Pan-Cancer Study Detects Novel Genetic Risk Variants and Shared Genetic Basis in Two Large Cohorts

Sara R. Rashkin,[1*] Rebecca E. Graff,[1,2*] Linda Kachuri,[1] Khanh K. Thai,[2] Stacey E. Alexeeff,[2] Maruta A. Blatchins,[2] Taylor B. Cavazos,[1,3] Douglas A. Corley,[2] Nima C. Emami,[1,3] Joshua D. Hoffman,[1] Eric Jorgenson,[2] Lawrence H. Kushi,[2] Travis J. Meyers,[1] Stephen K. Van Den Eeden,[2,4] Elad Ziv,[5,6,7] Laurel A. Habel,[2] Thomas J. Hoffmann,[1,2,5] Lori C. Sakoda,[2**] John S. Witte[1,4,5,7**]

[*] These authors contributed equally to this work
[**] Co-senior / corresponding authors

Affiliations:
[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

Corresponding Authors:
John S. Witte
Department of Epidemiology and Biostatistics
University of California, San Francisco
1450 3rd Street, Room 388, San Francisco, CA 94158
Phone: (415) 502-6882
Email: jwitte@ucsf.edu

Lori C. Sakoda
Division of Research
Kaiser Permanente Northern California
2000 Broadway, Oakland, CA 94612
Phone: (510) 891-3677
Email: lori.sakoda@kp.org

Abstract

Deciphering the shared genetic basis of distinct cancers has the potential to elucidate carcinogenic mechanisms and inform broadly applicable risk assessment efforts. However, no studies have investigated pan-cancer pleiotropy within single, well-defined populations unselected for phenotype. We undertook novel 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 (413,870 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 detected 21 novel genome-wide significant risk variants. In addition, numerous cancer sites exhibited clear heritability. Investigations of pleiotropy identified 12 cancer pairs exhibiting either positive or negative genetic correlations and 43 pleiotropic loci. We identified 158 pleiotropic variants, many of which were enriched for regulatory elements and influenced cross-tissue gene expression. Our findings demonstrate widespread pleiotropy and offer further insight into the complex genetic architecture of cross-cancer susceptibility.

Introduction

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.[1] 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,[2] families,[3] 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-CLPTM1L*)[9,12]; 6p21 (HLA complex)[15,16]; 7p15[17]; 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 conducted.

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 cancer and both lung and pancreatic cancers.[5] 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 cancers.[6] 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 with respect to 18 cancer types, based on 475,312 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 sought to detect novel risk SNPs and pleiotropic loci and variants and to estimate the heritability of and genetic correlations between cancer types. We then conducted *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 allowed 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.

Results

Genome-wide Association Analyses of Individual Cancers

We found 21 novel, independent variants that attained genome-wide significance at $P<5x10^{-8}$ upon meta-analysis of the UKB and GERA results (**Table 1**) and an additional 9 genome-wide significant variants that were only genotyped or imputed in one cohort (**Supplementary Table 1**). In addition, we detected 308 independent signals with $P<1x10^{-6}$ that confirmed risk SNPs identified by previous GWAS with $P<5x10^{-8}$ (**Supplementary Table 2**). Of the 21 novel genome-wide significant associations from the meta-analyses, 9 were in known susceptibility regions for the cancer of interest but independent of previously reported variants. The remaining 12 were in regions not previously associated with the cancer of interest in individuals of European ancestry. Fourteen of the 21 novel SNPs exhibited pleiotropy in that they were in regions previously associated with at least one of the other cancer types evaluated in this study.

In sensitivity analyses for the 21 novel SNPs, we did not detect any material differences in effect estimates across categories of age at diagnosis, Surveillance, Epidemiology, and End Results Program (SEER) grade, or SEER stage (Heterogeneity P>0.05, corrected for multiple testing). In genome-wide sensitivity analyses restricted to incident cases (i.e., excluding prevalent cases), our findings were essentially unchanged (Heterogeneity P>0.05, corrected for multiple testing; **Supplementary Figure 1**). Similarly, genome-wide sensitivity analysis results for esophageal and stomach cancers separately were materially comparable to those for the two phenotypes combined (Heterogeneity P>0.05, corrected for multiple testing; **Supplementary Figure 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] For example, our estimate for testicular cancer closely matches the previous family-based estimate of heritability (h^2=0.25; 95% CI: 0.15-0.37),[3] as well as a previous estimate of array-based heritability (h^2=0.30; 95% CI: 0.08-0.51).[4] For lung cancer, our estimated heritability of h^2=0.15 (95% CI: 0.10-0.20) approaches the twin-based estimate of h^2=0.18 (95% CI: 0.00-0.42),[2] exceeds the array-based estimate from a study using the same methodology (h^2=0.08; 95% CI: 0.05-0.10),[6] and is comparable to an earlier array-based estimate using individual-level data (h^2=0.21; 95% CI: 0.14-0.27).[4] For rectal (h^2=0.11; 95% CI: 0.07-0.16) and bladder (h^2=0.08; 95% CI: 0.04-0.12) cancers, our heritability estimates are close to those from twin/family studies – h^2=0.14 (95%CI: 0.00-0.50)[2] and h^2=0.07 (95% CI: 0.02-0.11),[3] respectively. 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=5.33x10^{-7}) were genetically correlated at a Bonferroni corrected significance threshold of P=0.05/153=3.27x10^{-4} (**Figure 1a-b; Supplementary Table 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 43 pleiotropic regions associated with more than one cancer ($P<1x10^{-6}$ for each cancer; regions defined using our linkage disequilibrium [LD] clumping procedure; see Methods; **Figure 1c; Supplementary Table 4**). Most were at known cancer pleiotropic loci: HLA (24 regions), 8q24 (10 regions), *TERT-CLPTM1L* (5 regions), and *TP53* (1 region). Twenty-two of the HLA regions were associated with both cervical cancer and NHL, and the remaining two were associated with (1) cervical and prostate cancers and (2) NHL and prostate cancer. Six regions in 8q24 were associated with prostate and colon cancers, and four were associated with prostate and breast cancers. Of the regions in *TERT-CLPTM1L*, two were associated with prostate cancer, one with breast cancer and one with testicular cancer; two were associated with melanoma, one with breast cancer and one with bladder cancer; and the last was associated with melanoma and cervical, lung, and pancreatic cancers. The *TP53* region, indexed by rs78378222, was associated with melanoma and lymphocytic leukemia. The remaining three pleiotropic regions were in loci previously associated with at least one cancer and were indexed by rs772695095 (*DIRC3* at 2q35; breast and thyroid cancers), rs11813268 (intergenic at 10q24.33; melanoma and prostate cancer), and rs6507874 (*SMAD7* at 18q21.1; colon and rectal cancers).

