Title

Assessment of genetic susceptibility to multiple primary cancers through whole-exome sequencing in two large multi-ancestry studies

Journal

BMC Medicine, 20(1)

ISSN

1741-7015

Authors

Cavazos, Taylor B
Kachuri, Linda
Graff, Rebecca E
et al.

Publication Date

2022

DOI

10.1186/s12916-022-02535-6

Peer reviewed

BMC Medicine

RESEARCH ARTICLE

Open Access

Assessment of genetic susceptibility to multiple primary cancers through whole-exome sequencing in two large multi-ancestry studies

Taylor B. Cavazos[1], Linda Kachuri[2,3], Rebecca E. Graff[2,4], Jovia L. Nierenberg[2,5], Khanh K. Thai[4], Stacey Alexeeff[4], Stephen Van Den Eeden[4], Douglas A. Corley[4], Lawrence H. Kushi[4], Regeneron Genetics Center[5], Thomas J. Hoffmann[2], Elad Ziv[5], Laurel A. Habel[4], Eric Jorgenson[6], Lori C. Sakoda[4,7] and John S. Witte[2,3,8*]

Abstract

Background: Up to one of every six individuals diagnosed with one cancer will be diagnosed with a second primary cancer in their lifetime. Genetic factors contributing to the development of multiple primary cancers, beyond known cancer syndromes, have been underexplored.

Methods: To characterize genetic susceptibility to multiple cancers, we conducted a pan-cancer, whole-exome sequencing study of individuals drawn from two large multi-ancestry populations (6429 cases, 165,853 controls). We created two groupings of individuals diagnosed with multiple primary cancers: (1) an overall combined set with at least two cancers across any of 36 organ sites and (2) cancer-specific sets defined by an index cancer at one of 16 organ sites with at least 50 cases from each study population. We then investigated whether variants identified from exome sequencing were associated with these sets of multiple cancer cases in comparison to individuals with one and, separately, no cancers.

Results: We identified 22 variant-phenotype associations, 10 of which have not been previously discovered and were significantly overrepresented among individuals with multiple cancers, compared to those with a single cancer.

Conclusions: Overall, we describe variants and genes that may play a fundamental role in the development of multiple primary cancers and improve our understanding of shared mechanisms underlying carcinogenesis.

Keywords: Multiple primary cancers, Pleiotropy, Whole-exome sequencing, Germline genetics

Background

The substantial global burden of cancer coupled with increasing survival due to improved screening, surveillance, and treatments has yielded a growing number of cancer survivors who are at risk of developing a second primary cancer in their lifetime [1, 2]. The prevalence of multiple primary cancers globally is estimated to range between 2 and 17%, with wide variation likely due to differences in cancer registration practices, case definitions, population characteristics, and follow-up times [1, 2]. Cancer predisposition syndromes, such as Li-Fraumeni, Lynch, and hereditary breast and ovarian cancer, are known to increase the risk of multiple primary cancers; however, less than 2% of all cancers are attributed to hereditary cancer syndromes [1]. Genetic risk factors for

*Correspondence: jswitte@stanford.edu

[3] Department of Epidemiology and Population Health, Stanford University, Alway Building, 300 Pasteur Drive, Stanford, CA 94305, USA
Full list of author information is available at the end of the article

multiple primary cancers beyond known syndromes are not well understood.

Genome-wide association studies (GWAS) have implicated many common, low penetrance variants in 5p15 (*TERT-CLPTM1L*) [3], 6p21 (*HLA*) [4, 5], 8q24 [6], and other loci in the risk of several cancer types. Additional studies have investigated pleiotropy in these regions or characterized cross-cancer susceptibility variants [7, 8]. A pleiotropic locus has the potential to not only affect the risk of many different cancer types, but also increase the likelihood that a single individual develops multiple primary cancers. In our prior work, we discovered that the rare pleiotropic variant *HOXB13* G84E had a stronger association with the risk of developing multiple primary cancers than of a single cancer [9]. This suggests that there may be increased power to detect pleiotropic variation in individuals with multiple primary cancers relative to those with only a single cancer. Identifying widespread pleiotropic signals is informative for understanding shared genetic mechanisms of carcinogenesis, toward the identification of informative markers for cancer prevention and precision medicine.

In this study, we survey the landscape of rare and common variations in individuals with multiple primary cancers, single cancers, and cancer-free controls through whole-exome sequencing (WES) in two large, multi-ancestry studies. We evaluate associations previously discovered in studies of individuals with a single cancer and find novel pleiotropic variation in individuals with multiple primaries.

Methods

Study populations and phenotyping

Our study included ancestrally diverse individuals with multiple primary cancers or no cancer from two large study populations, the Kaiser Permanente Research Bank (KPRB) [10] and the UK Biobank (UKB) [11]. It additionally included individuals with a single cancer in the UKB study population only. From the KPRB, we included individuals who were previously genotyped through the Research Program on Genes, Environment and Health (RPGEH) and the ProHealth Study. For the UKB, we specifically studied participants from the 200K release of WES data [11, 12].

For both study populations, ascertainment of cancer diagnoses has been previously described [7, 13]. Both studies included prevalent and/or incident diagnoses of malignant, borderline, and in situ primary tumors [13]. ICD codes indicating non-melanoma skin cancer or metastatic cancer were not considered primary tumors. Cancers were primarily defined according to the SEER site recode paradigm [14]. However, for hematologic cancers, we incorporated morphology following WHO classifications [15], placing cancers into three major subtypes: lymphoid neoplasms, myeloid neoplasms, and NK- and T-cell neoplasms (Additional file 1: Table S1). Cases were individuals with ICD-9 or ICD-10 codes for primary tumors at two or more distinct organ sites. In the KPRB, controls without a cancer diagnosis, at the last follow-up, were matched 1:1 to cases on age at specimen collection, sex, genotyping array, and reagent kit. In the UKB, controls included all individuals without a cancer diagnosis at the last follow-up.

In both study populations, we excluded duplicates/twins and first-degree relatives, retaining the individual from each related pair who had higher coverage at targeted sites. Following quality control (QC) of WES data (described below), the KPRB and UKB study populations used in this project included 3111 and 3318 cases with multiple primary cancers and 3136 and 162,717 cancer-free controls, respectively. The UKB also contributed 29,091 individuals with a single cancer diagnosis. While our study was primarily unselected for cancer type, prostate cancer cases were oversampled in the KPRB due to the inclusion of individuals from the ProHealth Study.

Genetic ancestry and principal component analysis

Genetic ancestry was defined using genome-wide, imputed array data that underwent extensive QC, as previously described [13]. Ancestry principal components (PCs) were computed using flashPCA2 [16] by projecting our study samples onto PCs defined by 1000G phase 3 reference populations [17]. Individuals were assigned to the closest reference population using distance from the top 10 PCs. Individuals with ancestral PCs greater than five standard deviations from the reference population mean were excluded. The final analytic dataset included individuals of European, African, East Asian, South Asian, and Hispanic/Latino ancestry; however, the analysis was largely biased toward individuals of European ancestry as they were overrepresented (Additional file 2: Fig. S1). A total of $N = 646$ (10.2%) and $N = 8739$ (5.26%) individuals were of non-European ancestry in the KPRB and UKB, respectively (Table 1).

