki, Or, Pr, Re	En, Pa
rs62516012	8q24.21	6.64 × 10 ⁻²⁸	2.29 × 10 ⁻²³	4.28 × 10 ⁻⁷	1.18	0.93	Or, Pr	Br, NHL, Th
rs117952826	8q24.21	9.76 × 10 ⁻⁹	4.36 × 10 ⁻⁶	9.93 × 10 ⁻⁵	1.30	0.86	Ki, Or, Pr, Th	Br, Ce, Lu
rs45631563	10q26.13	1.94 × 10 ⁻⁹	6.75 × 10 ⁻⁸	1.18 × 10 ⁻³	1.16	0.89	Br, Co, Le, Lu, Ov, Pa, Te, Th	En, Es, Ki, Or, Pr

OR odds ratio, Bl bladder, Br breast, Ce cervix, Co colon, En endometrium, Es esophagus/stomach, Ki kidney, Le lymphocytic leukemia, Lu lung, Me melanoma, NHL non-Hodgkin's lymphoma, Or oral cavity/pharynx, Ov ovary, Pa pancreas, Pr prostate, Re rectum, Te testis, Th thyroid.

the HLA locus, were associated with cervical cancer and NHL. The two cancers were weakly negatively correlated in the two cohorts combined and nominally significantly negatively correlated in the UKB alone (Supplementary Data 8). The difference may reflect better coverage and imputation of the HLA region in the UKB than in GERA.

Variant-specific analyses provided further evidence in support of locus-specific cancer pleiotropy, including validation of previously reported signals at 1q32 (refs. 7,8) and 2q33 (refs. 9,10) (*ALS2CR12*). Interestingly, our lead 1q32 variant (rs1398148) maps to *PIK3C2B* and is in LD ($r^2 > 0.60$) with known *MDM4* cancer risk variants^{7,8}, suggesting that the 1q32 locus may be involved in modulating both p53- and PI3K-mediated oncogenic pathways. The 100 independent pleiotropic variants (with overall pleiotropic $P < 5 \times 10^{-8}$) mapped to a total of 56 genomic locations (defined by cytoband), which included the six genomic locations to which all 25 of the regions identified from the locus-specific analysis map. Although 99 of the 100 variants showing one- or bidirectional pleiotropic associations are in regions previously associated with cancer, 83 of the 99 were associated with at least one cancer not previously reported.

Out of 100 independent variants identified from the variant-specific pleiotropy analyses, 17 were in 8q24 and 15 were in the HLA region. Different distributions of one- and bidirectional results highlight patterns of directional pleiotropy: of the 15 HLA variants, 7 were bidirectional, while only three of the 17 variants in 8q24 were bidirectional. The HLA region is critical for innate and adaptive immune response and has a complex relationship with cancer risk. Heterogeneous associations with HLA haplotypes have been reported for different subtypes of NHL⁴⁴ and lung cancer⁴⁵, suggesting that relevant risk variants are likely to differ within, as well as between, cancers. Studies have further demonstrated that somatic mutation profiles are associated with HLA class I (ref. 46) and class II alleles⁴⁷. Specifically, mutations that create neoantigens more likely to be recognized by specific HLA alleles are less likely to be present in tumors from patients carrying such alleles. It is thus possible that some of the positive and negative pleiotropy we identified is related to mutation type. These results reinforce the importance of the immune system playing a role in cancer susceptibility.

In contrast to the HLA region, the majority of the 8q24 pleiotropic variants had the same direction of effect for all associated cancers, implying the existence of shared genetic mechanisms driving tumorigenesis across sites. The proximity of the well-characterized *MYC* oncogene makes it a compelling candidate for such a consistent, one-directional effect. It could work via regulatory elements, such as acetylated and methylated histone marks⁴⁸. Consistent with this hypothesis, we observed heritability enrichment⁴⁹ for variants with the H3K27ac annotation for breast ($P = 3.09 \times 10^{-4}$), colon ($P = 4.44 \times 10^{-4}$), prostate ($P = 2.74 \times 10^{-5}$), and rectal ($P = 0.036$) cancers—all of which share susceptibility variants in 8q24, according to our analyses and previous studies⁴⁸.

In silico analyses found the 100 pleiotropic variants to be enriched across multiple regulatory domains compared to non-pleiotropic randomly selected variants and highlighted cross-cancer susceptibility loci. The 11q13.3 region includes rs12275055, which maps to active enhancers and is also an eQTL for *TPCN2*, a gene involved in controlling the angiogenic response to VEGF and extracellular vesicle trafficking in cancer cells^{50,51}. An additional interesting region, 22q13.2, is indexed by rs5759167, an intergenic variant linked to prostate and lung cancers risk. Its pleiotropic effects are likely mediated by regulation of *PACSIN2*, which codes for a cyclin D1 binding partner that serves as a brake for *CCND1*-mediated cellular migration⁵².