Genome-wide Variant-Specific Pleiotropy

We assessed variant-specific pleiotropy by testing all variants genome-wide using the summary statistics for each cancer. We found 137 independent one-directional pleiotropic variants with at least two associated cancers, the same direction of effect for all associated cancers, and an overall pleiotropic $P<1x10^{-6}$ (**Supplementary Table 5**), among which 85 attained genome-wide significance ($P<5x10^{-8}$). Of the 137 one-directional pleiotropic variants, there were 45 for which the overall pleiotropic *P* was smaller than the *P* for each of the associated cancers, 17 of which attained genome-wide significance (**Figure 2**). While 134 of the 137 one-directional pleiotropic variants were in regions that have previously been associated with cancer, 113 were associated with at least one new 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 21 such variants with $P<1x10^{-6}$, all of which were independent from one another and from the one-directional pleiotropic variants (**Figure 3; Supplementary Table 6**). Fifteen attained genome-wide significance. There were 13 variants where the overall pleiotropic *P* was smaller than the *P* for the associated cancers, eight of which attained genome-wide significance. While 20 of the 21 bidirectional pleiotropic variants were in regions that have previously been associated with cancer, 10 were independent of known risk variants, and all 20 were associated with at least one new cancer. The SNP in a novel region (*SSPN* at 12q12.1) was rs10842692, which was associated with five cancers in one direction and nine cancers in the other.

The number of one- and bidirectional SNPs shared by cancer pairs ranged from four (bladder and esophagus/stomach; colon and NHL) to 32 (pancreas and prostate; lymphocytic leukemia and prostate) (**Figure 4a; Supplementary Table 7**). For 19 cancer pairs, the shared associations had exclusively the same direction of effect (tabulating across both the one- and bidirectional analyses). For eight cancer pairs, most of the shared variants were associated in opposite directions.

For each of the 158 SNPs showing either one- or bidirectional pleiotropy, 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 multiple testing, only one 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 screen-detected prostate cancer.[31,32]

Functional Characterization of Pleiotropic Variants

The biological significance of 158 pleiotropic variants was evaluated using *in-silico* annotation tools (**Supplementary Table 8**).[33–35] Pleiotropic variants were enriched in intergenic and exonic regions, as well as non-coding RNA transcripts (P<0.018) (**Figure 4b)**. The distribution of DeepSea functional significance scores was skewed toward 0, indicating a higher likelihood of regulatory effects compared to a reference distribution of 1000 Genomes variants (**Figure 4b**). Suggestively functional variants (n=38, DeepSEA score<0.05) were also predicted to be

pathogenic by Combined Annotation-Dependent Depletion (CADD; median score of 10.99, corresponding to the top 10% of deleterious substitutions). The top-ranked variant was rs3862792 in 11q13.3 (*CCND1*; synonymous; CADD=19.55), associated with prostate cancer and lymphocytic leukemia. Forty-two of the 158 pleiotropic variants were characterized by active chromatin states, 51 were classified as enhancers, and 83 had significant (FDR<0.05) effects on gene expression (**Figure 4b**). Nine variants belonged to all three classes (**Figure 4b)**, including rs10842692, the lead variant in one of our novel pleiotropic regions (*SSPN* at 12q12.1).

Consistent with hypothesized pleiotropy, 75.9% of the 83 expression quantitative trait loci (eQTLs) identified among the pleiotropic variants had more than one target tissue, and 71.1% influenced the expression of more than one gene (**Supplementary Figure 3**), for a total of 586 significant SNP-gene pairs. The most common expression tissues for eQTLs among pleiotropic variants were whole blood (63%), followed by subcutaneous adipose (9.9%), and esophageal (5.3%). Regulatory effects mediated by chromatin looping were observed for 46 variants, which clustered in the HLA region (10 variants), 8q24 (3 variants), and 11q13.3 (3 variants; **Supplementary Figure 4**). Three enhancer-promoter links were also identified in 6p21.23 (rs535777, rs73728618) and 22q13.2 (rs5759167, *PACSIN2* promoter; **Supplementary Figure 4**). Notably, rs5759167 is also an eQTL for *PACSIN2* in whole blood (P=9.89x10^{-14}).

Genes represented by pleiotropic variants were significantly enriched for 67 pathways (**Supplementary Table 9**). Top-ranking pathways included inflammatory and immune response mechanisms (cytokines and chemokine targets of PSMD4: p=3.06x10^{-8}; interferon-γ signaling in endothelial cells: p=1.16x10^{-6}; antigen processing and presentation: p=2.53x10^{-6}) and cell cycle

checkpoint controls (G1/S: p=6.39x10^{-6}). There was also enrichment for PML targets with promoters bound by Myc (p=3.27x10^{-7}), transcription-factor networks implicated in neoplastic transformation, including FOXM1 (p=2.74x10^{-5}), and p53 signaling (p=2.35x10^{-4}).

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 novel insights into cancer susceptibility. Specifically, we detected 21 novel genome-wide significant risk variants across the 18 individual cancers. We also detected 158 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.[15] 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.[36] Several GWAS findings also provided evidence of pleiotropy, in that novel risk variants for one cancer had been previously associated with one or more other cancers. For instance, rs9818780 was associated with melanoma and has been implicated in sunburn risk.[37] This intergenic variant is an eQTL for *LINC00886* and *METTL15P1* in skin tissue. The former gene has previously been linked to breast cancer,[38] and both genes have been implicated in ovarian cancer.[39] Beyond novel

associations, our GWAS detected 308 independent associations with $P<1x10^{-6}$ that confirmed signals identified in previous GWAS with $P<5x10^{-8}$. 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.[40] We also identified several novel 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.[41,42] 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 previous consortia-based 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. 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 to examine several cancers that have not been previously studied in large consortia.

The assessment of pleiotropy at the locus level confirmed previously reported associations at 5p15.33, HLA, and 8q24.[9,12,15,16,18] Out of the 43 pleiotropic loci that we identified, over half, all in 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 Table 10**). The difference may reflect better coverage and imputation of the HLA region in the UKB than in GERA. Other findings support a pleiotropic role of several loci previously associated with specific cancers in separate studies. For example, we validated previous results showing that *DIRC3* is associated with breast and thyroid cancers.[14,43] Additionally, the intergenic region surrounding rs11813268 on 10q24.33 has not been previously associated with melanoma or prostate cancer (as it was in our study), although associations with other cancers have been reported, including kidney, lung, and thyroid cancers.[39,44–47] *SMAD7* has been previously linked to colorectal cancer,[48] and we confirmed its association with colon and rectal cancers separately.

Variant-specific analyses provided further evidence in support of locus-specific cancer pleiotropy, including validation of previously reported signals at 1q32[7,8] and 2q33[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 158 pleiotropic variants (with overall pleiotropic $P<1x10^{-6}$) mapped to a total of 78 genomic locations, which included all of the regions identified from the locus-specific analyses. Although 154 of the 158 variants showing one- or bidirectional pleiotropic associations are in regions previously associated with cancer, 133 of the 154 were associated with at least one new cancer. One variant (rs10842692) associated with 14 cancers in the bidirectional analysis mapped to a novel region at 12p12.1 and is an active enhancer and an eQTL for *SSPN* in multiple tissue types, including adipose. *SSPN* has been linked to waist circumference,[49] suggesting that increased adiposity may be one plausible mechanism underlying the pleiotropic associations observed for this locus.