Whole-exome sequencing and quality control

The Regeneron Genetics Center used the Illumina NovaSeq 6000 platform to perform WES for both study populations where the source of DNA was saliva for the KPRB and blood for the UKB. Sample preparation and QC were performed using a high-throughput, fully automated process that has been previously described in detail [18]. Briefly, following sequencing, reads were aligned to the GRCh38 reference genome and variants were called with WeCall [18] for the KPRB and DeepVariant [19] for the UKB. WeCall is a fast, accurate algorithm

Table 1 Characteristics of the Kaiser Permanente Research Bank and UK Biobank study populations by ancestry group. Cases are individuals with multiple primary cancers or a single cancer (for UK Biobank only). Controls are those without any cancer

Population: Kaiser Permanente Research Bank									
	Multiple-cancer cases						**Controls**		
Ancestry	*N*	Mean age	Female (%)				*N*	Mean age	Female (%)
AFR	99	70.5	33.3				100	70.4	32.0
EAS	95	69.7	49.5				91	69.5	49.5
EUR	2,786	72.8	43.0				2,815	72.9	43.3
LAT	131	69.5	46.6				130	69.5	45.4
SAS	-	-	-				-	-	-
Population: UK Biobank									
	Multiple-cancer cases			**Single-cancer cases**			**Controls**		
Ancestry	*N*	Mean age	Female (%)	*N*	Mean age	Female (%)	*N*	Mean age	Female (%)
AFR	29	55.9	51.7	426	56.5	51.4	3,292	51.8	60.4
EAS	10	58.8	80.0	88	55.2	76.1	1,009	52.6	66.9
EUR	3,249	61.9	51.7	27,902	59.4	57.5	154,047	56.6	54.6
LAT	5	63.8	80.0	273	56.0	59.7	334	51.8	62.6
SAS	25	58.2	60.0	402	57.6	58.7	4,035	53.3	47.0

that jointly identifies and infers genotypes at sites relative to a reference genome. DeepVariant is a computationally scalable deep neural network approach to calling variants [20]. WeCall was first used to call variants in the initial 50K release of the UKB whole-exome sequence data and in our KPRB sequence data. Later, DeepVariant was applied to the 200K release of the UKB WES data we use here after we had processed the KPRB data. Regardless, both algorithms have high sensitivity and specificity for calling genetic variants, so their findings should be comparable across the two studies. Finally, samples with gender discordance, 20× coverage at less than 80% of targeted sites, and/or contamination greater than 5% were excluded.

Additional QC was applied to filter low-quality variants and related individuals. First, genotype calls with low depth of coverage (DP) were updated to missing (DP < 7 for SNPs and DP < 10 for indels). Then, sites with low allele balance (AB) were removed. Specifically, variants without at least one sample having AB $\geq$ 15% for SNPs or AB $\geq$ 20% for indels were excluded. Following previous studies [18], we excluded variants with missingness > 10% and HWE p-value < 10^{-15}, computed across all individuals in each study population. After these steps, a total of ~3.51M high-quality sites were retained for the KPRB and ~15.92M were retained for the UKB; excluding singletons, there were ~1.36M and ~8.22M variants, respectively. In the UKB, the larger number of variants observed was due to rare variation present in the larger sample size; when restricting to common variants (MAF > 1%), there were ~186K and ~137K variants, respectively, for the KPRB and UKB.

Association analyses in individuals with multiple cancers versus cancer-free controls

Genetic association analyses of single variants and genes investigated the following cancer phenotypes: (1) diagnosis with at least two primary cancers across any of the 36 organ sites ("any 2+ primary cancers") and (2) groupings of individuals defined by a shared index cancer at one of 16 organ sites with at least 50 cases from each study population ("cancer-specific analyses"). Primary analyses compared multiple cancer cases to cancer-free controls. Within our cancer-specific analyses of 16 organ sites, there were cases shared across our index cancer groupings. For example, the set of individuals with at least one diagnosis of breast cancer overlaps with those having at least one ovarian cancer diagnosis.

Single-variant and gene-based association analyses were performed using REGENIE v2.2.4, a machine-learning approach for performing whole-genome regression to correct for cryptic population structure, as well as adjust for case-control imbalance by applying saddlepoint approximation when the standard case-control p-value is less than 0.05 [21]. We assessed single-variant associations for high-quality variants shared across both populations with minor allele count (MAC) > 2 across cancer phenotype cases and controls within each study. The number of variants tested in our single-variant analyses varied by cancer phenotype (~337K [other female genital cancer-specific analysis] to ~722K [any 2+ primary cancers]). WES variants were functionally annotated using SnpEff v5.0 [22] and dbNFSP v3.5 [23] accessed through ANNOVAR [24]. Missense variants were classified using five algorithms: (1) SIFT ("D"), (2) HDIV from

Polyphen2, (3) HVAR from Polyphen2, (4) LRT ("D"), and (5) MutationTaster ("A" or "D"). For our gene-based burden analyses, we restricted to rare variants with a MAF < 0.5%, including singletons, computed across all individuals within each study population. Following previous work, three gene-based models were evaluated and the model with the lowest p-value was selected [25]: (1) all rare variants with predicted loss of function (pLOF) by SnpEff, (2) pLOF and missense rare variants predicted to be deleterious by the above five classification algorithms, and (3) pLOF and missense rare variants predicted to be deleterious by at least one algorithm. In our gene-based and single-variant analyses, we adjusted for covariates including age, top 10 PCs, and sex (except for sex-specific index cancers of the breast, cervix, ovary, uterus, other female genital organs, and prostate). In the KPRB population, we additionally adjusted for genotyping array and reagent kit, as they were used to perform case-control matching. In the UKB, we adjusted for flow cell (S2 vs S4), which differed for the initial 50K and subsequent 150K release of WES samples.

Single-variant and gene-based burden analyses for each phenotype were combined across study populations in a fixed-effects meta-analysis using METASOFT [26] and metafor v3.0.2 [27], respectively. For our single-variant analyses, we report all suggestive, independent [linkage disequilibrium (LD) $r^2 < 0.2$] associations with $p < 5 \times 10^{-6}$. For our gene-based analyses, we report all associations adjusted for the number of genes tested ($p < 2.65 \times 10^{-6} = 0.05/18{,}842$). In both analyses, we report meta-analysis p-values.

Distinguishing susceptibility signals for multiple cancers versus single cancers

We also evaluated whether the variants and genes associated with the diagnosis of multiple primary cancers (versus non-cancer controls) remained associated when comparing individuals with multiple cancers to those diagnosed with a single cancer. These analyses assessed whether the variants or genes were pleiotropic for developing multiple cancers or general markers of susceptibility to a specific cancer. We undertook these analyses in the UKB sample only, since individuals diagnosed with a single primary cancer were not sequenced in the KPRB. Single-variant and gene-level analyses were implemented as described above. For each variant or gene of interest identified in our case-control analyses, we performed a case-case analysis comparing individuals diagnosed with multiple cancers to those diagnosed with a single cancer. For our cancer-specific analyses, we compared individuals diagnosed with the index cancer plus any other cancer to those diagnosed with the index cancer only. For example, for a finding discovered in our cancer-specific analysis of prostate cancer, we performed a case-case analysis comparing individuals diagnosed with prostate cancer plus any other cancer to individuals with only a prostate cancer diagnosis.

Results

Characterization of multiple primary cancer diagnoses in two large study populations

Our meta-analyses included 6429 cases with multiple primary cancers and 165,853 cancer-free controls (Table 1). All cases had at least two independent primary cancer diagnoses, and 656 cases had more than two diagnoses (Additional file 2: Fig. S2). In the KPRB, the maximum number of cancer diagnoses for an individual was 6 ($n = 1$), and in the UKB, the maximum number was 5 ($n = 2$). Overall, 36 unique cancer sites were represented across multiple cancer cases in the two study populations, with 180 unique pairs of sites (e.g., breast and melanoma) and 298 unique ordered pairs of sites by diagnostic sequence (e.g., breast followed by melanoma) (Additional file 1: Table S2). Only 51 of the 298 ordered pairs had at least 25 cancer cases when grouping individuals by first and second cancer diagnosis (i.e., ignoring any subsequent cancer diagnoses; Additional file 1: Table S2, Fig. 1). The top ordered pairs represented in the combined study populations were prostate then melanoma ($N = 221$), cervix then breast ($N = 202$), melanoma then prostate ($N = 180$), breast then melanoma ($N = 174$), and prostate then colorectal ($N = 170$). Prostate, breast, melanoma, colorectal, and cervix were the most common sites of first cancer diagnoses (Fig. 1). The prevalence of each cancer pair was similar in the KPRB and UKB (Additional file 2: Fig. S3). As most individual cancer pairs were underpowered for downstream analysis, we considered all multi-cancer cases combined, as well as groupings of individuals with a shared index cancer (16 cancers) (Additional file 2: Fig. S4, Additional file 1: Table S3). Among those with multiple cancers, the cancers with the largest number of cases were prostate ($N = 1977$; oversampled in KPRB), breast ($N = 1874$), melanoma ($N = 1443$), colorectal ($N = 1324$), and urinary bladder ($N = 829$).

