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

Sara R. Rashkin1,8, Rebecca E. Graff1,2,8, Linda Kachuri1, Khanh K. Thai2, Stacey E. Alexeeff2, Maruta A. Blatchins2, Taylor B. Cavazos1,3, Douglas A. Corley2, Nima C. Emami1,3, Joshua D. Hoffman1, Eric Jorgenson2, Lawrence H. Kushi2, Travis J. Meyers1, Stephen K. Van Den Eeden2,4, Elad Ziv5,6,7, Laurel A. Habel2, Thomas J. Hoffmann1,2,5, Lori C. Sakoda2,9 & John S. Witte1,4,5,7,9

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.

https://doi.org/10.1038/s41467-020-18246-6

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. 8These authors contributed equally: Sara R. Rashkin, Rebecca E. Graff. 9These authors jointly supervised this work: Lori C. Sakoda, John S. Witte. *Email: lori.sakoda@kp.org; jwitte@ucsf.edu

NATURE COMMUNICATIONS | (2020) 11:4423 | https://doi.org/10.1038/s41467-020-18246-6 | www.nature.com/naturecommunications
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 disease. 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 twins, families, and unrelated populations 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) [8, 9], 2q33 (CASP8-AL2SC2R12) [10, 11], 3q28 (TP63) [12, 13], 4q24 (TET2) [14, 15], 5p15 (TERT-CLPTM1L) [16, 17], 6p21 (HLA complex), 7p15 [18, 19], 8q24, 11q13 [18, 19], 11q13 (HLA complex) [15, 16], 7p15 [17], 8q24 [12, 18], 11q13 [18, 19], 17q12 (HNF1B) [20, 21], and 19q13 (MERIT) [22]. In addition, recent studies have tested single-nucleotide polymorphisms (SNPs) previously associated with one cancer to discover pleiotropic associations with other cancer types [23–25]. Consorls, 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 cancers [4]. Another study of six cancer types found correlations between colorectal, ovarian, and prostate cancers [26].