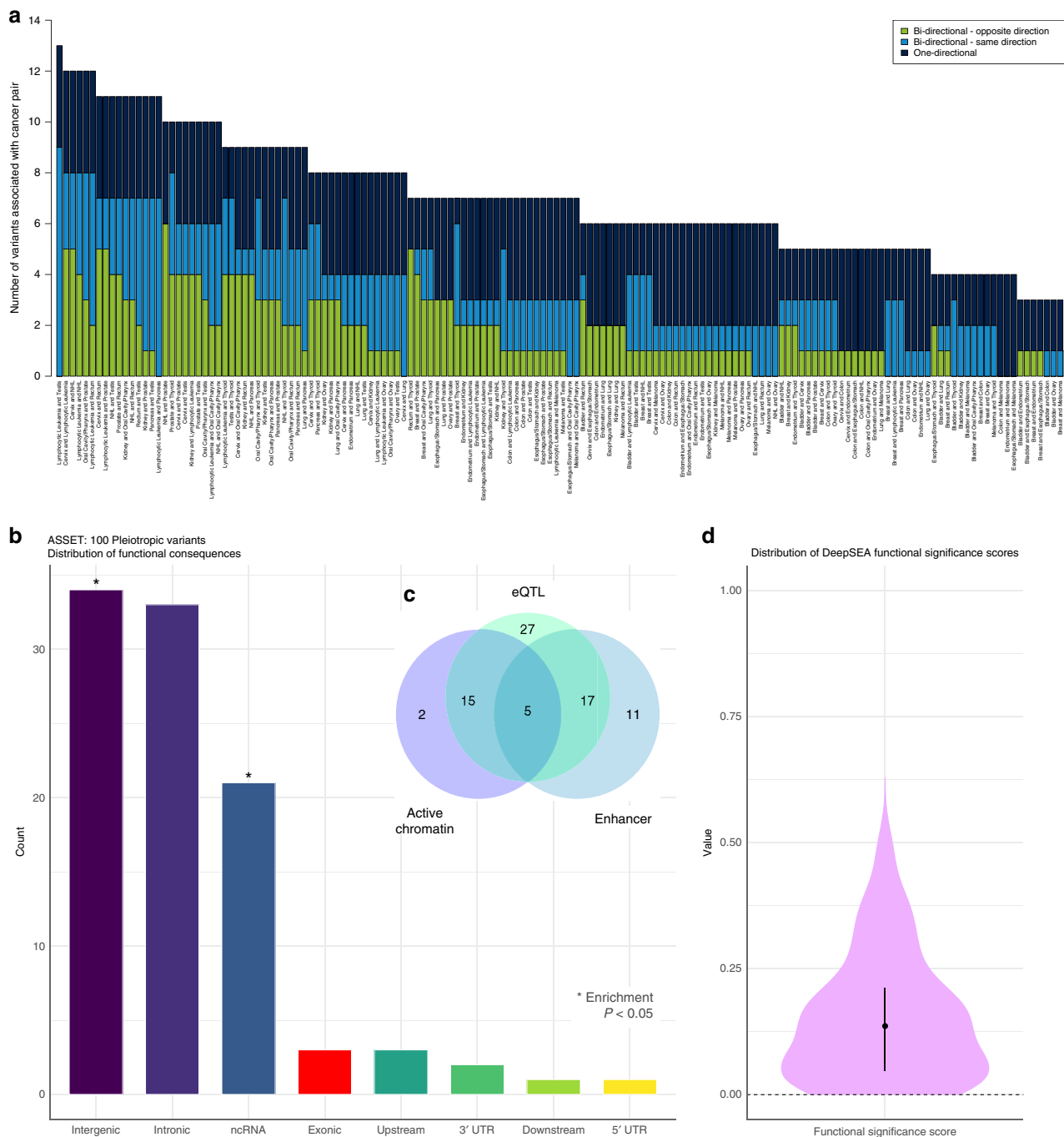


Fig. 4 Summary of cancer pairs associated with and functional consequences of the 100 one- and bidirectional pleiotropic variants. **a** The number of pleiotropic variants (of the independent 100 one- and bidirectional variants with overall pleiotropic $P < 5 \times 10^{-8}$) associated with each pair of cancers by type of pleiotropic effect for select cancer pairs using ASSET: SNPs identified in the one-directional analysis, where all associations are in the same direction (navy); SNPs identified in the bidirectional analysis, where both cancers in the pair are associated in the same direction (both risk increasing or both risk decreasing), even though at least one other cancer is associated in the opposite direction (blue); and SNPs identified in the bidirectional analysis, where the pair of cancers are associated in opposite directions (one risk increasing and one risk decreasing) (green). **b** The distribution of variant consequences and corresponding enrichment, calculated using Fisher’s exact test comparing the proportion of variants belonging to each functional class observed among the 100 ASSET variants to all variants in the UK Biobank. Pleiotropic variants were enriched in intergenic ($P = 0.043$) and non-coding RNA transcripts ($P = 0.015$). **c** Venn diagram summarizing the number of variants with specific regulatory elements, based on analyses of chromatin features from Roadmap and expression quantitative trait loci (eQTL) associations. **d** Distribution of DeepSEA functional significance scores, providing an integrated summary score based on evolutionary conservation and chromatin data, with 0 denoting variants most likely to be functional.