Exome-wide single-variant association analyses

We found two independent, genome-wide significant associations ($p < 5 \times 10^{-8}$) and 20 suggestive associations ($p < 5 \times 10^{-6}$) between individual variants and the multiple cancer phenotypes (i.e., either any 2+ primary cancers or cancer-specific analyses) (Fig. 2, Additional file 1: Table S4). We found an additional two significant and two suggestive associations (Additional file 2: Fig. S5) in our cancer-specific analyses of lymphoid and myeloid neoplasms; however, we assumed them to represent

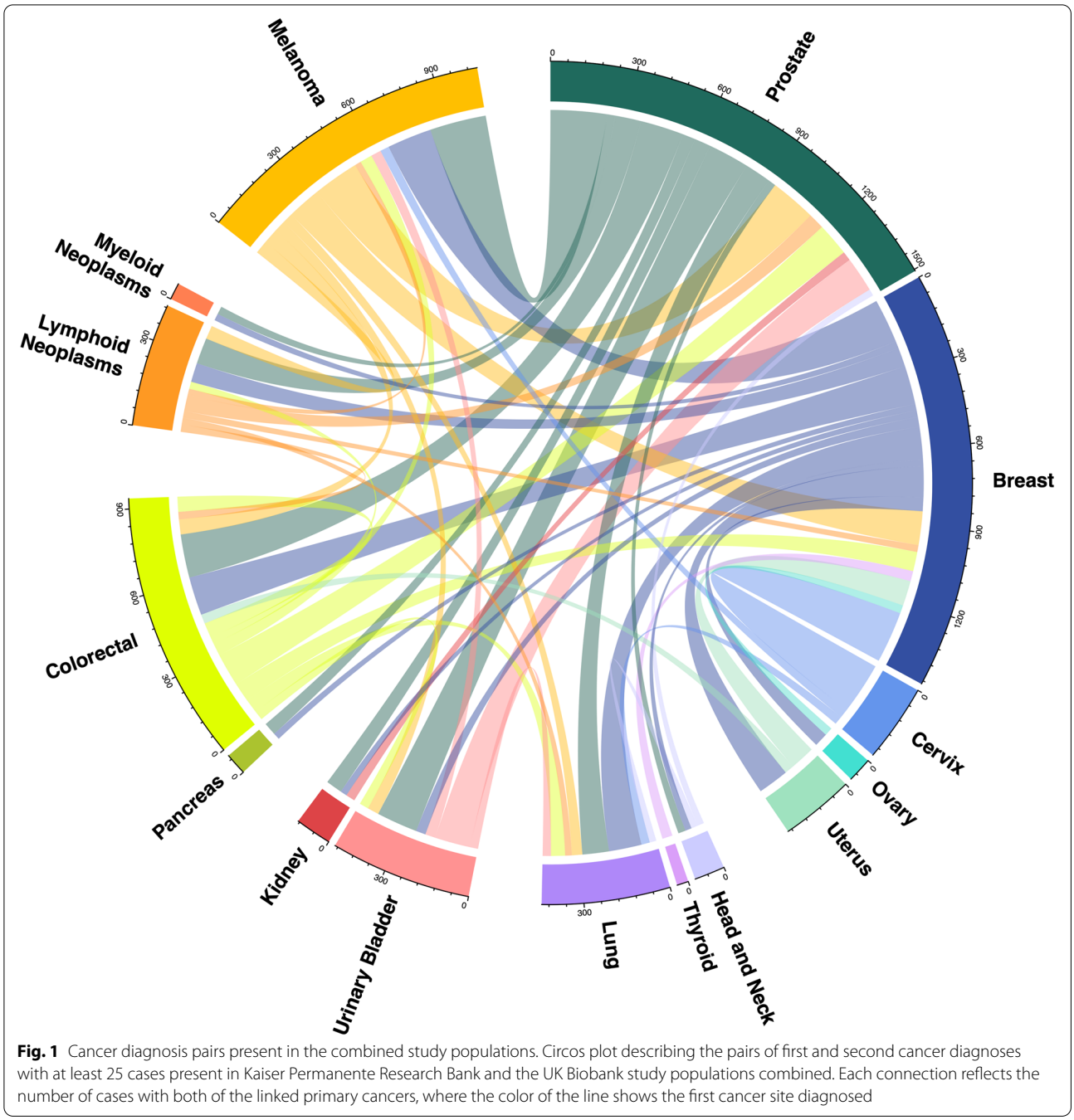

Fig. 1 Cancer diagnosis pairs present in the combined study populations. Circos plot describing the pairs of first and second cancer diagnoses with at least 25 cases present in Kaiser Permanente Research Bank and the UK Biobank study populations combined. Each connection reflects the number of cases with both of the linked primary cancers, where the color of the line shows the first cancer site diagnosed

somatic alterations in the blood as they had low allele balance across our heterogenous samples (Additional file 2: Fig. S6) and occur in genes known to be impacted by clonal hematopoiesis of indeterminate potential (CHIP) [28]. Results were relatively homogeneous across the KPRB and UKB study populations (Additional file 1: Table S4). When stratifying by sex, there were no clear material or statistically significant differences in the results; the associations remain in the same direction and were homogeneous across sex subgroups (Additional file 1: Table S4). Additionally, when restricting analyses to European-only individuals, we found 17 (of the 22) associations had only minor changes (<10%) in their effect estimates and corresponding slight decreases in their *p*-values (Additional file 1: Table S4). Thus, a large majority of our findings likely have similar effects across ancestries and including individuals of all ancestries improves statistical power. The five SNPs with ≥ 10% changes in