Out of 158 total variants identified from the variant-specific pleiotropy analyses, 20 were in 8q24 and 19 were in the HLA region. Different distributions of one- and bidirectional results highlight patterns of directional pleiotropy: of the 19 HLA variants, eight were bidirectional, while only three of the 20 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 NHL[50] and lung cancer,[51] 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[52] and class II alleles.[53] 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 marks.[54] Consistent with this hypothesis, we observed heritability enrichment[55] for variants with the H3K27ac annotation for breast (P = 3.09x10^{-4}), colon (P = 4.44x10^{-4}), prostate (P = 2.74x10^{-5}), and rectal (P = 0.036) cancers – all of which share susceptibility variants in 8q24, according to our analyses and previous studies.[54]

In-silico analyses found the 158 pleiotropic variants to be enriched across multiple regulatory domains and highlighted functional features relevant to cancer pleiotropy. The 11q13.3 region includes rs3862792, which is predicted to be in the top 1% of deleterious substitutions in the genome.[35] Although rs3862792 has been previously linked to prostate cancer,[56] our results suggest it may also be relevant for lymphocytic leukemia. *CCND1* is a hallmark cancer oncogene that plays a role in cell cycle transitions, cell invasion, and cell migration, making rs3862792 a highly plausible candidate for cross-cancer effects. Three pleiotropic variants in 11q13.3 mapped to enhancer regions, including an eQTL for *TPCN2*, which is part of a signaling pathway controlling the angiogenic response to VEGF.[57] The 22q13.2 region is indexed by rs5759167, an intergenic variant linked to prostate and lung cancers in our analysis. 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 migration.[58] This is consistent with our observation that that the risk-increasing G-allele was associated with increased *PACSIN2* expression in whole blood.[59] Lastly,

our pathway analysis indicated that pleiotropic variants as a group are enriched for canonical signaling pathways that control cell-cycle progression, apoptosis, and immune-related functions, the dysregulation of which is a hallmark of cancer.

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 some differences in their risk factor profiles. However, there is precedent for using a composite phenotype,[60] and analyses of stomach and esophageal tumors suggest that they have many overlapping molecular features.[61,62] In addition, sensitivity analyses for each cancer alone gave similar results, suggesting that they may have similar genetic bases despite having different environmental risk factors. Finally, we focused solely on individuals of European ancestry. Further analyses are needed to accurately assess patterns of pleiotropy in non-Europeans.

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 unprecedented 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 to 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.[63] 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.[64,65]

Cancer cases in the UKB were identified via linkage to various national cancer registries established in the early 1970s.[63] 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 paradigm.[66] 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. 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).[63] 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) protocol.[63] 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 supplementary data.[63] Ancestry principal components (PCs) were computed using *fastPCA*[67] based on a set of 407,219 unrelated samples and 147,604 genetic markers.[63]

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.[65,68] 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,[69] IMPUTE2 v2.3.1 was used to impute SNPs relative to the cosmopolitan reference panel from 1000 Genomes.[70–72] Ancestry PCs were computed using Eigenstrat v4.2, as previously described.[64]

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 ≥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[73] 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 413,870 UKB participants and 66,526 GERA participants. We excluded SNPs with imputation quality score <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 $<1x10^{-5}$, and/or MAF <0.01, leaving 8,876,519 variants for analysis for the UKB and 8,973,631 for GERA.

Genome-Wide Association Analyses of Individual Cancers

We used PLINK[74,75] 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 3,978 (breast cancer) in GERA (**Supplementary Table 11**). Control counts were 359,825 (189,855 female) and 50,525 (29,801 female) 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 Figure 5**).

To determine independent signals in our union set of SNPs, we implemented the LD clumping procedure in PLINK[74,75] 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. To identify all potential signals, in each clump, index SNPs had a suggestive association based on $P<1x10^{-6}$, and SNPs were added if they were marginally significant with $P<0.05$, were within

500kb 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,[76,77] performing stepwise selection of the index SNPs within a +/- 1000kb region of one another. SNPs were deemed independent if they maintained $P<1x10^{-6}$ 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.[78] As an additional filter for novelty, we again used COJO[76,77] 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 $P<1x10^{-6}$ 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[79] 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],[38] prostate [~80,000 cases],[80] and ovarian [~30,000

cases][39]), we tested our potentially novel SNPs with $P<1x10^{-6}$ for replication (defined as having the same direction of effect and $P<0.05$). Tested SNPs that did not replicate were not considered novel.

We considered whether clinical characteristics of the cases were informative about associated phenotypes by examining SEER stage and grade (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 ≥ median. The case counts for all cancer-outcome strata are tabulated in **Supplementary Table 12**. For each of the novel 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 ($P<1x10^{-6}$) 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 $P<1x10^{-6}$ 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.

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.[81,82] 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 13**).[83] 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 $P<1x10^{-6}$. Non-overlapping regions were iteratively formed around index SNPs associated with any cancer, beginning with the SNP associated with the smallest *P*. We used a suggestive threshold of $P<1x10^{-6}$ to assess whether suggestive regions for one cancer might also be informative for another. SNPs were added to a region if they were associated with any cancer with $P<1x10^{-6}$, were within 500kb 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 (*as*sociation analysis based on sub*sets*).[84] 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 statistic.[84] 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).[84] 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 traits.[84]

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 $P<1x10^{-6}$), and other SNPs were clumped with the lead variant if they had overall $P<0.05$, were within 500kb of the index SNP, and had $r^2>0.1$ with the index SNP. 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 $<1x10^{-6}$ and it was associated with at least two cancers. A SNP was determined to have a

bidirectional pleiotropic association if the overall *P* was $<1x10^{-6}$ 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 158 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 in the reference panel of UKB European descent individuals.

Overall functional significance was assessed using DeepSEA,[33] 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 deleteriousness.[35] 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).[35]

To assess more specific functional features, we annotated each SNP according to Roadmap's 15-core chromatin states across 127 cell or tissue types.[34,85] Chromatin state was assigned by taking the most common state across 127 cell or tissue types, with values ≤7 indicating open, accessible chromatin regions. Three-dimensional chromatin interactions were explored to identify significant interaction and enhancer-promoter links. Lastly, we explored associations with gene expression in blood and non-neurological tissues (since we did not investigate brain tumors) using data from the GTEx v7[86] and BIOS QTL[59] databases.

We conducted pathway analyses based on gene set collections from the Molecular Signatures Database (MSigDB)[87] with an FDR $q<0.05$ significance threshold. Pathway enrichment analyses focused on the C2 MSigDB collection, which includes curated signatures of genetic and chemical perturbations, as well as canonical pathways from BioCarta, KEGG, Reactome, and other databases.

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.

Data Availability

Our GWAS summary statistics are publicly available via direct request and will be deposited into dbGaP. The UKB cohort data is publicly available from the UKB access portal at https://www.ukbiobank.ac.uk. The UKB cancer phenotyping we performed to define cases will be provided to the UKB for public use. The Kaiser Permanente data are available via application with a local collaborator at: https://researchbank.kaiserpermanente.org/our-research/for-researchers/.

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).

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., 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.

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**, (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 the iCOGS custom genotyping array. *Nat. Genet.* **45**, 385–391, 391e1-2 (2013).
9. Barrett, J. H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.* **43**, 1108–1113 (2011).

10. 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).

11. Ellinghaus, E. *et al.* Identification of germline susceptibility loci in ETV6-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. 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).

20. 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).

21. 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).

22. Setiawan, V. W. *et al.* Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. *Carcinogenesis* **35**, 2068–2073 (2014).

23. Rafnar, 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**, (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. Biomarkers Prev.* **26**, 816–825 (2017).

30. Karami, S. *et al.* Telomere structure and maintenance gene variants and risk of five cancer types. *Int. J. Cancer* **139**, 2655–2670 (2016).