This is consistent with our observation that that the risk-increasing G-allele is associated with increased *PACSN2* expression in whole blood⁵³. Lastly, our pathway analysis indicated that pleiotropic variants as a group are enriched for genes

involved in immune regulation and infection, as well as cancer development and progression. Our in silico findings highlight loci that are good candidates for investigation in future in vivo studies.

It is important to acknowledge some limitations of our study. First, counts for some of the cancer types were limited. However, small sample sizes are partially offset by the advantages of using two population-based cohorts. Second, due to the complexity of the LD structure in the HLA region, we may have overestimated the number of distinct, independent signals. Slight overestimation, however, does not affect our overall conclusions regarding the pleiotropic nature of this region. Third, our analyses included both prevalent and incident cases. Nevertheless, sensitivity analyses restricted to incident cancers yielded comparable results. Fourth, we grouped esophageal and stomach cancers despite possible differences in their risk factor profiles. However, there is precedent for using a composite phenotype⁵⁴, and analyses of stomach and esophageal tumors suggest that they have many overlapping molecular features^{55,56}. In addition, sensitivity analyses for each cancer alone gave similar results, suggesting that they may have similar genetic bases despite potentially having different environmental risk factors. Fifth, we focused solely on individuals of European ancestry. Further analyses are needed to accurately assess patterns of pleiotropy in non-Europeans. Finally, the two distinct cohorts studied here—the UKB and GERA—were recruited from different populations and time periods and were genotyped with different versions of Axiom GWAS arrays. Only variants genotyped or well-imputed across the cohorts were combined in our meta-analysis. Moreover, studying two cohorts provides complementary evidence for pleiotropy.

The characterization of pleiotropy is fundamental to understanding the genetic architecture of cross-cancer susceptibility and its biological underpinnings. The availability of two large, independent cohorts provided an opportunity to efficiently evaluate the shared genetic basis of many cancers, including some not previously studied together. The result was a multifaceted assessment of common genetic factors implicated in carcinogenesis, and our findings illustrate the importance of investigating different aspects of cancer pleiotropy. Broad analyses of genetic susceptibility and targeted analyses of specific loci and variants may both contribute insights into different dimensions of cancer pleiotropy. Future studies should consider the contribution of rare variants to cancer pleiotropy and aim to elucidate the functional pathways mediating associations observed at pleiotropic regions. Such research, combined with our findings, has the potential to inform drug development, risk assessment, and clinical practice toward reducing the burden of cancer.

Methods

Study populations and phenotyping. The UKB is a population-based cohort of 502,611 individuals in the United Kingdom. Study participants were aged 40–69 at recruitment between 2006 and 2010, at which time all participants provided detailed information about lifestyle and health-related factors and provided biological samples⁵⁷. GERA participants were drawn from adult Kaiser Permanente Northern California (KPNC) health plan members who provided a saliva sample for the Research Program on Genes, Environment and Health (RPGEH) between 2008 and 2011. Individuals included in this study were selected from the 102,979 RPGEH participants who were successfully genotyped as part of GERA and answered a baseline survey concerning lifestyle and medical history^{58,59}.

Cancer cases in the UKB were identified via linkage to various national cancer registries established in the early 1970s⁵⁷. Data in the cancer registries are compiled from hospitals, nursing homes, general practices, and death certificates, among other sources. The latest cancer diagnosis in our data from the UKB occurred in August 2015. GERA cancer cases were identified using the KPNC Cancer Registry, including all diagnoses captured through June 2016. Following SEER standards, the KPNC Cancer Registry contains data on all primary cancers (i.e., cancer diagnoses that are not secondary metastases of other cancer sites; excluding non-melanoma skin cancer) diagnosed or treated at any KPNC facility since 1988.

In both cohorts, individuals with at least one recorded prevalent or incident diagnosis of a borderline, in situ, or malignant primary cancer were defined as cases for our analyses. Individuals with multiple cancer diagnoses were classified as a case only for their first cancer. For the UKB, all diagnoses described by International Classification of Diseases (ICD)-9 or ICD-10 codes were converted

into ICD-O-3 codes; the KPNC Cancer Registry already included ICD-O-3 codes. We then classified cancers according to organ site using the SEER site recode paradigm⁶⁰. We grouped all esophageal and stomach cancers and, separately, all oral cavity and pharyngeal cancers to ensure sufficient statistical power. The 18 most common cancer types (except non-melanoma skin cancer) were examined. Testicular cancer data were obtained from the UKB only due to the small number of cases in GERA.