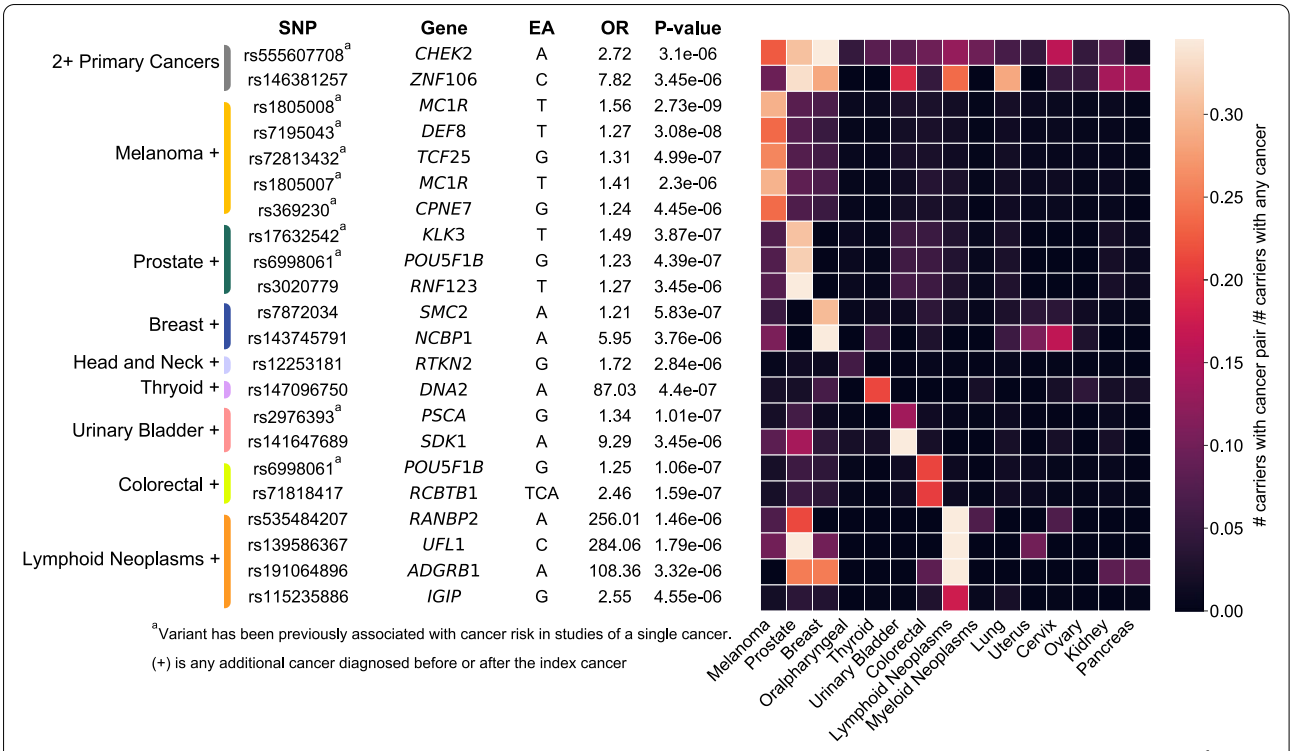

Fig. 2 Germline single-variant association results for multiple primary cancers combined or grouped by organ site. Suggestive ($p < 5 \times 10^{-6}$) germline variant associations with multiple cancer phenotypes versus cancer-free controls ($n = 165{,}853$) following a fixed-effects meta-analysis of Kaiser Permanente Research Bank and UK Biobank WES data. Associations were detected for any 2+ primary cancers ($n = 6429$) and with groups of cases defined by a shared index cancer, at any time point, plus any other cancer diagnosis: melanoma + ($n = 1443$), prostate + ($n = 1977$), breast + ($n = 1874$), head and neck + ($n = 283$), thyroid + ($n = 198$), urinary bladder + ($n = 829$), colorectal + ($n = 1324$), and lymphoid neoplasms + ($n = 728$). Variants that have been previously associated in single cancer studies have superscript (a). The heatmap reflects the number of carriers with the risk-increasing allele for each associated variant with the index (*y*-axis) and additional (*x*-axis) cancer over the total number of carriers, restricting to cancer cases. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers. Abbreviations: SNP, single nucleotide polymorphism; EA, effect allele; OR, odds ratio

their corresponding effects when restricting to the European population may have been driven in part by the non-European ancestry individuals.

Of our 22 findings, two variants were suggestively associated with any 2+ primary cancers, rs555607708 (OR [95% CI] = 2.72 [1.79, 4.15], $p = 3.10 \times 10^{-6}$), a frameshift variant in *CHEK2* known to be associated with risk at many cancer sites [29], and rs146381257 (OR [95% CI] = 7.82 [3.28, 18.62], p = 3.45×10^{-6}), a 5′upstream variant in *ZNF106*. The risk-increasing allele for rs555607708 (*CHEK2*) was most commonly found among individuals with at least one breast cancer (41.9%), prostate cancer (30.6%), melanoma (22.6%), or cervical cancer (16.1%) (Fig. 2). For rs146381257 (*ZNF106*), frequencies were increased in prostate cancer (33.3%), lung cancer (28.6%), breast cancer (28.6%), lymphoid neoplasms (23.8%), urinary bladder cancer (19.0%), pancreatic cancer (14.3%), and kidney cancer (14.3%).

An additional 10 of our findings were previously reported risk variants for a single cancer (Fig. 2). Notably, we detected an association with the *MC1R* variant rs1805008 for melanoma [30] (OR [95% CI] = 1.56 [1.35, 1.81], $p = 2.73 \times 10^{-9}$), when comparing all individuals with at least one melanoma diagnosis plus any other cancer diagnosis to cancer-free controls. We also replicated the previously associated prostate-specific antigen (PSA) variant, rs17632542 [31] (*KLK3*, OR [95% CI] = 1.49 [1.28, 1.73], $p = 3.87 \times 10^{-7}$) in individuals with at least one prostate cancer diagnosis. In addition, we replicated associations between missense risk variant rs6998061 (8q24 locus, *POU5F1B*) and multiple tumor types in both our prostate cancer-specific analysis [32] (OR [95% CI] = 1.23 [1.13, 1.33], $p = 4.39 \times 10^{-7}$) and our colorectal cancer-specific analysis [33] (OR [95% CI] = 1.25 [1.15, 1.37], $p = 1.06 \times 10^{-7}$).

The remaining variants demonstrating associations with multiple cancer phenotypes were not previously associated with any single cancer (Fig. 2). They included a variant discovered in our breast cancer-specific analysis, rs143745791 (*NCBP1*, OR [95% CI]

= 5.95 [2.79, 12.67], $p = 3.76 \times 10^{-6}$), for which 16.2% of carriers, restricted to cases, had a breast and cervical cancer diagnosis, and a variant discovered in our urinary bladder cancer-specific analysis, rs141647689 (*SDK1*, OR [95% CI] = 9.29 [3.63, 23.80], $p = 3.45 \times 10^{-6}$), for which 14.3% of carriers also had prostate cancer (Fig. 2). Three variants found in our lymphoid neoplasm-specific analysis had increased frequencies in cases who also had a diagnosis of prostate cancer: rs535484207 (*RANBP2*, OR [95% CI] = 256.01 [26.82, 2442.95], $p = 1.46 \times 10^{-6}$), rs139586367 (*UFL1*, OR [95% CI] = 284.06 [27.95, 2886.15], $p = 1.79 \times 10^{-6}$), and rs191064896 (*ADGRB1*, OR [95% CI] = 108.36 [15.02, 781.08], $p = 3.32 \times 10^{-6}$), where 21.4%, 40.0%, and 25.0% of carriers for the risk-increasing allele, for each respective variant, had both cancers. The *ADGRB1* variant was also present at increased frequencies among individuals with a lymphoid neoplasm and breast cancer diagnosis (25.0%, Fig. 2). Additionally, we identified a single variant in our head and neck cancer-specific analysis, rs12253181 (*RT*KN2, OR [95% CI] = 1.99 [1.67, 2.37]). Colocalization analyses, within a 500-kb region of the risk SNP, with ezQTL [34] detected a negative correlation between ARID5B expression in whole blood and effects on cancer risk (Additional file 2: Fig. S7). However, these findings should be interpreted with caution since R2 may not adequately control for LD between rare variants and only captures cis-eQTLs in coding regions.

Gene-based analyses of multiple cancers

Out of 18,842 genes tested, we found 10 significant associations ($p < 2.65 \times 10^{-6}$) across our analyses of any 2+ primary cancers and our cancer-specific analyses (Fig. 3, Additional file 1: Table S5). An additional four CHIP genes (*ASXL1*, *TET2*, *JAK2*, and *DDX41*) were significantly associated with myeloid neoplasms and are likely driven by somatic alterations (Additional file 2: Fig. S8).

In our analyses of any 2+ primary cancers and our breast cancer-specific analysis, we replicated associations for known pleiotropic genes, *BRCA2* (pLOF, $p = 3.76 \times 10^{-11}$ and 1.91×10^{-9}) and *CHEK2* (pLOF + missense, $p = 2.95 \times 10^{-11}$ and 1.67×10^{-8}) (Fig. 3). *BRCA2* also emerged in our ovarian cancer-specific analysis (pLOF, $p = 1.91 \times 10^{-9}$). We found associations between the known prostate cancer gene *ATM* and any 2+ primary cancers and in our prostate cancer-specific analysis (pLOF + missense, $p = 9.84 \times$