Genome-wide association analyses of individual cancers have identified loci associated with multiple cancer types, including those that are less common [31]. We detected 25 pleiotropic regions—defined using our linkage disequilibrium (LD) clumping procedure; see “Methods” [32] and Fig. 1b and Supplementary Data 3). However, at a nominal threshold of P = 0.05, we observed suggestive relationships between 11 other pairs. Among cancers, only colon and rectal cancers (r_g = 0.85, P = 5.33 × 10^{-7}) were genetically correlated at a Bonferroni-corrected significance threshold of P = 0.05/153 = 3.27 × 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 = 0.00089), 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 pleotropic regions associated with more than one cancer (P < 5 × 10^{-8}) for each cancer; independent regions were defined using our linkage disequilibrium (LD) clumping procedure; see “Methods” [33].
| Cancer site | SNP | Chromosome | Position | Gene | REF/ALT | MAF UKB | MAF GERA | OR UKB | OR GERA | OR Meta | Meta P |
|-------------|-----|------------|----------|------|---------|---------|----------|--------|---------|---------|-------|
| Bladder     | rs76088467 | 6 | 21795787 | CASC15 | G/A | 0.025 | 0.030 | 0.67 | 0.59 | 0.64 | 2.34 × 10^{-8} |
| Breast      | rs6752414 | 2 | 121425339 | Intergenic | T/C | 0.077 | 0.083 | 0.89 | 0.87 | 0.88 | 1.81 × 10^{-9} |
| Breast      | rs382027702 | 15 | 49725815 | FAM227B | A/C | 0.48 | 0.48 | 1.06 | 1.08 | 1.06 | 2.68 × 10^{-8} |
| Cervix      | rs10754676 | 2 | 13398492 | PAX8 | A/G | 0.36 | 0.37 | 1.16 | 1.08 | 1.15 | 7.71 × 10^{-14} |
| Cervix      | rs23056433 | 6 | 32157364 | PFK2 | A/G | 0.063 | 0.047 | 0.76 | 0.88 | 0.77 | 1.24 × 10^{-15} |
| Colon       | rs735166742 | 8 | 103561978 | Upstream of ODF1 | G/C | 0.015 | 0.017 | 0.65 | 0.61 | 0.64 | 1.27 × 10^{-8} |
| Colon       | rs88114648 | 20 | 7833046 | Intergenic | G/A | 0.14 | 0.15 | 0.83 | 0.84 | 0.83 | 2.30 × 10^{-9} |
| Esophagus/  | rs5450256 | 2 | 106687838 | C2orf40 | G/A | 0.024 | 0.022 | 0.52 | 0.67 | 0.53 | 1.04 × 10^{-8} |
| Stomach     | rs11224829 | 2 | 5100352 | ROAR | A/G | 0.024 | 0.025 | 0.53 | 0.62 | 0.55 | 3.36 × 10^{-9} |
| Lung        | rs10868399 | 1 | 21666218 | Intergenic | A/G | 0.42 | 0.42 | 1.23 | 1.09 | 1.18 | 3.50 × 10^{-8} |
| Lung        | rs11490990 | 5 | 12835592 | Intergenic | A/G | 0.224 | 0.228 | 0.67 | 0.57 | 0.64 | 3.51 × 10^{-8} |
| Lymphocytic | rs15434860 | 8 | 13012376 | DCT | C/T | 0.17 | 0.16 | 1.30 | 1.18 | 1.26 | 2.86 × 10^{-8} |
| Lymphocytic | rs14490818 | 3 | 12609901 | Intergenic | A/G | 0.022 | 0.011 | 0.53 | 0.48 | 0.48 | 3.95 × 10^{-8} |
| Leukemia    | rs61954738 | 13 | 95571786 | Intergenic | T/C | 0.023 | 0.023 | 0.52 | 0.44 | 0.49 | 3.95 × 10^{-8} |
| Leukemia    | rs71378222 | 17 | 7517152 | Intergenic | G/T | 0.12 | 0.14 | 0.44 | 0.34 | 0.40 | 1.89 × 10^{-9} |
| Leukemia    | rs9818790 | 3 | 154492758 | Intergenic | C/T | 0.49 | 0.48 | 0.92 | 0.89 | 0.91 | 3.16 × 10^{-9} |
| Melanoma    | rs12186662 | 5 | 90356197 | ADGRV1 | G/A | 0.32 | 0.36 | 0.90 | 0.89 | 0.90 | 1.09 × 10^{-8} |
| Melanoma    | rs55797833 | 9 | 21990044 | Intergenic | G/T | 0.023 | 0.022 | 1.71 | 1.72 | 1.71 | 6.71 × 10^{-12} |
| Melanoma    | rs78378222 | 17 | 7517152 | Intergenic | G/T | 0.022 | 0.014 | 0.70 | 0.63 | 0.67 | 118 × 10^{-8} |
| Rectum      | rs14653185 | 2 | 23465764 | Intergenic | C/T | 0.013 | 0.018 | 0.57 | 0.50 | 0.55 | 4.36 × 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.
| 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)\(^f\)               |
| Cervix                      | 0.07 (0.02–0.12)           |                                  | 0.09 (0.07–0.11)\(^g\)             | 0.13 (0.06–0.15)\(^h\)               |
| Colon                       | 0.07 (0.04–0.10)           | 0.09 (0.07–0.11)\(^i\)           | 0.15 (0.00–0.45)                   |                                       |
| Endometrium                 | 0.13 (0.07–0.18)           | 0.18 (0.09–0.27)                 | 0.07 (0.02–0.12)\(^d\)             |                                       |
| Esophagus/stomach           | 0.14 (0.07–0.21)           | 0.38 (0.17–0.59)\(^j\)           | 0.22 (0.00–0.55)\(^k\)             |                                       |
| 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)\(^l\)           | 0.09 (0.09–0.16)\(^m\)             |                                       |
| 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)\(^n\)           | 0.10 (0.08–0.10)\(^o\)             |                                       |
| 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)\(^p\)           | 0.10 (0.04–0.16)                   | 0.57 (0.51–0.63)                     |
| Prostate                    | 0.16 (0.13–0.20)           | 0.22 (0.16–0.28)\(^q\)           | 0.09 (0.09–0.16)\(^m\)             |                                       |
| Rectum                      | 0.11 (0.07–0.16)           | 0.18 (0.14–0.22)                 | 0.10 (0.04–0.16)                   | 0.57 (0.51–0.63)                     |
| Testis                      | 0.26 (0.15–0.38)           | 0.30 (0.08–0.51)                 | 0.25 (0.15–0.37)\(^d\)             | 0.53 (0.52–0.53)\(^d\)               |
| Thyroid                     | 0.21 (0.09–0.33)           |                                  |                                  |                                       |