Controls were restricted to individuals who had no record of any cancer in the relevant registries, who did not self-report a prior history of cancer (other than non-melanoma skin cancer), and, if deceased, who did not have cancer listed as a cause of death. Individuals whose first cancer diagnosis was for a cancer not among our 18 cancers of interest were excluded. For analyses of sex-specific cancer sites (breast, cervix, endometrium, ovary, prostate, and testis), controls were restricted to individuals of the appropriate sex.

Quality control. For the UKB population, genotyping was conducted using either the UKB Axiom array (436,839 total; 408,841 self-reported European) or the UK BiLEVE array (49,747 total; 49,746 self-reported European)⁵⁷. The former is an updated version of the latter, such that the two arrays share over 95% of their marker content. UKB investigators undertook a rigorous quality control (QC) protocol⁵⁷. Genotype imputation was performed using the Haplotype Reference Consortium as the main reference panel and the merged UK10K and 1000 Genomes phase 3 reference panels for additional data, resulting in a unified set of 93,095,623 imputed SNPs⁵⁷, which is used for all analyses. Ancestry principal components (PCs) were computed using fastPCA based on a set of 407,219 unrelated samples and 147,604 genetic markers⁵⁷.

For GERA participants, genotyping was performed using an Affymetrix Axiom array (Affymetrix, Santa Clara, CA, USA) optimized for individuals of European race/ethnicity. Details about the array design, estimated genome-wide coverage, and QC procedures have been published previously^{59,61}. The genotyping produced high-quality data with average call rates of 99.7% and average SNP reproducibility of 99.9%. Variants that were not directly genotyped (or that were excluded by QC procedures) were imputed to generate genotypic probability estimates. After pre-phasing genotypes with SHAPE-IT v2.5, IMPUTE2 v2.3.1 was used to impute SNPs relative to the cosmopolitan reference panel from 1000 Genomes. Ancestry PCs were computed based on 144,799 high-performing SNPs using the smartpca program in the EIGENSOFT4.2 software package⁵⁸.

For both cohorts, analyses were limited to self-reported European ancestry individuals for whom self-reported and genetic sex matched. To further minimize potential population stratification, we excluded individuals for whom either of the first two ancestry PCs fell outside five standard deviations of the mean of the population. Based on a subset of genotyped autosomal variants with minor allele frequency (MAF) ≥ 0.01 and genotype call rate $\geq 97\%$, we excluded samples with call rates $< 97\%$ and/or heterozygosity more than five standard deviations from the mean of the population. With the same subset of SNPs, we used KING to estimate relatedness among the samples. We excluded one individual from each pair of first-degree relatives, first prioritizing on maximizing the number of the cancer cases relevant to these analyses and then maximizing the total number of individuals in the analyses. Our study population ultimately included 408,786 UKB participants and 66,526 GERA participants. We excluded SNPs with imputation quality score ($r^2_{\text{INFO}} < 0.3$, call rate $< 95\%$ (alternate allele dosage required to be within 0.1 of the nearest hard call to be non-missing; UKB only), Hardy–Weinberg equilibrium P among controls $< 1 \times 10^{-5}$, and/or MAF < 0.01 , leaving 8,876,519 variants for analysis for the UKB and 8,973,631 for GERA.

For indels, the r^2_{INFO} scores indicated extremely high accuracy, ranging from 0.81 to 0.99 in the UKB (median = 0.99) and from 0.72 to 0.99 in GERA (median = 0.99) (Supplementary Data 1–2). In addition, the correlation was very high between imputed and sequenced genotypes for 44 EUR samples from the 1000 Genomes Project genotyped with the Axiom UK Biobank array and imputed using the 1KGP WGS Phase 3 reference panel: the average r^2 was 0.97 for SNPs and 0.90 for indels (MAF > 0.01 ; Jeremy Gollub, Personal Communication).

Genome-wide association analyses of individual cancers. We used PLINK to implement within-cohort logistic regression models of additively modeled SNPs genome-wide, comparing cases of each cancer type to cancer-free controls. All models were adjusted for age at specimen collection, sex (non-sex-specific cancers only), first ten ancestry PCs, genotyping array (UKB only), and reagent kit used for genotyping (Axiom v1 or v2; GERA only). Case counts ranged from 471 (pancreatic cancer) to 13,903 (breast cancer) in the UKB and from 162 (esophageal/stomach cancer) to 3978 (breast cancer) in GERA (Supplementary Table 3). Control counts were 359,825 (189,855 females) and 50,525 (29,801 females) in the UKB and GERA, respectively. After separate GWAS were conducted in each cohort, association results for the 7,846,216 SNPs in both cohorts were combined via meta-analysis. For variants that were only examined in one cohort (22% of the total 10,003,934 SNPs analyzed), original summary statistics were merged with the meta-analyzed SNPs to create a union set of SNP statistics for each cancer for use in downstream analyses (Supplementary Fig. 6).