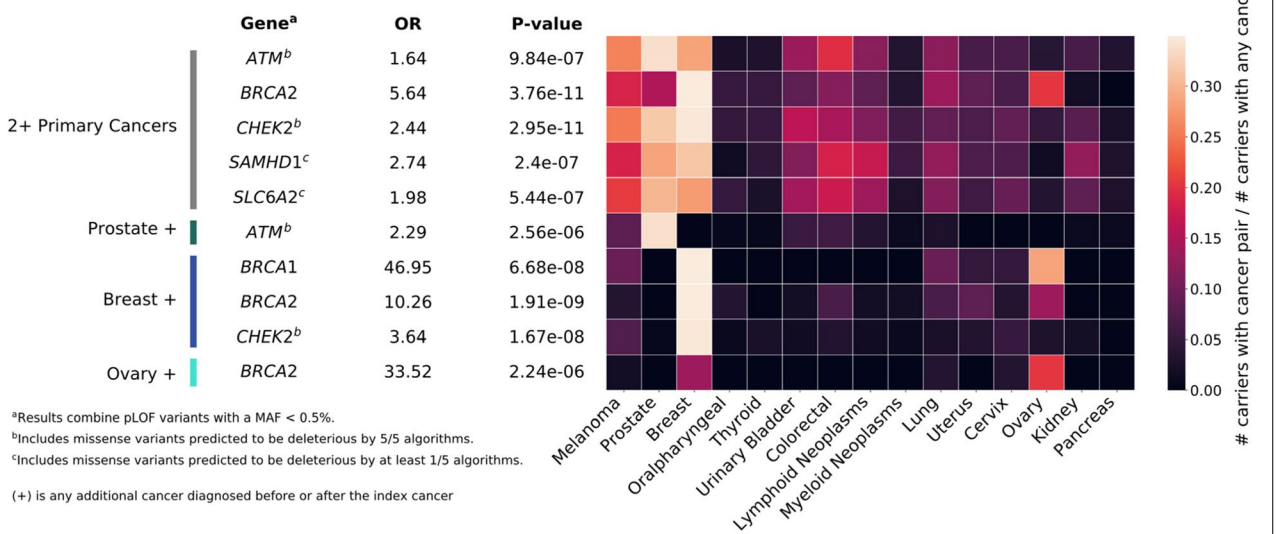

Fig. 3 Germline gene-based association results for multiple primary cancers combined or grouped by organ site. Burden tests were performed combining variants defined as pLOF with or without deleterious missense variants, defining deleteriousness by at least one (1/5) or all five (5/5) prediction algorithms used (Methods), at a MAF < 0.5%. Following a fixed-effects meta-analysis of Kaiser Permanente Research Bank and UK Biobank data, Bonferroni significant associations ($p < 2.65 \times 10^{-6} = 0.05/18{,}842$) corrected for the number of genes tested were found for comparisons of cancer-free controls ($n = 165{,}853$) with all cases with any 2+ primary cancers ($n = 6429$) and with groups of cases defined by an index cancer for the following phenotypes: prostate + ($n = 1977$), breast + ($n = 1874$), and ovary + ($n = 239$). For each gene, the variant grouping with the smallest *p*-value was selected. The heatmap reflects the number of carriers of each associated variant, with the index (*y*-axis) and additional (*x*-axis) cancer over the total number of carriers, where the carrier is defined as having at least one alternate allele across all variants in a given gene, restricting to cancer cases. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers. Abbreviations: OR, odds ratio; pLOF, predicted loss of function

10^{-7} and 2.56×10^{-6}). Additional associations were observed between *SAMHD1* and *SLC642* and any 2+ primary cancers (pLOF + missense, $p = 2.40 \times 10^{-7}$ and $p = 5.44 \times 10^{-7}$, respectively). *BRCA1* also surfaced in the breast cancer-specific analysis (pLOF, $p = 6.68 \times 10^{-8}$).

Predicted loss of function variants in *BRCA1* and *BRCA2* were present at increased frequencies in individuals with a breast cancer diagnosis and ovary as an additional cancer site (Fig. 3), such that 28.6% and 13.6% of individuals, respectively, were a carrier for at least one variant in the burden set. For *BRCA1*, there was also an increase of carriers with an additional melanoma (9.52%) or lung cancer (9.52%) diagnosis. For *BRCA2*, there was an increase of carriers with an additional uterine (8.47%), lung (6.78%), or colorectal cancer (6.78%).

Comparison of mutation burden in individuals with multiple versus single cancers

Out of the 22 associated variants (Fig. 2), 10 remained associated when comparing individuals with multiple cancers to those with single cancers (Additional file 1: Table S6; $p < 0.05$). Two of these variants were positively associated in our analysis of any 2+ primary cancers: rs555607708 (*CHEK2*; OR [95% CI] = 1.57 [1.09, 2.25], $p = 0.015$) and rs146381257 (*ZNF106;* OR [95% CI] = 5.38 [1.07, 27.18], $p = 0.042$). The other eight variants were positively associated with the diagnosis of a specific index cancer plus any other cancer versus the specific cancer alone (Additional file 1: Table S6). Two of these eight variants were associated in our breast cancer-specific case-case analysis: rs7872034, a missense variant in *SMC2* (OR [95% CI] = 1.16 [1.05, 1.27], $p = 0.0025$), and rs143745791, a missense variant in *NCBP1* (OR [95% CI] = 3.71 [2.08, 6.61], $p = 8.37 \times 10^{-6}$).

Of the 10 findings from the gene-level burden analyses (Fig. 3), eight remained positively associated with multiple cancers in comparison with single cancers ($p < 0.05$; Additional file 1: Table S7). Five of these genes were discovered in our case-case analysis of any 2+ primary cancers: *SLC6A2* (OR [95% CI] = 1.86 [1.42, 2.41], $p = 3.90 \times 10^{-6}$), *ATM* (OR [95% CI] = 1.42 [1.15, 1.77], $p = 1.10 \times 10^{-3}$), *CHEK2* (OR [95% CI] = 1.56 [1.23, 1.98], $p = 2.31 \times 10^{-4}$), *SAMHD1* (OR [95% CI] = 1.56 [1.14, 2.13], $p = 5.34 \times 10^{-3}$), and *BRCA2* (OR [95% CI] = 1.86 [1.31, 2.65], $p = 5.42 \times 10^{-4}$). *ATM* (OR [95% CI] = 1.82 [1.20, 2.75], $p = 4.64 \times 10^{-3}$) was positively associated in our prostate cancer-specific case-case analysis, and the two remaining genes were positively associated in our breast cancer-specific case-case analysis: *BRCA1* (OR [95% CI] = 2.38 [1.07, 5.30], $p = 0.0340$) and *BRCA2* (OR [95% CI] = 1.97 [1.22, 3.18], $p = 5.50 \times 10^{-3}$).

Discussion

We investigated the genetic basis of carcinogenic pleiotropy through whole-exome sequencing of individuals diagnosed with multiple primary cancers from two large, multi-ancestry study populations. Comparing individuals with multiple cancers to cancer-free controls uncovered 22 independent, suggestively associated variants, ten of which remained associated when comparing individuals with multiple cancers to those with a single cancer. Across our multiple cancer phenotypes, we also recapitulated previously known gene-based associations in *ATM*, *BRCA1/2*, and *CHEK2* and found potentially novel associations in *SAMHD1* and *SLC6A2*. These genes remained associated with multiple cancer diagnoses when comparing to individuals with a single cancer. These findings offer insights into germline exome variants that increase an individual's risk of developing multiple primary cancers.

Compelling findings from our analyses of all individuals with more than one cancer diagnosis include associations with the rare variant rs146381257 in *ZNF106*. Carriers of the rs146381257 risk allele (C) were primarily overrepresented in individuals with at least one prostate, breast, lung, or urinary bladder cancer and in individuals with lymphoid neoplasms. Carriers also demonstrated an increased risk of developing multiple cancers compared to individuals with a single cancer. *ZNF106* is an RNA binding protein involved in post-transcriptional regulation and insulin receptor signaling. Although germline variation in *ZNF106* has not previously been associated with cancer risk, a recent study found it to be associated with worse urinary bladder cancer survival [35].