\(^a\)Taken from ref. 6, 95% CI calculated from provided standard error.
\(^b\)Taken from ref. 4.
\(^c\)Taken from ref. 2, except where not included in analysis or 95% CI range was >0.60; remaining taken from ref. 3, as marked.
\(^d\)Taken from ref. 3, family-based not twin.
\(^e\)Estrogen receptor negative (ER\(^{-}\)).
\(^f\)For in situ (invasive: \(h^{2} = 0.22 \ [0.14–0.27]\)).
\(^g\)Colorectal.
\(^h\)For esophageal in Asian population (stomach in Asian population: \(h^{2} = 0.25 \ [0.00–0.52]\)).
\(^i\)Stomach.
\(^j\)For chronic lymphocytic leukemia.
\(^k\)Age >15 years.
\(^l\)For diffuse large B cell lymphoma.
\(^m\)Taken 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.
Supplementary Table 1). Most were at known cancer pleiotropic loci: HLA (14 regions), 8q24 (7 regions), TERT-CLPTM1L (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-CLPTM1L, 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, which is 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 cancer32,33.

Genome-wide variant specific pleiotropy was defined 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

| Cancer site 1         | Cancer site 2          | $r_p$ (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)     | $5.33 \times 10^{-7}$ |
| 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.
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         | OR     | Associated cancer sites |
|-------------------|---------------|--------|-------------------------|
| rs6578551         | 1q21.3        | 2.23×10⁻⁹ | Bl, Ki, Lu, Me, Pa, Pr   |
| rs1398148         | 1q32.1        | 2.82×10⁻¹⁰| Bl, Ki, Me, Pr, Th      |
| rs2349073         | 2q33.1        | 3.78×10⁻¹⁸| Br, En, Es, Le, Me, NHL, Ov, Pa, Te |
| rs2293607         | 3q26.2        | 2.46×10⁻¹¹| Bl, Co, En, Es, Ki, Le, Me, Pr, Th |
| rs148297846       | 5p15.33       | 2.42×10⁻⁸ | Bl, En, Ki, Lu, Pr     |
| rs13007I          | 6p21.33       | 3.26×10⁻¹¹| Co, Es, NHL, Or, Re, Te|
| rs239519I         | 6p21.32       | 5.93×10⁻¹⁰| Le, NHL, Or, Th        |
| rs73728618        | 6p21.32       | 5.83×10⁻³³| Ce, NHL                |
| rs113661590       | 8q24.21       | 1.06×10⁻⁸ | Br, Es, Or, Pr, Re, Th |
| rs6893267         | 8q24.21       | 7.52×10⁻⁷ | Co, Or, Re             |
| rs21964323_CA_C   | 9p21.3        | 2.08×10⁻⁸ | Br, En, Ki, Lu, Me, Or, Ov, Pa, Pr, Te, Th |
| rs18132628        | 10q24.33      | 2.27×10⁻¹⁷| Bl, Ki, Lu, Me, Or, Ov, Pr, Th |
| rs78378222        | 17p13.1       | 2.20×10⁻¹²| Es, Ki, Le, Me, Pa, Pr, Re |
| rs11263763        | 17q12         | 3.78×10⁻⁴⁷| En, Pr, Te             |
| rs2532389         | 17q21.31      | 1.17×10⁻⁹ | Br, Ce, Lu, Me, Te     |
| rs4939827         | 18q21.1       | 2.65×10⁻¹⁰| Co, Re                 |
| rs34978822        | 20q13.33      | 8.21×10⁻¹⁰| 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.