To determine independent signals in our union set of SNPs, we implemented the LD clumping procedure in PLINK based on genotype hard calls from a reference panel comprised of a downsampled subset of 10,000 random UKB

participants. For each cancer separately, LD clumps were formed around index SNPs with the smallest P not already assigned to another clump. While only variants with $P < 5 \times 10^{-8}$ were considered significant, to also identify suggestive variants for supplementary results, in each clump, index SNPs had a suggestive association based on $P < 1 \times 10^{-6}$, and SNPs were added if they were marginally significant with $P < 0.05$, were within 500 kb of the index SNP, and had $r^2 > 0.1$ with the index SNP. To confirm independence, we implemented GCTA's conditional and joint analysis (COJO) method with the aforementioned downsampled subset of UKB participants as a reference panel, performing stepwise selection of the index SNPs within a ± 1000 kb region of one another. SNPs were deemed independent if they maintained a $P < 1 \times 10^{-6}$ in the joint model. The remaining independent variants were determined to be novel if they were independent of previously reported risk variants in European ancestry populations (as described below).

To identify SNPs previously associated with each cancer type, we abstracted all genome-wide significant SNPs from relevant GWAS published through June 2018. We determined that a SNP was potentially novel if it had LD $r^2 < 0.1$ with all previously reported SNPs for the relevant cancer based on both the UKB reference panel and the 1000 Genomes EUR superpopulation via LDlink. As an additional filter for novelty, we again used COJO to condition each potentially novel SNP on previously reported SNPs for the relevant cancer using the UKB reference panel, and SNPs were not considered novel if they did not maintain $P < 1 \times 10^{-6}$ in the joint model. To confirm novelty and consider pleiotropy, we conducted an additional literature review to investigate whether these SNPs had previously been reported for the same or other cancers, including those not attaining genome-wide significance and those in non-GWAS analyses. For this additional review, we used the PhenoScanner database to search for SNPs of interest and variants in LD in order to comprehensively scan previously reported associations. We then supplemented with more in-depth PubMed searches to determine if the genes in which novel SNPs were located had previously been reported for the same or other cancers. Finally, for cancers with publicly available summary statistics (breast [$>120,000$ cases]³⁹, prostate [$\sim 80,000$ cases]⁶², and ovarian [$\sim 30,000$ cases]⁴⁰), we tested our potentially previously unreported SNPs with $P < 1 \times 10^{-6}$ for replication (defined as having the same direction of effect and $P < 0.05$). Tested SNPs that did not replicate were not considered previously unreported.

We considered whether clinical characteristics of the cases were informative about associated phenotypes by examining SEER stage and grade (both GERA only) and age at cancer diagnosis (UKB and GERA). For each clinical variable, we decomposed cases into one of two categories: grade 1–2 (well or moderately differentiated) or grade 3–4 (poorly or undifferentiated); stage 0–1 (in situ or localized) or stage 2–7 (regional or distant metastases); age $<$ median or age \geq median. The case counts for all cancer-outcome strata are tabulated in Supplementary Table 4. For each of the previously unreported GWAS SNPs, we conducted logistic regression comparing controls to each of the relevant case subtypes. We then compared the effect estimates across the strata for each clinical variable (e.g., for each relevant SNP–cancer pair, we compared the OR for grade 1–2 with the OR for grade 3–4) and calculated Cochran's Q statistic to test for heterogeneity, adjusting for multiple testing for the number of strata and SNPs tested.

To assess whether our results were influenced by factors associated with survival, we conducted sensitivity analyses restricted to incident cases in the larger UKB cohort. For each cancer, we compared the independent SNPs that were suggestively associated in the analysis using both prevalent and incident cases ($P < 1 \times 10^{-6}$) with those in the incident only analysis. We assessed whether the effect sizes varied by calculating Cochran's Q statistic to test for heterogeneity, adjusting for multiple testing across the number of SNPs tested for each cancer. Additional sensitivity analyses evaluated esophageal and stomach cancers as separate phenotypes in the UKB cohort. For independent SNPs with $P < 1 \times 10^{-6}$ in the analysis of the composite phenotype in UKB alone, we compared effect sizes for the composite phenotype to effect sizes for esophageal and stomach cancers separately and calculated Cochran's Q statistic to test for heterogeneity, adjusting for multiple testing across the number of SNPs tested. For both of these sensitivity analyses, we assessed all SNPs with $P < 1 \times 10^{-6}$ to allow for a sufficient number of variants for comparison.

Genome-wide heritability and genetic correlation. We used LD-score regression (LDSC) on summary statistics from the union set of all SNPs genome-wide to calculate the genome-wide liability-scale heritability of each cancer type and the genetic correlation between each pair of cancer types. Internal LD scores were calculated using the aforementioned downsampled subset of UKB participants. To convert to liability-scale heritability, we adjusted for lifetime risks of each cancer based on SEER 2012–2014 estimates (Supplementary Table 5)⁶³. LDSC was unable to estimate genetic correlations for testicular cancer with both oral cavity/pharyngeal and pancreatic cancers, likely due to small sample sizes.