Additional noteworthy findings from our analyses of all multiple primary cancers combined include cancer susceptibility signals in *SAMHD1* and *SLC6A2*, both having a significantly higher risk being diagnosed with multiple cancers compared to single cancers. Germline *SAMHD1* mutations are implicated in Aicardi-Goutieres syndrome (AGS) [36], an autosomal recessive condition that results in autoimmune inflammatory encephalopathy. Most cancer-related studies have focused on the role of somatic alternations in *SAMHD1* [37]; however, a study of chronic lymphoid leukemia (CLL) proposed an oncogenic role of germline *SAMHD1* variation mediated by DNA repair mechanisms [38]. Consistent with this hypothesis, we also found increased *SAMHD1* variation in individuals with lymphoid neoplasms, as well as with prostate, breast, colorectal, and lung cancers. *SLC6A2*, also

known as *NAT1*, has been found to be prognostic for colon cancer [39], and both in vivo and in vitro studies have linked expression to survival in many cancer types, including prostate [40] and breast [41]. Polymorphisms in *SLC6A2* may also interact with smoking exposure to modulate the risk for tobacco-related cancers [42].

Because we compared multiple primary cancers with both cancer-free controls and individuals diagnosed with a single cancer, we were well positioned to explore patterns of pleiotropy and disentangle variation likely to be driven by single cancers. For example, we identified two variants, rs7872034 (missense variant in *SMC2*) and rs143745791 (missense variant in *NCBP1*), suggestively associated with a diagnosis of at least one breast cancer (plus any other cancer) versus no cancer. These variants remained associated with a diagnosis of breast and another cancer when comparing to individuals diagnosed with a single breast cancer. While rs7872034 is in high LD ($r^2 = 0.98$) with a known breast cancer risk variant (rs4742903; *SMC2* intron) [43], it may also increase the risk of developing multiple cancers. Regarding rs143745791, germline variants in *NCBP1* have not been previously associated with cancer; because it is rare (MAF < 0.2%), larger sequencing efforts may be necessary to identify variation in studies of individuals with a single cancer. Expression of this gene has been found to promote lung cancer growth and poor prognosis [44], and *NCBP1* is overexpressed in basal-like and triple-negative breast cancers [45]. Similarly, *BRCA1/2* germline variants are prevalent among these subtypes; however, in our study populations, *BRCA1/2* carriers were more common among those with an additional ovarian cancer whereas *NCBP1* carriers more frequently had an additional cervical cancer.

In our prostate cancer-specific analysis comparing individuals with multiple cancers versus those with only a single cancer, we discovered a suggestive association with rs3020779, an eQTL for *RNF123* (also known as *KPC1*), which is a gene involved in p50 mediation and downstream stimulation of multiple tumor suppressors [46]. In our analysis of head and neck cancer, we detected an association with rs12253181, located in the 3′-UTR of *RTNK2*. Integration of whole blood gene expression data at this locus determined that another nearby gene, *ARID5B*, may be a more likely candidate. Expression of *ARID5B* was negatively correlated with the cancer susceptibility signal in this region. While this gene has not previously been associated with head and neck cancer risk, germline variation in *ARID5B* has been implicated in acute lymphoblastic leukemia (ALL) [47], as well as treatment resistance and higher rates of relapse [48]. Genetic variants in *ARID5B* have also been linked to autoimmune diseases [49, 50], suggesting that immune dysregulation may be a plausible pleiotropic mechanism at this locus, especially given the infectious etiology of oropharyngeal carcinoma [51, 52].

Our findings have potential implications for improving our understanding of the shared mechanisms of carcinogenesis. With further replication, they may also enable prevention (e.g., smoking cessation) and screening strategies that prioritize individuals at risk for developing additional cancers. For example, women who carry the rare missense variant in *NCBP1* (rs143745791) were estimated to have an approximately sixfold higher risk of developing breast and other cancers in comparison with no cancer and an approximately threefold higher risk in comparison with women diagnosed with breast cancer alone. If replicated, such findings suggest that the pleiotropic variants reported here could have clinical significance for preventative cancer screening and early detection among individuals with a previous cancer diagnosis.

Limitations of our study included the identification of variants that were likely somatic in our analyses of hematologic cancers due to an expansion of hematopoietic clonal populations with the same acquired mutation (i.e., CHIP). Confounding of germline testing by CHIP has been reported in *TP53* [53] and *TET2* [54], so careful interpretation is critical to avoid unnecessary clinical intervention. An additional limitation of our, and other, studies are obtaining accurate effect estimates for rare variants and the reliance on available annotations for inclusion into gene-based tests. Although heterogeneity was minimal in our study, differences in effects across populations may reflect differences in population characteristics and sample size. Replication of rare findings in larger cohorts and optimization of functional impact annotations could lead to more precise results. Also, our approach did not allow for formal replication, due to the limited sample size of each cohort. In order to identify signals for our largely understudied phenotype, we combined the two cohorts in a meta-analysis rather than undertaking underpowered replication. Finally, while all individuals with multiple cancers were included in our study regardless of genetic ancestry, individuals of non-European ancestry were underrepresented; larger, more diverse cohorts will be needed to fully explore the genetic basis of multiple cancers.

Selection bias and phenotypic misclassification may also have biased our results. We combined prevalent and incident cancer cases together to maximize statistical power for detecting potential associations. The prevalent cases may include fewer individuals with worse prognosis since these individuals may be less likely included in the

study. If any pleiotropic variants reflect more aggressive disease, this could lead to underestimating their potential associations, and vice-versa. Also, the controls' disease status is conditional on their being cancer free at the last follow-up. If some controls would eventually be diagnosed with cancer, then any associations would be underestimated. There is the potential that recurrences arising from the first cancer may have been misclassified as second primaries. If so, this may overestimate pleiotropic associations. In our study, 10.3% and 17.6% of second primaries that occurred within 1 year of the index cancer in the KPRB and UKB respectively may represent recurrences. However, the average age at diagnosis between first and second cancers was 8.3 years (median = 7) in the KPRB and 9.5 years (median = 6.5) in the UKB, suggesting that the majority of multiple cancer cases were most likely second primaries.

Strengths of this work include studying individuals of multiple ancestries who were largely unselected for specific cancer phenotypes. We also performed the first ever exome-wide study of genetic susceptibility to multiple primary cancers, using two large multi-ancestry study populations. Our study design allowed us to characterize variation across multiple primary cancers representing 36 unique sites, as well as to conduct cancer-specific analyses of 16 sites. Using this approach, we confirmed many known single-variant and gene-based findings, strengthening and supporting our novel results reported for individual cancers through our cancer-specific analyses.

In summary, by undertaking an exome-wide survey of common and rare variations in two large study populations, we identified several variant and gene-based associations that may increase the risk of developing multiple cancers within individuals. Future studies should aim to replicate our findings and undertake experiments that validate the functionality of the discovered pleiotropic variants. Combined with future research, our results have the potential to inform genetic counseling, improve risk prediction for multiple cancers, and guide novel treatment and drug development.

Conclusions

This study examines the genetic underpinnings of multiple primary cancers in two large, multi-ancestry population-based cohorts. Analyses of single-variant and gene-level associations identified novel patterns of cross-cancer pleiotropy and confirmed results in key cancer genes.

Abbreviations

AB: Allele balance; DP: Depth of coverage; EA: Effect allele; HWE: Hardy-Weinberg equilibrium; KPRB: Kaiser Permanente Research Bank; LD: Linkage disequilibrium; MAC: Minor allele count; MAF: Minor allele frequency; OR: Odds ratio; PC: Principal component; pLOF: Predicted loss of function; QC: Quality control; SNP: Single nucleotide polymorphism; UKB: UK Biobank; WES: Whole-exome sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-022-02535-6.

Additional file 1: Table S1. Cancer Site Coding Following SEER and WHO Guidelines. **Table S2.** Unique Cancer Pairs and Sample Counts in the Kaiser Permanente Research Bank and UK Biobank. **Table S3.** Sample Counts for Shared Index Cancers Across All Diagnoses in the Kaiser Permanente Research Bank and UK Biobank. **Table S4.** Single Variant Association Summary Statistics for Multiple Primary Cancers Combined or Grouped by Organ Site. **Table S5.** Gene-Based Association Summary Statistics for Multiple Primary Cancers Combined or Grouped by Organ Site. **Table S6.** Single Variant Burden in Individuals with Multiple Versus Single Cancers. **Table S7.** Gene-Based Burden in Individuals with Multiple Versus Single Cancers.