31. Kote-Jarai, Z. *et al.* Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. *Hum. Genet.* **129**, 687–694 (2011).

32. 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).

33. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

34. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

35. 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).

36. 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).

37. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

38. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).

39. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).

40. 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).

41. Hemminki, K. & Chen, B. Familial risks in testicular cancer as aetiological clues. *Int. J. Androl.* **29**, 205–210 (2006).

42. Zhang, L. *et al.* Familial Associations in Testicular Cancer with Other Cancers. *Sci. Rep.* **8**, 10880 (2018).

43. Gudmundsson, J. *et al.* Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nat. Genet.* **44**, 319–322 (2012).

44. Scelo, G. *et al.* Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* **8**, (2017).

45. Chahal, H. S. *et al.* Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nat. Commun.* **7**, (2016).

46. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).

47. Gudmundsson, J. *et al.* A genome-wide association study yields five novel thyroid cancer risk loci. *Nat. Commun.* **8**, 14517 (2017).

48. 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).

49. Graff, M. *et al.* Genome-wide physical activity interactions in adiposity - A meta-analysis of 200,452 adults. *PLoS Genet.* **13**, e1006528 (2017).

50. 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).

51. Ferreiro-Iglesias, A. *et al.* Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat. Commun.* **9**, 3927 (2018).

52. Marty, R. *et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **171**, 1272-1283.e15 (2017).

53. Marty Pyke, R. *et al.* Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* **175**, 416-428.e13 (2018).

54. Grisanzio, C. & Freedman, M. L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555–559 (2010).

55. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

56. Amin Al Olama, A. *et al.* Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. *Hum. Mol. Genet.* **24**, 5589–5602 (2015).

57. Favia, A. *et al.* VEGF-induced neoangiogenesis is mediated by NAADP and two-pore channel-2-dependent Ca2+ signaling. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4706-4715 (2014).

58. 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 Georget. Tex* **10**, 73–81 (2011).

59. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).

60. Okines, A. F. C. *et al.* Biomarker analysis in oesophagogastric cancer: Results from the REAL3 and TransMAGIC trials. *Eur. J. Cancer Oxf. Engl. 1990* **49**, 2116–2125 (2013).

61. Barra, W. F. *et al.* GEJ cancers: gastric or esophageal tumors? searching for the answer according to molecular identity. *Oncotarget* **8**, 104286–104294 (2017).

62. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).

63. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

64. 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).

65. 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).

66. Site Recode ICD-O-3/WHO 2008 Definition. Available from URL: https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html [accessed June 30, 2017].

67. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).

68. 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).

69. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).

70. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5**, e1000529 (2009).

71. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3 Genes Genomes Genet.* **1**, 457–470 (2011).

72. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

73. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

74. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

75. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

76. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

77. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1-3 (2012).

78. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

79. Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

80. 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).

81. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

82. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

83. 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.

84. 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).

85. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

86. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

87. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

Tables and Figures

Table 1. Novel genome-wide significant loci from meta-analysis of UK Biobank (UKB) and Genetic Epidemiology Research on Adult Health and Aging (GERA) SNPs for each cancer site.

						MAF**		Odds Ratio (OR)			
Cancer Site	**SNP**	**Chromosome**	**Position**	**Gene**	**REF/ALT***	**UKB**	**GERA**	**UKB**	**GERA**	**Meta**	**Meta P**
Bladder	rs76088467†,+	6	21795787	*CASC15*	**G**/A	0.025	0.030	0.67	0.59	0.64	$2.34x10^{-8}$
Breast	rs6752414†,+	2	121425339	Intergenic	**T**/C	0.077	0.083	0.89	0.87	0.88	$1.81x10^{-9}$
Breast	rs8027730§,+	15	49872585	*FAM227B*	**A**/C	0.48	0.48	1.06	1.08	1.06	$2.68x10^{-8}$
Cervix	rs10175462§,+	2	113988492	*PAX8*	**A**/G	0.36	0.37	1.16	1.08	1.15	$7.71x10^{-14}$
Cervix	rs2856437†,+	6	32157364	*PBX2*	**A**/G	0.063	0.047	0.76	0.88	0.77	$1.24x10^{-15}$
Colon	rs71518872§,+	8	103561978	Upstream of *ODF1*	**G**/C	0.015	0.017	0.65	0.61	0.64	$1.27x10^{-8}$
Colon	rs8114643§	20	7833046	Intergenic	**G**/A	0.14	0.15	0.83	0.84	0.83	$2.10x10^{-9}$
Esophagus/Stomach	rs75460256†,+	2	106687838	*C2orf40*	**G**/A	0.024	0.022	0.52	0.67	0.53	$1.04x10^{-8}$
Kidney	rs112248293§,+	15	61500352	*RORA*	**A**/G	0.024	0.025	0.53	0.62	0.55	$3.36x10^{-9}$
Lung	rs10863899§	1	211666218	5'UTR of *RD3*	**G**/A	0.42	0.42	1.23	1.09	1.18	$1.91x10^{-8}$
Lung	rs146099759§	5	12883592	Intergenic	**A**/G	0.024	0.028	0.69	0.57	0.64	$3.50x10^{-8}$
Lung	rs12543486†,+	8	13012376	*DLC1*	**C**/T	0.17	0.16	1.30	1.18	1.26	$3.51x10^{-8}$
Lymphocytic Leukemia	rs114490818†	3	126099101	Intergenic	**A**/G	0.022	0.011	0.48	0.53	0.48	$2.86x10^{-8}$
Lymphocytic Leukemia	rs61965473§,+	13	95571786	Intergenic	**T**/C	0.023	0.023	0.52	0.44	0.49	$3.95x10^{-8}$
Lymphocytic Leukemia	rs78378222†,+	17	7571752	3'UTR of *TP53*	**G**/T	0.012	0.014	0.44	0.34	0.40	$1.89x10^{-9}$
Melanoma	rs9818780§,+	3	156492758	Intergenic	**C**/T	0.49	0.48	0.92	0.89	0.91	$3.16x10^{-8}$
Melanoma	rs12186662§	5	90356197	*ADGRV1*	G/**A**	0.32	0.36	0.90	0.89	0.90	$1.09x10^{-8}$
Melanoma	rs55797833†,+	9	21995044	5'UTR OF *CDKN2A*	**G**/T	0.023	0.021	1.71	1.72	1.71	$6.71x10^{-12}$
Melanoma	rs78378222†,+	17	7571752	3'UTR of *TP53*	**G**/T	0.012	0.014	0.70	0.63	0.67	$1.18x10^{-8}$
Rectum	rs145503185§	9	23455764	Intergenic	**C**/T	0.013	0.018	0.57	0.50	0.55	$4.36x10^{-8}$
Thyroid	2:173859846_TA_T§	2	173859846	*RAPGEF4*	T/**TA**	0.25	0.26	1.45	1.15	1.36	$3.49x10^{-8}$

*REF is reference allele and ALT allele is effect allele; bold allele is minor allele
**MAF = minor allele frequency calculated in all controls

§ Indicates SNPs in loci not previously associated with the cancer of interest in European ancestry

† 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)

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

Table 2. Heritability estimates (h^2) and 95% confidence intervals (CIs) for each cancer based on the union set of UK Biobank (UKB) and Genetic Epidemiology Research on Adult Health and Aging (GERA) SNPs, compared with previous estimates from other array-based and twin/family-based studies.