Locus-specific pleiotropy. Using our union set of SNP-based summary statistics, we constructed pleiotropic regions of SNPs associated with more than one cancer with $P < 5 \times 10^{-8}$. Non-overlapping regions were iteratively formed around index SNPs associated with any cancer, beginning with the SNP associated with the smallest P . SNPs were added to a region if they were associated with any cancer

with $P < 5 \times 10^{-8}$, were within 500 kb of the index SNP, and had LD $r^2 > 0.5$ with the index SNP. We used a larger threshold for assessing pleiotropic regions ($r^2 > 0.5$) than for identifying truly independent signals ($r^2 > 0.1$; above) to ensure that all SNPs within a region were in LD. If all SNPs in a region were associated with the same cancer, the region was not considered pleiotropic.

Genome-wide variant-specific pleiotropy. We quantified one-directional and, separately, bidirectional variant-specific pleiotropy via the R package ASSET (association analysis based on subsets)⁶⁴. Briefly, ASSET explores all possible subsets of traits for the presence of association signals, resulting in the best combination of traits to maximize the test statistic⁶⁴. ASSET has two procedures: in one, all traits are assumed to be associated with a variant in the same effect direction (one-directional pleiotropy); in the other, variants can be associated with traits in opposite directions (bidirectional pleiotropy)⁶⁴. In the one-directional pleiotropy analysis, an overall P across the selected traits is provided, and in the bidirectional pleiotropy analysis, a P for each direction is provided as well as an overall P for the total association signal for both directions combined. ASSET corrects for the internal multiple testing burden accrued by iterating through all possible trait subsets for each variant as well as controlling for shared samples among the traits⁶⁴.

Genome-wide ASSET analyses were conducted on the union sets of summary statistics for all 18 cancers. Independent variants were determined via LD clumping, where index SNPs were suggestively significant (overall $P < 1 \times 10^{-6}$), and other SNPs were clumped with the lead variant if they had overall $P < 0.05$, were within 500 kb of the index SNP, and had $r^2 > 0.1$ with the index SNP. While we only considered variants with an overall $P < 5 \times 10^{-8}$ significant, we used a suggestive significance threshold to comprehensively assess all potentially pleiotropic variants. A SNP was determined to have a one-directional pleiotropic association if the overall P was $< 1 \times 10^{-6}$ and it was associated with at least two cancers. A SNP was determined to have a bidirectional pleiotropic association if the overall P was $< 1 \times 10^{-6}$ and the P for each direction was < 0.05 . For one- and bidirectional SNPs in LD with each other, the SNP with the smaller overall P was retained. We deconstructed bidirectional associations into cancers with risk-increasing effects and cancers with risk-decreasing effects.

To assess whether clinical aspects of the cases could be informative about the pleiotropic variants, for each of the one-directional and bidirectional pleiotropic SNPs, we conducted logistic regression comparing controls to each of the relevant case subtypes described above and calculated Cochran's Q statistic to test for heterogeneity between estimates across the strata for each clinical variable.

Functional characterization of pleiotropic variants. Functional consequences for the 100 pleiotropic variants identified in the ASSET analysis were obtained from ANNOVAR. Enrichment of functional classes was evaluated using Fisher's exact test, comparing the distribution observed among the pleiotropic variants to that of all variants with INFO > 0.90 in the reference panel of UKB European descent individuals (16,972,700 SNPs total).

Overall functional significance was assessed using DeepSEA, a deep learning tool that prioritizes functional variants by integrating regulatory binding and ENCODE modification patterns of ~ 900 cell-factor combinations with evolutionary conservation features. Resulting functional significance scores, ranging from 0 to 1, represent the degree of deviation from a reference distribution of 1000 Genomes variants, with lower scores indicating a higher likelihood of functional significance. We also report CADD scores, which combine over 60 diverse annotations to predict deleteriousness³⁶. CADD scores are transformed into a log10-derived rank score based on the genome-wide distribution of scores for 8.6 billion single-nucleotide variants in GRCh37/hg19 (i.e., CADD = 10 corresponds to top 10% most deleterious substitutions)³⁶.

To assess more specific functional features, we annotated each SNP according to Roadmap's 15-core chromatin states across 127 cell or tissue types^{35,65}. Chromatin state was assigned by taking the most common state, with values ≤ 7 indicating open, accessible chromatin regions. Three-dimensional chromatin interactions were explored to identify significant interaction and enhancer-promoter links. We also explored associations with gene expression in using data from the GTEx v8 and BIOS QTL databases. The distribution of functional features among pleiotropic cancer risk variants was compared to a random sample of the same number of SNPs. Chromatin features and BIOS QTL annotations were obtained from the FUMA (Functional Mapping and Annotation) database. Differences in the proportion of variants belonging to each functional class were tested using a two-sample chi-squared test. Lastly, after annotating variants to their nearest gene, we conducted gene-set pathway enrichment analyses using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database⁶⁶ with an FDR $q < 0.05$ significance threshold.