Additional file 2: Figure S1. Genetic Ancestry in the Kaiser Permanente Research Bank and UK Biobank. **Figure S2.** Time Intervals Between Multiple Cancer Diagnoses in the Kaiser Permanente Research Bank and UK Biobank. **Figure S3.** Circos Plots of Cancer Pairs in the Kaiser Permanente Research Bank and UK Biobank. **Figure S4.** Cancers Represented in the Kaiser Permanente Research Bank and UK Biobank with Sufficient Sample Size for Exome-wide Association Analyses. **Figure S5.** Significant Single-Variant Association Results Due to Clonal Hematopoiesis of Indeterminate Potential. **Figure S6.** Allele Balance for Findings Related to Lymphoid and Myeloid Neoplasms. **Figure S7.** Z-Z plot for expression at ARID5B. **Figure S8.** Significant Gene-Based Association Results Due to Clonal Hematopoiesis of Indeterminate Potential.

Acknowledgements

We are grateful to the Kaiser Permanente Northern California members who have generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment, and Health and the ProHealth Study. The authors also thank the Regeneron Genetics Center for covering the costs of whole-exome sequencing of the Kaiser Permanente Research Bank study participants.

Regeneron Genetics Center author list and contribution

RGC Management and Leadership Team

Goncalo Abecasis, D.Phil., Aris Baras, M.D., Michael Cantor, M.D., Giovanni Coppola, M.D., Andrew Deubler, Aris Economides, Ph.D., Katia Karalis, Ph.D., Luca A. Lotta, M.D., Ph.D., John D. Overton, Ph.D., Jeffrey G. Reid, Ph.D., Katherine Siminovitch, M.D., Alan Shuldiner, M.D.

Sequencing and Lab Operations

Christina Beechert, Caitlin Forsythe, M.S., Erin D. Fuller, Zhenhua Gu, M.S., Michael Lattari, Alexander Lopez, M.S., John D. Overton, Ph.D., Maria Sotiropoulos Padilla, M.S., Manasi Pradhan, M.S., Kia Manoochehri, B.S., Thomas D. Schleicher, M.S., Louis Widom, Sarah E. Wolf, M.S., Ricardo H. Ulloa, B.S.

Clinical Informatics

Amelia Averitt, Ph.D., Nilanjana Banerjee, Ph.D., Michael Cantor, M.D., Dadong Li, Ph.D., Sameer Malhotra, M.D., Deepika Sharma, MHI, Jeffrey Staples, Ph.D.

Genome Informatics

Xiaodong Bai, Ph.D., Suganthi Balasubramanian, Ph.D., Suying Bao, Ph.D., Boris Boutkov, Ph.D., Siying Chen, Ph.D., Gisu Eom, B.S., Lukas Habegger, Ph.D., Alicia Hawes, B.S., Shareef Khalid, Olga Krasheninina, M.S., Rouel Lanche, B.S., Adam J. Mansfield, B.A., Evan K. Maxwell, Ph.D., George Mitra, B.A., Mona Nafde, M.S., Sean O'Keeffe, Ph.D., Max Orelus, B.B.A., Razvan Panea, Ph.D., Tommy Polanco, B.A., Ayesha Rasool, M.S., Jeffrey G. Reid, Ph.D., William Salerno, Ph.D. , Jeffrey C. Staples, Ph.D., Kathie Sun, Ph.D., Jiwen Xin, Ph.D.

Analytical Genomics and Data Science
Goncalo Abecasis, D.Phil., Joshua Backman, Ph.D., Amy Damask, Ph.D., Lee Dobbyn, Ph.D., Manuel Allen Revez Ferreira, Ph.D., Arkopravo Ghosh, M.S., Christopher Gillies, Ph.D., Lauren Gurski, B.S., Eric Jorgenson, Ph.D., Hyun Min Kang, Ph.D., Michael Kessler, Ph.D., Jack Kosmicki, Ph.D., Alexander Li, Ph.D., Nan Lin, Ph.D., Daren Liu, M.S., Adam Locke, Ph.D., Jonathan Marchini, Ph.D., Anthony Marcketta, M.S., Joelle Mbatchou, Ph.D., Arden Moscati, Ph.D., Charles Paulding, Ph.D., Carlo Sidore, Ph.D., Eli Stahl, Ph.D., Kyoko Watanabe, Ph.D., Bin Ye, Ph.D., Blair Zhang, Ph.D., Andrey Ziyatdinov, Ph.D.

Research Program Management & Strategic Initiatives
Marcus B. Jones, Ph.D., Jason Mighty, Ph.D., Lyndon J. Mitnaul, Ph.D.

Authors' contributions
TBC contributed to the conception and design of the work, analyzed and interpreted the data, wrote the initial draft of the article, and participated in drafting and revising it critically. LK and JLN contributed to data analysis and participated in writing and drafting the article and revising it critically. REG, LCS, and JSW contributed to the conception and design of the work, data acquisition, writing and drafting the article, and revising it critically. KKT, SA, SV, DAC, LHK, LH, and EJ contributed to the data acquisition and revising the article critically. TJH and EZ revised the article critically. The authors read and approved the final manuscript.

Authors' information
Not applicable

Funding
This material is based upon work supported by NIH grant R01 CA201358, RC2 AG036607, and the National Science Foundation Graduate Research Fellowship Program under Grant No. 1650113. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support for study enrollment, survey administration, and biospecimen collection of Kaiser Permanente Research Bank participants 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. Additionally, LK is supported by funding from the NCI (K99CA246076) and REG is supported by a Young Investigator Award from the Prostate Cancer Foundation. This research has been conducted using the UK Biobank Resource under Application Number 14015.

Availability of data and materials
All results generated from this study are included in the published article or Supplementary Materials. The UK Biobank cohort data is publicly available from the UK Biobank access portal at https://www.ukbiobank.ac.uk. The Kaiser Permanente Research Bank data are available on dbGaP (phs002809.v1.p1). All remaining relevant data are available in the article, supplementary information, or from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
The research was conducted with approved access to UK Biobank data under application number 14105 (PI: Witte) and in accordance with the UK Biobank Ethics and Governance Framework. UK Biobank data are publicly available by request from https://www.ukbiobank.ac.uk. The de-identified participants from the Kaiser Permanente Research Bank were obtained with approval under UCSF IRB number 16-19699 (PI: Witte). Access to and use of data from the Kaiser Permanente Research Bank (KPRB) were approved by the KPRB Access Review Committee. The study was conducted under a waiver of informed consent approved by the Kaiser Permanente Northern California Institutional Review Board. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.
I confirm that all necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived and that any patient/participant/sample identifiers included were not known to anyone (e.g., hospital staff, patients, or participants themselves) outside the research group so cannot be used to identify individuals.

Consent for publication
Not applicable

Competing interests
J.S. Witte is a non-employee, cofounder of Avail Bio. E. Jorgenson and additional authors listed under "Regeneron Genetics Center" are full-time employees of Regeneron Pharmaceuticals. No disclosures were reported for the other authors.

Author details
[1]Biological and Medical Informatics, University of California San Francisco, San Francisco, CA 94158, USA. [2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [3]Department of Epidemiology and Population Health, Stanford University, Alway Building, 300 Pasteur Drive, Stanford, CA 94305, USA. [4]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612, USA. [5]Regeneron Genetics Center, Tarrytown, NY 10591, USA. [6]Department of Medicine, University of California San Francisco, San Francisco, CA 94158, USA. [7]Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA 91101, USA. [8]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA.