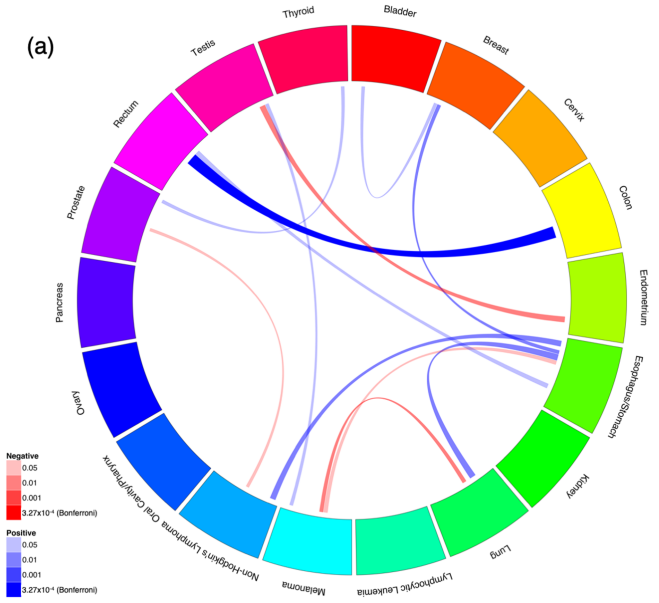
	Array-Based Heritability			Twin/Family-Based Heritability
Cancer Site	**Current Study**	**Jiang *et al.*[a]**	**Sampson *et al.*[c]**	**Mucci *et al.*[d]**
Bladder	0.08 (0.04-0.12)		0.12 (0.09-0.16)	0.07 (0.02-0.11)[e]
Breast	0.10 (0.08-0.13)	0.14 (0.12-0.16)	0.10 (0.00-0.20)[**]	0.31 (0.11-0.51)
Cervix	0.07 (0.02-0.12)			0.13 (0.06-0.15)[e,+]
Colon	0.07 (0.04-0.10)	0.09 (0.07-0.11)[*]		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)[***]	0.22 (0.00-0.55)[++]
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)[†]	0.09 (0.09-0.16)[e,+++]
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)[††]	0.10 (0.08-0.10)[e]
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)[b]	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)[e]
Thyroid	0.21 (0.09-0.33)			0.53 (0.52-0.53)[e]

[*] Colorectal
[**] Estrogen receptor negative (ER-)
[***] For esophageal in Asian population (stomach in Asian population: $h^2 = 0.25$ [0.00-0.52])
[†] For chronic lymphocytic leukemia
[††] For diffuse large B-cell lymphoma
[+] For in situ (invasive: $h^2 = 0.22$ [0.14-0.27])
[++] Stomach
[+++] Age >15 years

[a] Taken from Jiang, et al. (2019),[6] 95% CI calculated from provided standard error
[b] Taken from Lindström, et al. (2017)[5]
[c] Taken from Sampson, et al. (2015)[4]
[d] Taken from Mucci, et al. (2016),[2] except where not included in analysis or 95% CI range was > 0.60; remaining taken from Czene, et al. (2002),[3] as marked
[e] Taken from Czene, et al. (2002),[3] family-based not twin

Figure 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 the point estimate, and shading is proportional to strength of association according to P, where the Bonferroni-corrected threshold is $0.05/153=3.27\text{x}10^{-4}$; (b) genetic correlation, 95% confidence interval (CI), and P for all cancer pairs with $P<0.05$; (c) 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 $P<1\text{x}10^{-6}$ for any cancer and SNPs are added if they have $P<1\text{x}10^{-6}$ for any cancer, are within 500kb of the index SNP, and have LD $r^2>0.5$ with the index SNP.

(a)



(b)

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)	$5.33\text{x}10^{-7}$
Endometrium	Testis	-0.41 (-0.70 – -0.11)	0.0064
Esophagus/Stomach	Lung	0.44 (0.15 – 0.74)	0.0035
Esophagus/Stomach	Melanoma	-0.27 (-0.53 – -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 – -0.08)	0.0048
Melanoma	Testis	0.23 (0.03 – 0.44)	0.028
Non-Hodgkin's Lymphoma	Prostate	-0.21 (-0.37 – -0.05)	0.012
Prostate	Thyroid	0.23 (0.05 – 0.41)	0.013

(c)

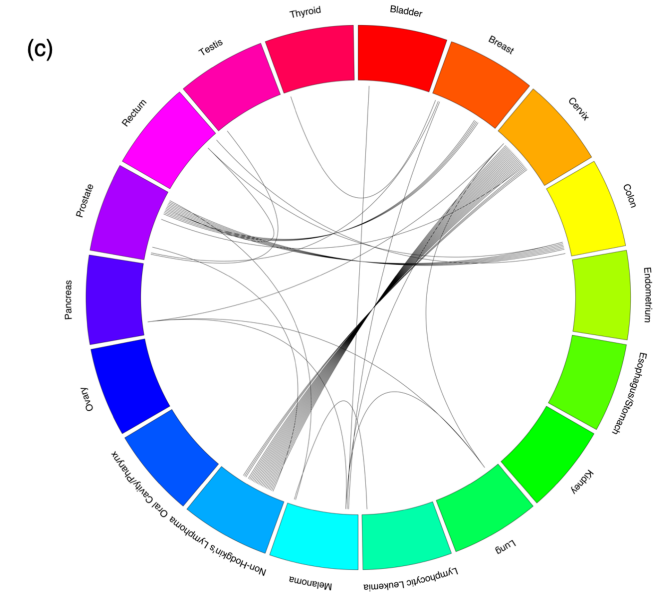
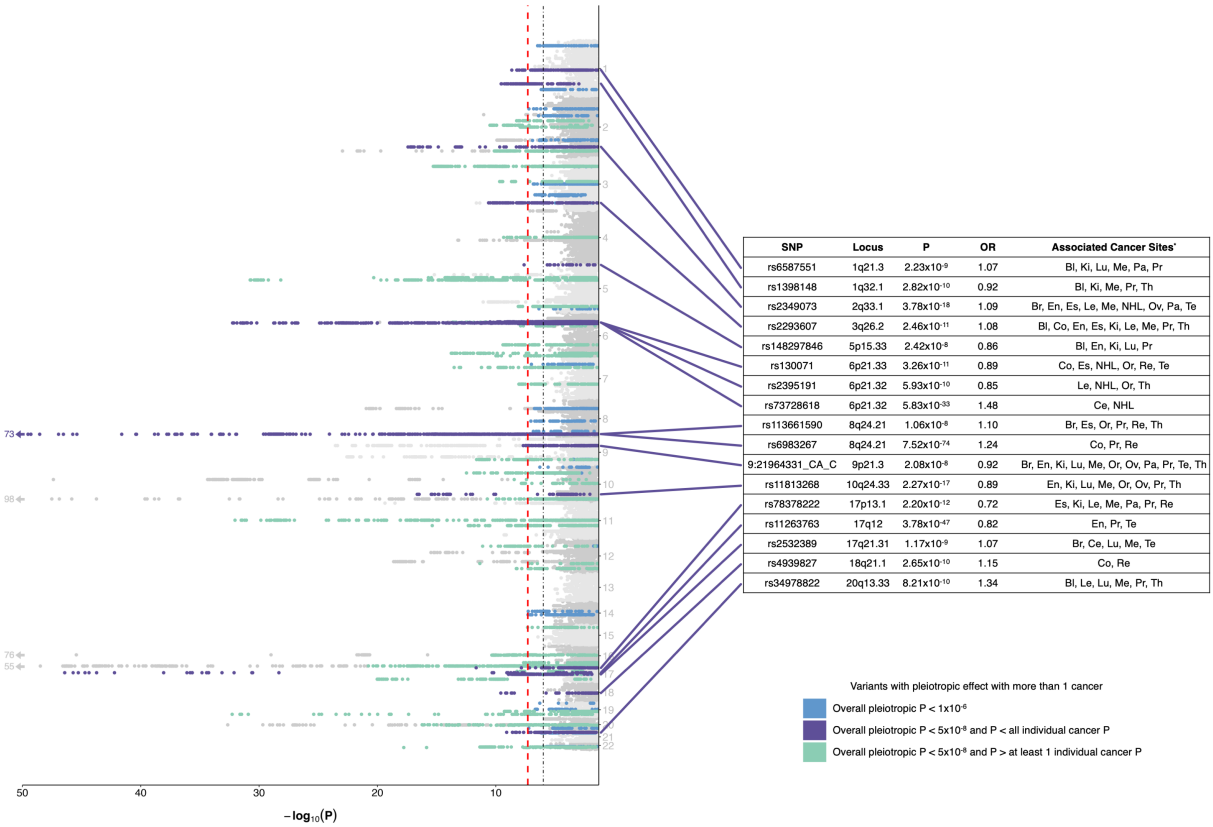