Ethics. The study was approved by the University of California and KPNC Institutional Review Boards and the UKB data access committee, and informed consent was obtained from all participants.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Our meta-analysis summary statistics are publicly available at https://github.com/Wittelab/pancancer_pleiotropy. The UKB cohort data is publicly available from the UKB access portal at <https://www.ukbiobank.ac.uk>. The Kaiser Permanente data are available via application with a local collaborator at <https://researchbank.kaiserpermanente.org/our-research/for-researchers/>. All remaining relevant data are available in the article, supplementary information, or from the corresponding author upon reasonable request.

Received: 2 January 2020; Accepted: 13 August 2020;

Published online: 04 September 2020

References

- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
- Mucci, L. A. et al. Familial risk and heritability of cancer among twins in Nordic countries. *JAMA* **315**, 68–76 (2016).
- Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* **99**, 260–266 (2002).
- Sampson, J. N. et al. Analysis of heritability and shared heritability based on genome-wide association studies for 13 cancer types. *J. Natl Cancer Inst.* **107**, djv279 (2015).
- Lindström, S. et al. Quantifying the genetic correlation between multiple cancer types. *Cancer Epidemiol. Biomark. Prev.* **26**, 1427–1435 (2017).
- Jiang, X. et al. Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **10**, 431 (2019).
- Couch, F. J. et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.* **9**, e1003212 (2013).
- Eeles, R. A. et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45**, 385–391, 391e1–2 (2013).
- Barrett, J. H. et al. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.* **43**, 1108–1113 (2011).
- Broeks, A. et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.* **20**, 3289–3303 (2011).
- Ellinghaus, E. et al. Identification of germline susceptibility loci in ETV6-RUNX1-rearranged childhood acute lymphoblastic leukemia. *Leukemia* **26**, 902–909 (2012).
- Rothman, N. et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* **42**, 978–984 (2010).
- Eeles, R. A. et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).
- Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
- Bahrami, A. et al. Genetic susceptibility in cervical cancer: from bench to bedside. *J. Cell Physiol.* **233**, 1929–1939 (2017).
- Smedby, K. E. et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS Genet.* **7**, e1001378 (2011).
- Jin, G. et al. Genetic variants at 6p21.1 and 7p15.3 are associated with risk of multiple cancers in Han Chinese. *Am. J. Hum. Genet.* **91**, 928–934 (2012).
- Eeles, R. A. et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).
- Purdue, M. P. et al. Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nat. Genet.* **43**, 60–65 (2011).
- Spurdle, A. B. et al. Genome-wide association study identifies a common variant associated with risk of endometrial cancer. *Nat. Genet.* **43**, 451–454 (2011).
- Couch, F. J. et al. Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of breast cancer and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers. *Cancer Epidemiol. Biomark. Prev.* **21**, 645–657 (2012).
- Setiawan, V. W. et al. Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. *Carcinogenesis* **35**, 2068–2073 (2014).
- Rafnar, T. et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat. Genet.* **41**, 221–227 (2009).
- Cheng, I. et al. Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia. *Gut* **63**, 800–807 (2014).
- Jones, C. C. et al. Cross-cancer pleiotropic associations with lung cancer risk in African Americans. *Cancer Epidemiol. Biomark. Prev.* **28**, 715–723 (2019).
- Hung, R. J. et al. Cross cancer genomic investigation of inflammation pathway for five common cancers: lung, ovary, prostate, breast, and colorectal cancer. *J. Natl. Cancer Inst.* **107**, djv246 (2015).
- Qian, D. C. et al. Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions. *Hum. Mol. Genet.* **24**, 7406–7420 (2015).
- Fehringer, G. et al. Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. *Cancer Res.* **76**, 5103–5114 (2016).
- Toth, R. et al. Genetic variants in epigenetic pathways and risks of multiple cancers in the GAME-ON consortium. *Cancer Epidemiol. Biomark. Prev.* **26**, 816–825 (2017).
- Karami, S. et al. Telomere structure and maintenance gene variants and risk of five cancer types. *Int. J. Cancer* **139**, 2655–2670 (2016).
- Broderick, P. et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317 (2007).
- Kote-Jarai, Z. et al. Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. *Hum. Genet.* **129**, 687–694 (2011).
- Parikh, H. et al. Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels. *Hum. Genet.* **129**, 675–685 (2011).
- Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
- Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
- Rentsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
- Han, J. et al. Expression quantitative trait loci in long non-coding RNA PAX8-AS1 are associated with decreased risk of cervical cancer. *Mol. Genet. Genomics* **291**, 1743–1748 (2016).
- Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
- Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
- Phelan, C. M. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
- Graff, R. E. et al. Familial risk and heritability of colorectal cancer in the Nordic Twin Study of Cancer. *Clin. Gastroenterol. Hepatol.* **15**, 1256–1264 (2017).
- Hemminki, K. & Chen, B. Familial risks in testicular cancer as aetiological clues. *Int. J. Androl.* **29**, 205–210 (2006).
- Zhang, L. et al. Familial associations in testicular cancer with other cancers. *Sci. Rep.* **8**, 10880 (2018).
- Wang, S. S. et al. HLA Class I and II diversity contributes to the etiologic heterogeneity of non-Hodgkin lymphoma subtypes. *Cancer Res.* **78**, 4086–4096 (2018).
- Ferreiro-Iglesias, A. et al. Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat. Commun.* **9**, 3927 (2018).
- Marty, R. et al. MHC-I genotype restricts the oncogenic mutational landscape. *Cell* **171**, 1272–1283.e15 (2017).
- Marty Pyke, R. et al. Evolutionary pressure against MHC class II binding cancer mutations. *Cell* **175**, 416–428.e13 (2018).
- Grisanzio, C. & Freedman, M. L. Chromosome 8q24-associated cancers and MYC. *Genes Cancer* **1**, 555–559 (2010).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Favia, A. et al. VEGF-induced neoangiogenesis is mediated by NAADP and two-pore channel-2-dependent Ca²⁺ signaling. *Proc. Natl Acad. Sci. USA* **111**, E4706–E4715 (2014).
- Sun, W. & Yue, J. TPC2 mediates autophagy progression and extracellular vesicle secretion in cancer cells. *Exp. Cell Res.* **370**, 478–489 (2018).
- Meng, H. et al. PACSIN 2 represses cellular migration through direct association with cyclin D1 but not its alternate splice form cyclin D1b. *Cell Cycle* **10**, 73–81 (2011).
- Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).