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 population. 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 direction. Several GWAS findings also provided evidence of pleiotropy, in that previously unreported variant associations 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 risk. This intergenic variant is an eQTL for LINC00886 and METTL15P1 in skin tissue. The former gene has previously been linked to breast cancer, and both genes have been implicated in ovarian cancer. Beyond the previously unreported associations, our GWAS detected 308 independent associations with $P < 10^{-6}$ that confirmed signals identified in previous GWAS with $P < 5 	imes 10^{-8}$. 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 study. 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. 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; however, 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.
While such meta-analyses can be effective at reducing residual population stratification, our extensive quality control processes also seemingly mitigated population stratification; the mean $\lambda_{GC}$ 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            | Overall Increasing OR (95% CI) | Overall Decreasing OR (95% CI) | Increasing OR (95% CI) | Decreasing OR (95% CI) | Associated cancer sites |
|-----------------|------------------|-------------------------------|-------------------------------|-------------------------|------------------------|------------------------|
| rs17190106      | 6p21.33          | $1.61 \times 10^{-10}$        | $8.57 \times 10^{-10}$        | 1.17                    | 0.85                   | Bl, Co, Ki, Pr, Th En, Le, Pa, Re, Te |
| rs9266766       | 6p21.33          | $2.22 \times 10^{-4}$         |                               | 1.13                    | 0.96                   | Pr, Re Bl, Br, Ce, Ki, Le, Me, Th |
| rs114060326     | 6p21.33          | $9.12 \times 10^{-10}$        |                               | 1.17                    | 0.85                   | Ce, Le, Lu, NHL, Re, Th |
| rs34563311      | 6p21.32          |                               |                               |                         |                        | Br, Lu, NHL, Or, Pa, Th Te |
| rs9270747       | 6p21.32          |                               |                               |                         |                        | Ce, Lu, NHL, Re, Th |
| rs28807379      | 8q24.21          |                               |                               |                         |                        | Or, Pr Br, NHL, Th |
| rs45631663      | 8q24.21          |                               |                               |                         |                        | Ce, Lu, NHL, Re, Th |
| rs78035626      | 8q24.21          |                               |                               |                         |                        | Ce, Lu, NHL, Re, Th |

**Table 5 Continued**

| SNP             | Locus            | Overall Increasing OR (95% CI) | Overall Decreasing OR (95% CI) | Increasing OR (95% CI) | Decreasing OR (95% CI) | Associated cancer sites |
|-----------------|------------------|-------------------------------|-------------------------------|-------------------------|------------------------|------------------------|
| rs17190106      | 6p21.33          |                               |                               |                         |                        | Br, Lu, NHL, Or, Pa, Th Te |

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. 28 and 33) (AL2CR12). Interestingly, our lead 1q32 variant (rs1398148) maps to PIK3CB and is in LD ($r^2 = 0.60$) with known MDM4 cancer risk variants48, 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 cells50,51. 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 PAXSIN2, which codes for a cyclin D1 binding partner that serves as a brake for CCND1-mediated cellular migration52.
This is consistent with our observation that the risk-increasing G-allele is associated with increased PACSIN2 expression in whole blood\textsuperscript{33}. 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 phenotype and analyses of stomach and esophageal tumors suggest that they have many overlapping molecular features. 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-European populations. Finally, the two genotyping arrays—those of 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 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 pleotropic 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 samples. GERA participants were drawn from the Kaiser Permanente Northern California (KPNC) health plan members who provided a saliva sample for the Research Program on Genes, Environment and Health (RPGHE) between 2008 and 2011. Individuals included in this study were selected from the 102,977 RPGHE participants who were successfully genotyped as part of GERA and answered a baseline survey concerning lifestyle and medical history. GPA participants were recruited from adult Kaiser Permanente Northern California (KPNC) health plan members who provided a saliva sample for the Research Program on Genes, Environment and Health (RPGHE) between 2008 and 2011. Individuals included in this study were selected from the 102,977 RPGHE participants who were successfully genotyped as part of GERA and answered a baseline survey concerning lifestyle and medical history.

Cancer cases in the UKB were identified via linkage to various national cancer registries established in the early 1970s. Data on 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 record 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 reclassification algorithm. We grouped all esophageal and stomach cancers, 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 cancers of interest were censored. 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). The former is an updated version of the latter, such that the two arrays share over 95% of their content. UKB investigators undertook a rigorous quality control (QC) protocol. 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 SNPs, which is used for all analyses. Ancestry principal components (PCs) were calculated using PCA based on a set of 407,219 unrelated samples and 147,604 genetic markers.

For GERA participants, genotyping was performed using an Affymetrix Axiom array (Affymetrix, Santa Clara, CA, USA) optimized for individuals of European ancestry. To define the reference panel from 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 package.

For both cohorts, analyses were limited to self-reported European ancestry individuals for whom self-reported and genetic sex matched. To further minimize population stratification, we excluded individuals for whom the first two ancestry PCs fell outside five standard deviations from the mean of the population. Based on a subset of genotyped autosomal variants with minor allele frequency (MAF) > 0.01 and genotype call rate ≥ 0.99, 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 10,703 GERA participants. We included SNPs with minor allele frequency (MAF) > 0.01 and genotype call rate ≥ 0.97%, we excluded samples with MAF < 0.01, and/or genotype call rate ≥ 0.99, leaving 8,876,519 variants for analysis for the UKB and 8,973,631 for GERA.