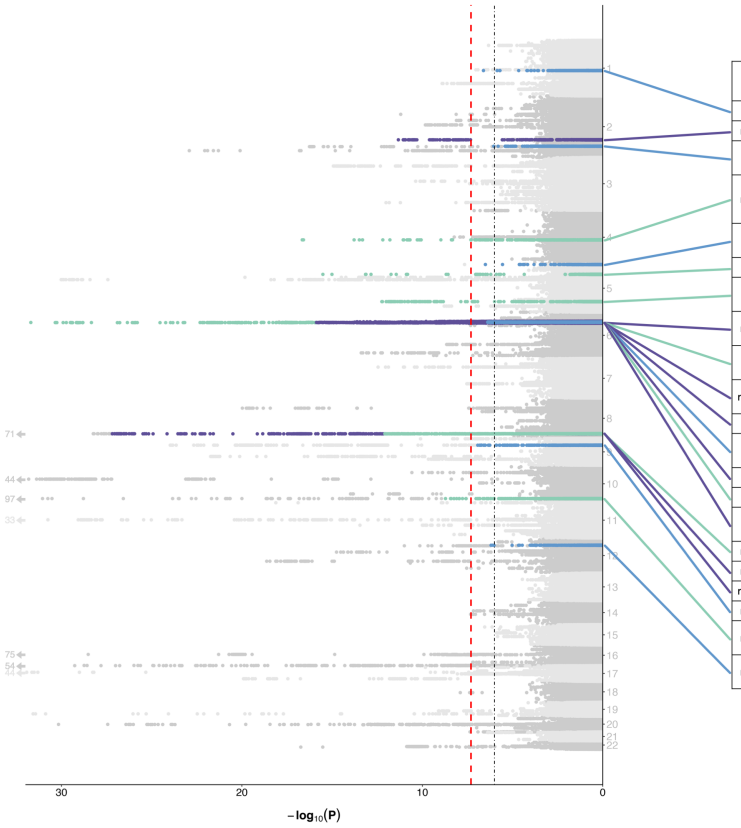
Figure 2. Manhattan plot displaying one-directional variant-specific pleiotropy, where the effect is maximized across all possible subsets of 18 cancers and assumes the same direction of effect across all selected cancers. The red dashed line represents the genome-wide significance threshold ($P < 5x10^{-8}$), and the black dotted line represents a suggestive threshold ($P < 1x10^{-6}$). Highlighted in purple are genome-wide significant loci where the overall pleiotropic *P* is less than all individual *P* for the selected cancers; details for the strongest signal at each locus are provided in the table, including the overall *P* and odds ratio (OR). 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, and highlighted in blue are loci with overall pleiotropic $P < 1x10^{-6}$. All highlighted loci are independent of bidirectional SNPs with smaller overall *P*.



SNP	Locus	P	OR	Associated Cancer Sites*
rs6587551	1q21.3	$2.23x10^{-9}$	1.07	Bl, Ki, Lu, Me, Pa, Pr
rs1398148	1q32.1	$2.82x10^{-10}$	0.92	Bl, Ki, Me, Pr, Th
rs2349073	2q33.1	$3.78x10^{-18}$	1.09	Br, En, Es, Le, Me, NHL, Ov, Pa, Te
rs2293607	3q26.2	$2.46x10^{-11}$	1.08	Bl, Co, En, Es, Ki, Le, Me, Pr, Th
rs148297846	5p15.33	$2.42x10^{-8}$	0.86	Bl, En, Ki, Lu, Pr
rs130071	6p21.33	$3.26x10^{-11}$	0.89	Co, Es, NHL, Or, Re, Te
rs2395191	6p21.32	$5.93x10^{-10}$	0.85	Le, NHL, Or, Th
rs73728618	6p21.32	$5.83x10^{-33}$	1.48	Ce, NHL
rs113661590	8q24.21	$1.06x10^{-8}$	1.10	Br, Es, Or, Pr, Re, Th
rs6983267	8q24.21	$7.52x10^{-74}$	1.24	Co, Pr, Re
9:21964331_CA_C	9p21.3	$2.08x10^{-8}$	0.92	Br, En, Ki, Lu, Me, Or, Ov, Pa, Pr, Te, Th
rs11813268	10q24.33	$2.27x10^{-17}$	0.89	En, Ki, Lu, Me, Or, Ov, Pr, Th
rs78378222	17p13.1	$2.20x10^{-12}$	0.72	Es, Ki, Le, Me, Pa, Pr, Re
rs11263763	17q12	$3.78x10^{-47}$	0.82	En, Pr, Te
rs2532389	17q21.31	$1.17x10^{-9}$	1.07	Br, Ce, Lu, Me, Te
rs4939827	18q21.1	$2.65x10^{-10}$	1.15	Co, Re
rs34978822	20q13.33	$8.21x10^{-10}$	1.34	Bl, Le, Lu, Me, Pr, Th

* 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

Figure 3. Manhattan plot displaying bidirectional variant-specific pleiotropy, where the effect is maximized across all possible subsets of 18 cancers and allows for different directions of effect for selected cancers. The red dashed line represents the genome-wide significance threshold ($P < 5x10^{-8}$), and the black dotted line represents a suggestive threshold ($P < 1x10^{-6}$). Highlighted are loci with overall pleiotropic $P < 1x10^{-6}$, 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, 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, and loci in blue have $P < 1x10^{-6}$. Details for the strongest signal at each highlighted locus are provided in the table, including overall P and odds ratio (OR).