54. Okines, A. F. C. et al. Biomarker analysis in oesophagogastric cancer: results from the REAL3 and TransMAGIC trials. *Eur. J. Cancer Oxf. Engl.* **49**, 2116–2125 (2013).
55. Barra, W. F. et al. GEJ cancers: gastric or esophageal tumors? searching for the answer according to molecular identity. *Oncotarget* **8**, 104286–104294 (2017).
56. Cancer Genome Atlas Research Network et al. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
57. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
58. Banda, Y. et al. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285–1295 (2015).
59. Kvale, M. N. et al. Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1051–1060 (2015).
60. Site Recode ICD-O-3/WHO 2008 Definition. https://seer.cancer.gov/siterecode/icdo3_dwhohome/index.html. Accessed 30, 2017.
61. Hoffmann, T. J. et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).
62. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
63. Howlander, N. et al. SEER Cancer Statistics Review, 1975–2014, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2014/, based on November 2016 SEER data submission, posted to the SEER web site, April 2017.
64. Bhattacharjee, S. et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
65. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
66. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

Acknowledgements

This research was supported by the following National Institutes of Health grants: R01CA088164, R01CA201358, R25CA112355, K07CA188142, K24CA169004, and U01CA127298, and the UCSF Goldberg-Benioff Program in Cancer Translational Biology. The UK Biobank analyses were conducted using the UKB Resource under application number 14105. Support for participant enrollment, survey completion, and biospecimen collection for the RPGEH was provided by the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, and Kaiser Permanente national and regional community benefit programs. Genotyping of the GERA cohort was funded by a grant from the National Institute on Aging, the National Institute of Mental Health, and the NIH Common Fund (RC2 AG036607). We thank the Breast Cancer Association Consortium (BCAC) for breast cancer summary statistics (<http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/>), the Ovarian Cancer Association Consortium (OCAC) for

ovarian cancer summary statistics (<http://ocac.ccge.medschl.cam.ac.uk/data-projects/results-lookup-by-region/>), and the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium for prostate cancer summary statistics (http://practical.icr.ac.uk/blog/?page_id=8164). We thank Drs. Jeremy Gollub and Anuradha Mittal at Thermo Fisher Scientific for providing information on the correlation between imputed and sequenced genotypes on the Axiom UKBiobank array.

Author contributions

S.R.R. and R.E.G. contributed by designing presented idea, conducting the analyses, and writing the manuscript. L.K. contributed to writing the manuscript. K.K.T. contributed by conducting analyses. S.E.A., M.A.B., T.B.C., D.A.C., N.C.E., J.D.H., E.J., L.H.K., T.J.M., S.K.V.D.E., E.Z., L.A.H., and T.J.H. aided in data acquisition, provided critical feedback, and helped shape the research, analysis, and manuscript. L.C.S. and J.S.W. contributed to study conception and design, supervised the project, and writing the manuscript.

Competing interests

J.S.W. is a non-employee co-founder of Avail.bio and serves as an expert witness for Pfizer and Sanofi. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-18246-6>.

Correspondence and requests for materials should be addressed to L.C.S. or J.S.W.

Peer review information *Nature Communications* thanks Angela Cox, Marza de Andrade, and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020