For the UKB, models, the R2INFO score had 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 1000 Genomes Project genotyped array.

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/cancer) to 13,903 (breast cancer) in GERA. For the UKB, models, the R^2INFO score had 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 1000 Genomes Project 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/cancer) to 13,903 (breast cancer) in GERA. For the UKB, models, the R^2INFO score had 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 1000 Genomes Project 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).
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 suggestive 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 clumping and index SNP (COHO) method33 on the aforementioned downsampled subset of UKB participants as a reference panel, performing stepwise selection of the index SNP within a \( \pm 100 \) kb region of one another. SNPs were deemed independent if they maintained a \( P < 1 \times 10^{-8} \) 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 \( LD^2 < 0.1 \) with all previously reported SNPs for the relevant cancer based on both the UKB reference panel and the 10,000 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 \( P < 1 \times 10^{-8} \) 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 Phenoscan database to search for SNPs of interest and variants in LD in order to comprehensively scan previously reported associations. We 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 \( P < 1 \times 10^{-8} \) for replication (defined as having both the direction of effect and \( P < 0.05 \)). Test 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 (poorly or undifferentiated); stage 0–1 (in situ or localized) or stage 2–7 (regional or distant metastases); age < median or age ≥ 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 suggestive and associated in the analysis using prevalent and incident cases. We tested cases (\( P < 1 \times 10^{-8} \)) 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 examined 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 \( P < 1 \times 10^{-4} \) 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 decompose cases into one of two categories: grade 1 (well or moderately differentiated) or grade 3 (poorly or undifferentiated); stage 0–1 (in situ or localized) or stage 2–7 (regional or distant metastases); age < median or age ≥ 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 suggestive and associated in the analysis using prevalent and incident cases. We tested cases (\( P < 1 \times 10^{-8} \)) 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 examined 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 \( P < 1 \times 10^{-4} \) to allow for a sufficient number of variants for comparison.

**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)48. 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 statistic48. ASSET has two procedures: in one, all traits are assumed to be associated with a variant in the same direction (one-directional pleiotropy); in the other, variants can be associated with traits in opposite directions (bidirectional pleiotropy)49. 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 incurred by iterating through all possible trait subsets for each variant as well as controlling for shared samples among the traits48.

Genome-wide ASSET analyses were conducted on the union sets of summary statistics for all 18 cancers. Independent variables were determined via LD clumping, where index SNPs were suggestively significant (overall \( P < 1 \times 10^{-6} \)), and SNPs were added to a region if they were in LD with the lead variant (\( r^2 > 0.05 \), 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 <\( 1 \times 10^{-8} \) and it was associated with at least two cancers. A SNP was determined to have a bidirectional pleiotropic association if the overall \( P \) was <\( 1 \times 10^{-8} \) 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 considered to have pleiotropy. 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).

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 100 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 deleteriousness46. 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)46.

To assess more specific functional features, we annotated each SNP according to Roadmap’s 15-core chromatin states across 127 cell or tissue types52,53. Chromatin state was assigned by taking the most common state, with values >7 indicating open/accessible chromatin regions. Once possible functional annotations were explored to identify significant interaction and enhancer-promoter links. We also explored collaborations 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 GWAS variants from the same regions 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) database with an \( FDR < 0.05 \) 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.
Received: 2 January 2020; Accepted: 13 August 2020; Published online: 04 September 2020