Received: 16 March 2022 Accepted: 17 August 2022
Published online: 06 October 2022

References

1. Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, et al. Multiple primary tumours: challenges and approaches, a review. ESMO Open. 2017;2:e000172.
2. Copur MS, Manapuram S. Multiple primary tumors over a lifetime. Oncology (Williston Park). 2019;33:629384.
3. Gaspar TB, Sá A, Lopes JM, Sobrinho-Simões M, Soares P, Vinagre J. Telomere maintenance mechanisms in cancer. Genes. 2018;9:241.
4. Smedby KE, Foo JN, Skibola CF, Darabi H, Conde L, Hjalgrim H, 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. 2011;7:e1001378.
5. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Sci Transl Med. 2017;9:eaai8708.
6. Huppi K, Pitt JJ, Wahlberg BM, Caplen NJ. The 8q24 gene desert: an oasis of non-coding transcriptional activity. Front Genet. 2012;3:69.
7. Rashkin SR, Graff RE, Kachuri L, Thai KK, Alexeeff SE, Blatchins MA, et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat Commun. 2020;11:4423.
8. Lindström S, Finucane H, Bulik-Sullivan B, Schumacher FR, Amos CI, Hung RJ, et al. Quantifying the genetic correlation between multiple cancer types. Cancer Epidemiol Biomark Prev. 2017;26:1427–35.
9. Hoffmann TJ, Sakoda LC, Shen L, Jorgenson E, Habel LA, Liu J, et al. Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. PLoS Genet. 2015;11:e1004930.
10. Witte JS, Van Den Eeden S, Chao CR, Ghai NR, Hoffmann TJ, Risch N, et al. ProHealth: Kaiser Permanente genome-wide association study of prostate cancer: dbGaP; 2020. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001221.v1.p1
11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779 http://www.ukbiobank.ac.uk/.
12. Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet. 2021;53:942–8.

13. Graff RE, Cavazos TB, Thai KK, Kachuri L, Rashkin SR, Hoffman JD, et al. Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. Nat Commun. 2021;12:970.
14. Adamo M, Groves C, Dickie L, Ruhl J. SEER program coding and staging manual 2021. Bethesda: National Cancer Institute; 2020. p. 20892.
15. Harris NL, Jaffe ES, Diebold J, Flandrin G, Muller-Hermelink HK, Vardiman J, et al. The World Health Organization classification of neoplasms of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee Meeting – Airlie House, Virginia, November, 1997. Hematol J. 2000;1:53–66.
16. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33:2776–8.
17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
18. Geisinger-Regeneron DiscovEHR Collaboration, Regeneron Genetics Center, Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586:749–56.
19. Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2021;36:5582–9.
20. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.
21. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021;53:1097–103.
22. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
23. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24:2125–37.
24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
25. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021;599:628–34.
26. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88:586–98.
27. Viechtbauer W. Conducting meta-analyses in *R* with the metafor package. J Stat Softw. 2010;36:1–48.
28. Steensma DP, Bejar R, Jaiswal S, Lindsley RC, Sekeres MA, Hasserjian RP, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood. 2015;126:9–16.
29. Cybulski C, Górski B, Huzarski T, Masojć B, Mierzejewski M, Dębniak T, et al. CHEK2 is a multiorgan cancer susceptibility gene. Am J Hum Genet. 2004;75:1131–5.
30. Amos CI, Wang L-E, Lee JE, Gershenwald JE, Chen WV, Fang S, et al. Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. Hum Mol Genet. 2011;20:5012–23.
31. Li H, Fei X, Shen Y, Wu Z. Association of gene polymorphisms of KLK3 and prostate cancer: a meta-analysis. Adv Clin Exp Med. 2020;29:1001–9.
32. Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, Coetzee SG, et al. Comprehensive functional annotation of 77 prostate cancer risk loci. PLoS Genet. 2014;10:e1004102.
33. Hutter CM, Slattery ML, Duggan DJ, Muehling J, Curtin K, Hsu L, et al. Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. BMC Cancer. 2010;10:670.
34. Zhang T, Klein A, Sang J, Choi J, Brown KM. ezQTL: a web platform for interactive visualization and colocalization of quantitative trait loci and GWAS. Genomics Proteomics Bioinformatics. 2022;S1672-0229(22):00069.
35. Wu Y, Liu Z, Wei X, Feng H, Hu B, Liu B, et al. Identification of the functions and prognostic values of RNA binding proteins in bladder cancer. Front Genet. 2021;12:574196.
36. Martinez-Lopez A, Martin-Fernandez M, Buta S, Kim B, Bogunovic D, Diaz-Griffero F. SAMHD1 deficient human monocytes autonomously trigger type I interferon. Mol Immunol. 2018;101:450–60.
37. Mauney CH, Hollis T. SAMHD1: recurring roles in cell cycle, viral restriction, cancer, and innate immunity. Autoimmunity. 2018;51:96–110.
38. Clifford R, Louis T, Robbe P, Ackroyd S, Burns A, Timbs AT, et al. SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. Blood. 2014;123:1021–31.
39. Shi C, Xie L, Tang Y, Long L, Li J, Hu B, et al. Hypermethylation of N-acetyltransferase 1 is a prognostic biomarker in colon adenocarcinoma. Front Genet. 2019;10:1097.
40. Tiang JM, Butcher NJ, Cullinane C, Humbert PO, Minchin RF. RNAi-mediated knock-down of arylamine N-acetyltransferase-1 expression induces E-cadherin up-regulation and cell-cell contact growth inhibition. PLoS One. 2011;6:e17031.
41. Minchin RF, Butcher NJ. Trimodal distribution of arylamine N-acetyltransferase 1 mRNA in breast cancer tumors: association with overall survival and drug resistance. BMC Genomics. 2018;19:513.
42. McKay JD, Hashibe M, Hung RJ, Wakefield J, Gaborieau V, Szeszenia-Dabrowska N, et al. Sequence variants of NAT1 and NAT2 and other xenometabolic genes and risk of lung and aerodigestive tract cancers in Central Europe. Cancer Epidemiol Biomark Prev. 2008;17:141–7.
43. kConFab Investigators, ABCTB Investigators, EMBRACE Study, GEMO Study Collaborators, Zhang H, Ahearn TU, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet. 2020;52:572–81.
44. Zhang H, Wang A, Tan Y, Wang S, Ma Q, Chen X, et al. NCBP1 promotes the development of lung adenocarcinoma through up-regulation of CUL4B. J Cell Mol Med. 2019;23:6965–77.
45. Wang L, Wrobel JA, Xie L, Li D, Zurlo G, Shen H, et al. Novel RNA-affinity proteogenomics dissects tumor heterogeneity for revealing personalized markers in precision prognosis of cancer. Cell Chem Biol. 2018;25:619–633.e5.
46. Kravtsova-Ivantsiv Y, Goldhirsh G, Ivantsiv A, Ben Itzhak O, Kwon YT, Pikarsky E, et al. Excess of the NF-κB p50 subunit generated by the ubiquitin ligase KPC1 suppresses tumors via PD-L1– and chemokines-mediated mechanisms. Proc Natl Acad Sci U S A. 2020;117:29823–31.
47. Wang P, Deng Y, Yan X, Zhu J, Yin Y, Shu Y, et al. The role of ARID5B in acute lymphoblastic leukemia and beyond. Front Genet. 2020;11:598.
48. Xu H, Zhao X, Bhojwani D, Goodings C, Zhang H, et al. ARID5B influences antimetabolite drug sensitivity and prognosis of acute lymphoblastic leukemia. Clin Cancer Res. 2020;26:256–64.
49. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, et al. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet. 2012;44:511–6.
50. Yang W, Tang H, Zhang Y, Tang X, Zhang J, Sun L, et al. Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. Am J Hum Genet. 2013;92:41–51.
51. Elrefaey S, Massaro MA, Chiocca S, Chiesa F, Ansarin M. HPV in oropharyngeal cancer: the basics to know in clinical practice. Acta Otorhinolaryngol Ital. 2014;34:299–309.
52. Ferreiro-Iglesias A, McKay JD, Brenner N, Virani S, Lesseur C, Gaborieau V, et al. Germline determinants of humoral immune response to HPV-16 protect against oropharyngeal cancer. Nat Commun. 2021;12:5945.
53. Weitzel JN, Chao EC, Nehoray B, Van Tongeren LR, LaDuca H, Blazer KR, et al. Somatic TP53 variants frequently confound germ-line testing results. Genet Med. 2018;20:809–16.
54. Tulstrup M, Soerensen M, Hansen JW, Gillberg L, Needhamsen M, Kaastrup K, et al. TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. Nat Commun. 2021;12:6061.

Publisher's Note