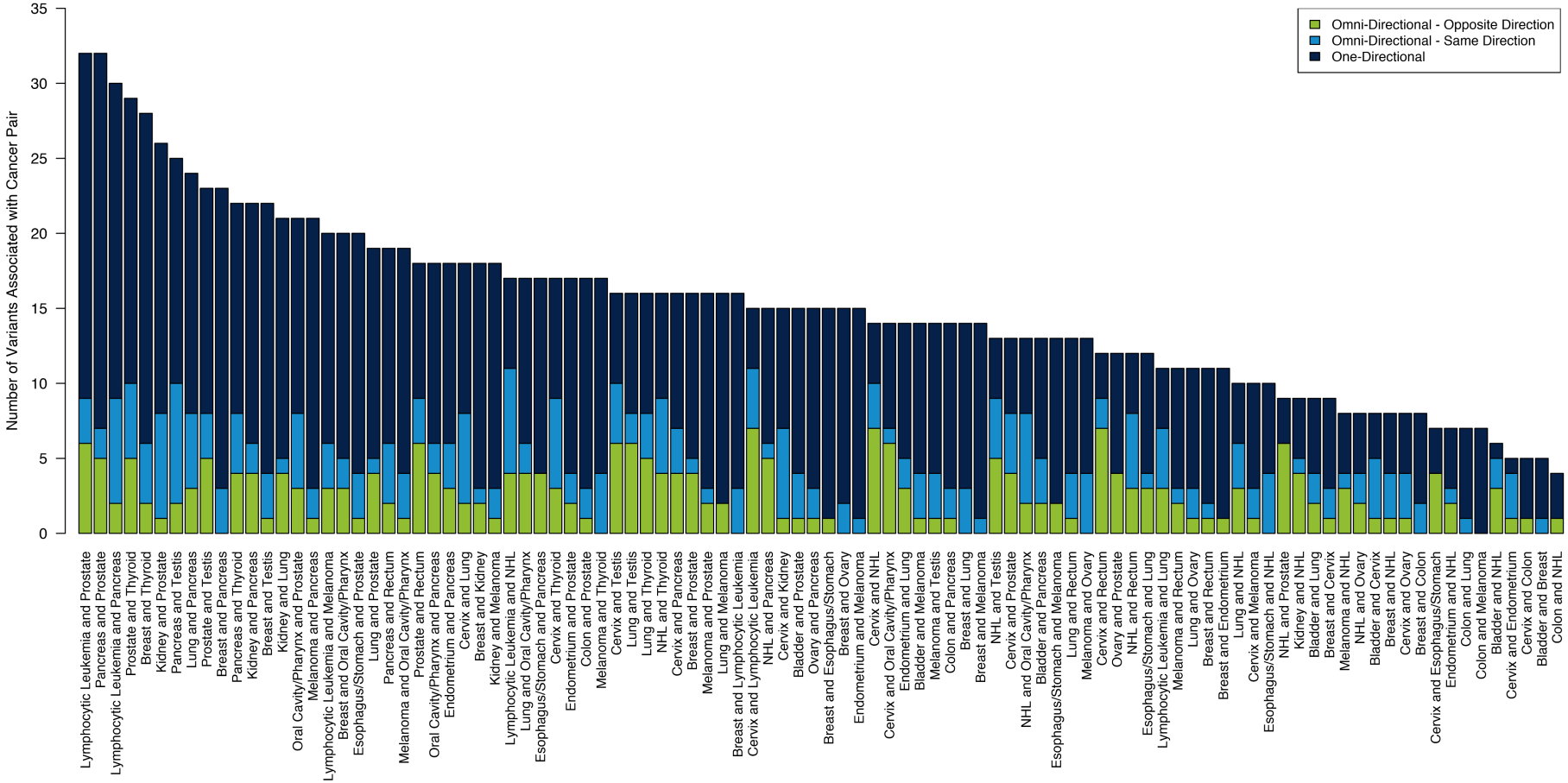
SNP	Locus	P			OR		Associated Cancer Sites*	
		Overall	Increasing	Decreasing	Increasing	Decreasing	Increasing	Decreasing
rs1870940	1q21.3	$2.44x10^{-7}$	$1.78x10^{-5}$	$7.16x10^{-4}$	1.08	0.92	Es, Pr	Le, Me, Ov, Re, Th
rs77282844	2q31.1	$4.68x10^{-12}$	$1.79x10^{-10}$	$8.57x10^{-4}$	1.17	0.85	Bl, Co, Ki, Pr, Th	En, Le, Pa, Re, Te
rs260751	2q33.1	$9.11x10^{-7}$	$9.82x10^{-4}$	$5.22x10^{-5}$	1.05	0.89	Ce, Ki, Le, Pr, Te, Th	Lu, Pa, Re
rs10007915	4q24	$2.19x10^{-17}$	$9.12x10^{-16}$	$5.58x10^{-4}$	1.13	0.96	Pr, Re	Bl, Br, Ce, Ki, Le, Me, NHL, Or, Ov, Te, Th
rs2735944	5p15.33	$3.13x10^{-7}$	$6.38x10^{-4}$	$2.60x10^{-5}$	1.09	0.74	Bl, Ce, En, Lu, Pr	Pa, Te
rs35407	5p13.2	$3.11x10^{-16}$	$2.91x10^{-2}$	$2.64x10^{-16}$	1.32	0.55	Ki, Le, Pa, Te	Me, Or, Ov
rs7717417	5q31.3	$5.70x10^{-13}$	$1.03x10^{-2}$	$1.69x10^{-12}$	1.04	0.64	Br, Lu, NHL, Or, Pa, Th	Te
rs17190106	6p21.33	$1.61x10^{-10}$	$5.37x10^{-10}$	$1.11x10^{-2}$	1.24	0.92	Ce, Lu, NHL, Re	Bl, Le, Me, Or, Pa, Pr, Te, Th
rs9266766	6p21.33	$2.22x10^{-9}$	$1.29x10^{-8}$	$7.14x10^{-3}$	1.20	0.95	Ce, NHL, Th	Br, Co, Le, Pa, Pr, Re, Te
rs114060326	6p21.33	$3.00x10^{-10}$	$4.53x10^{-5}$	$2.53x10^{-7}$	1.30	0.65	Bl, Ce, Le, Lu, Pa, Re, Te	Es, NHL
rs2763979	6p21.33	$1.73x10^{-14}$	$8.60x10^{-10}$	$5.56x10^{-7}$	1.08	0.89	Ce, Ki, Pr	Le, NHL, Or, Re
rs3128761	6p21.33	$4.21x10^{-7}$	$5.40x10^{-5}$	$4.19x10^{-4}$	1.09	0.89	Ce, En, Ki, Lu, Ov, Th	Es, Le, NHL, Or, Re, Te
rs34563311	6p21.32	$1.36x10^{-16}$	$8.76x10^{-9}$	$3.76x10^{-10}$	1.19	0.80	Ce, Le, Lu, Pa	En, NHL, Th
rs9270747	6p21.32	$1.99x10^{-32}$	$1.05x10^{-32}$	$2.41x10^{-2}$	1.26	0.92	Ce	Le, NHL, Re, Te
rs535777	6p21.32	$3.65x10^{-12}$	$7.97x10^{-5}$	$1.49x10^{-9}$	1.08	0.86	Ce, Ki, Pr, Th	Es, Le, NHL, Or, Ov, Re, Te
rs78809737	8q24.21	$8.35x10^{-13}$	$8.57x10^{-13}$	$3.02x10^{-2}$	1.29	0.80	Es, Ki, Or, Pr, Re	En, Pa
rs62516012	8q24.21	$6.64x10^{-28}$	$2.29x10^{-23}$	$4.28x10^{-7}$	1.18	0.93	Or, Pr	Br, NHL, Th
rs117952826	8q24.21	$9.76x10^{-9}$	$4.36x10^{-6}$	$9.93x10^{-5}$	1.30	0.86	Ki, Or, Pr, Th	Br, Ce, Lu
rs55797833	9p21.3	$1.19x10^{-7}$	$1.54x10^{-7}$	$3.87x10^{-2}$	1.71	0.81	Me	Bl, Le, NHL
rs45631563	10q26.13	$1.94x10^{-9}$	$6.75x10^{-8}$	$1.18x10^{-3}$	1.16	0.89	Br, Co, Le, Lu, Ov, Pa, Te, Th	En, Es, Ki, Or, Pr
rs10842692	12p12.1	$6.46x10^{-7}$	$3.27x10^{-2}$	$1.09x10^{-6}$	1.06	0.94	Es, Le, Lu, NHL, Or	Bl, Ce, En, Ki, Me, Ov, Pa, Te, Th