References

1. 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).
2. Mucci, L. A. et al. Familial risk and heritability of cancer among twins in Nordic countries. JAMA 315, 68–76 (2016).
3. 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).
4. 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).
5. Lindström, S. et al. Quantifying the genetic correlation between multiple cancer types. Cancer Epidemiol. Biomark. Prev. 26, 1427–1435 (2017).
6. Jiang, X. et al. Shared heritability and functional enrichment across six solid cancers. Nat. Commun. 10, 431 (2019).
7. 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).
8. Eeles, R. A. et al. Identification of 23 new prostate cancer susceptibility loci using fine-mapping custom genotyping arrays. Nat. Genet. 45, 385–391, 391e2 (2013).
9. Barrett, J. H. et al. Genome-wide association study identifies three new melanoma susceptibility loci. Nat. Genet. 43, 1108–1113 (2011).
10. Brooks, 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).
11. Ellinghaus, E. et al. Identification of germline susceptibility loci in ETVel-RUNX1-rearranged childhood acute lymphoblastic leukemia. Leukemia 26, 902–909 (2012).
12. Rothman, N. et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat. Genet. 42, 978–984 (2010).
13. 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).
14. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. 45, 353–361 (2013).
15. Bahrami, A. et al. Genetic susceptibility in cervical cancer: from bench to bedside. J. Cell Physiol. 233, 1929–1939 (2017).
16. 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).
17. 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).
18. Eeles, R. A. et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat. Genet. 40, 316–321 (2008).
19. Purdie, M. P. et al. Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p12 and 11q13.3. Nat. Genet. 43, 60–65 (2011).
20. Spurde, A. B. et al. Genome-wide association study identifies a common variant associated with risk of endometrial cancer. Nat. Genet. 43, 451–454 (2011).
21. Couch, F. J. et al. Common variants at the 19p13.1 and 20p13.3 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).
22. Setaiwan, V. W. et al. Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. Carcinogenesis 35, 2068–2073 (2014).
23. Rafaer, T. et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat. Genet. 41, 221–227 (2009).
24. 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).
25. Jones, C. C. et al. Cross-cancer pleiotropic associations with lung cancer risk in African Americans. Cancer Epidemiol. Biomark. Prev. 28, 715–723 (2019).
26. 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).
27. 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).
28. 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).
29. Toth, R. et al. Genetic variants in epigenetic pathways and risks of multiple cancers in the GAME-ON consortium. Cancer Epidemiol. Prev. Biomark. 26, 816–825 (2017).
30. Karam, S. et al. Telomere structure and maintenance gene variants and risk of five cancer types. Int. J. Cancer 139, 2655–2670 (2016).
31. Broderick, P. et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat. Genet. 39, 1313–1317 (2007).
32. Kote-Jarai, Z. et al. Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. Hum. Genet. 129, 687–694 (2011).
33. 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).
34. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence models. Nat. Methods 12, 931–934 (2015).
35. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. 8, 1826 (2017).
36. Rentzsch, 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).
37. Han, J. et al. Expression quantitative trait loci in non-long coding RNA PAX8-AS1 are associated with decreased risk of cervical cancer. Mol. Genet. Genomics 291, 1743–1748 (2016).
38. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. 104, 65–75 (2019).
39. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92–94 (2017).
40. Phelan, C. M. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat. Genet. 49, 680–691 (2017).
41. Graff, R. et al. Familial risk and heritability of colorectal cancer in the Nordic Twin Study of Cancer. Clin. Gastroenterol. Hepatol. 15, 1256–1264 (2017).
42. Hemminki, K. & Chen, B. Familial risks in testicular cancer as aetiological clues. Int. J. Androl. 29, 205–210 (2006).
43. Zhang, L. et al. Familial associations in testicular cancer with other cancers. Sci. Rep. 8, 10880 (2018).
44. 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).
45. Ferreiro-Iglesias, A. et al. Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. Nat. Commun. 9, 3927 (2018).
46. Marty, R. et al. MHC-I genotype restricts the oncogenic mutational landscape. Cell 171, 1272–1283.e15 (2017).
47. Marty Pyke, R. et al. Evolutionary pressure against MHC class II binding cancer mutations. Cell 175, 416–428.e13 (2018).
48. Grisanzio, C. & Freedman, M. L. Chromosome 8q24-associated cancers and lung cancer mutations. Proc. Natl Acad. Sci. USA 111, E4706–E4715 (2014).
54. Okines, A. F. C. et al. Biomarker analysis in oesophagogastric cancer: results from the REAL3 and TransMAGIC trials. Eur. J. Cancer. 49, 2116–2125 (2013).
55. Barra, W. F. et al. GEJ cancers: gastric or esophageal tumours? 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. Recófi, 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_dwhome/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. Howlader, 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 UKR 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 (R22 AG036607). We thank the Breast Cancer Association Consortium (BCAC) for breast cancer summary statistics (http://bcac.ccge.medschl.cam.ac.uk/bcacdata/uncordarray/gwas-icogs-and-oncarray-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 (PRAGICAL) 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 UKitobank 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.I.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.