Variants with pleiotropic effect with more than 1 cancer
Overall pleiotropic $P < 1x10^{-6}$
Overall pleiotropic $P < 5x10^{-8}$ and $P <$ all individual cancer P
Overall pleiotropic $P < 5x10^{-8}$ and $P >$ at least 1 individual cancer P

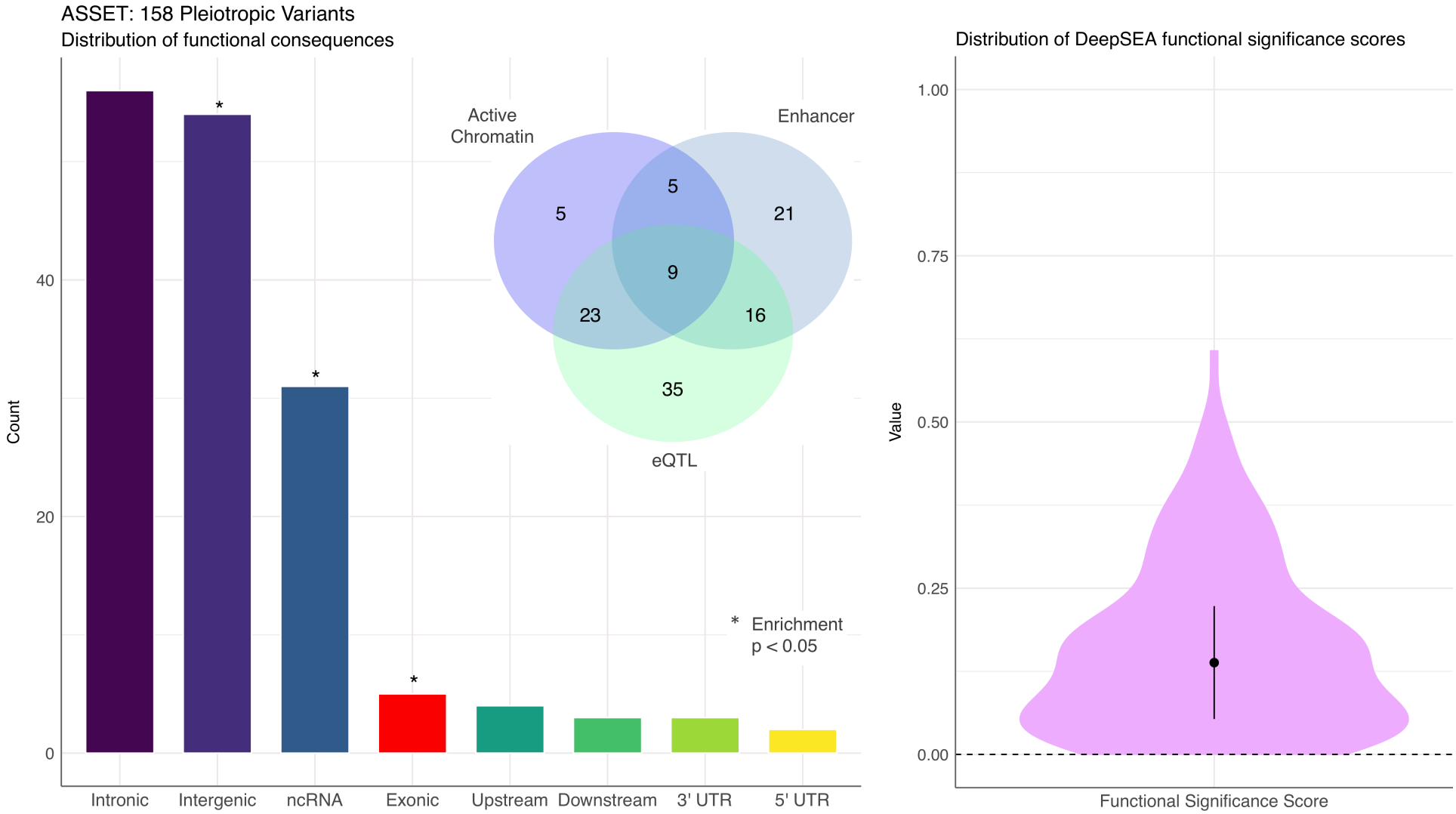
* Bl=Bladder, Br=Breast, Ce=Cervix, Co=Colon, En=Endometrium, Es=Esophagus/Stomach, Ki=Kidney, Le=Lymphocytic Leukemia, Lu=Lung, Me=Melanoma, NHL=Non-Hodgkin's Lymphoma, Or=Oral Cavity/Pharynx, Ov=Ovary, Pa=Pancreas, Pr=Prostate, Re=Rectum, Te=Testis, Th=Thyroid

Figure 4. Summary of the cancer pairs associated with and the functional consequences of the 158 one- and bidirectional pleiotropic variants. (a) The number of pleiotropic variants (of the 158 one- and bidirectional variants with overall pleiotropic $P<1x10^{-6}$) associated with each pair of cancers by type of pleiotropic effect for select cancer pairs. Variants were counted for a cancer pair if they were associated with both cancers using ASSET. For bidirectional variants, effects in the same direction and in opposite directions were tabulated separately. Included are all cancer pairs with any of the following cancer sites: breast, cervix, lung, melanoma, Non-Hodgkin's Lymphoma (NHL), pancreas, and prostate. (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 158 ASSET variants to all variants in the UK Biobank. The Venn diagram summarizes the number of variants with specific regulatory elements, based on analyses of chromatin features from Roadmap and expression quantitative trait loci (eQTL) associations. DeepSEA functional significance score provides an integrated summary score based on evolutionary conservation and chromatin data, with 0 denoting variants most likely to be functional.

(a)



(b)


ASSET: 158 Pleiotropic Variants
Distribution of functional consequences
Count
0
20
40
Intronic
Intergenic
ncRNA
Exonic
Upstream
Downstream
3' UTR
5' UTR
*
*
*
Active Chromatin
Enhancer
eQTL
5
5
21
9
23
16
35
* Enrichment
p < 0.05
Distribution of DeepSEA functional significance scores
Value
0.00
0.25
0.50
0.75
1.00
Functional Significance Score