Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset

Sven E. Ojavee[1,2,†,*], Ekaterina S. Maksimova[3,†], Kristi Läll[4], Marie C. Sadler[1,2,5], Reedik Mägi[4], Zoltan Kutalik[1,2,5], Matthew R. Robinson[3,*]

[1] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3] Institute of Science and Technology Austria, Klosterneuburg, Austria.
[4] Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[5] University Center for Primary Care and Public Health, Lausanne, Switzerland

† These authors contributed equally to this work.
*corresponding authors: svenerik.ojavee@unil.ch and matthew.robinson@ist.ac.at

Abstract

Genome-wide association studies seek to attribute disease risk to DNA regions and facilitate subject-specific prediction and patient stratification. For later-life disease, inference from case-control studies is hampered by the uncertainty that control group subjects might later be diagnosed. Time-to-event analysis treats controls as right-censored, making no additional assumptions about future disease occurrence and representing a more sound conceptual alternative for more accurate inference. Here, using data on 11 common cancers from the UK and Estonian Biobank studies, we provide empirical evidence that discovery and genomic prediction are greatly improved by analysing age-at-diagnosis, compared to a case-control model of association. We replicate previous findings from large-scale case-control studies and find an additional 59 previously unreported independent genomic regions, out of which 16 replicated in independent data (an increase of 18% and 6% over current findings). Our novel discoveries provide new insights into underlying cancer pathways, and our model yields a better understanding of the polygenicity and genetic architecture of the 11 tumours. We find that heritable germline genetic variation plays a vital role in cancer occurrence, with risk attributable to many thousands of underlying genomic regions. Finally, we show that Bayesian modelling strategies utilising time-to-event data increase prediction accuracy. As sample size increases, incorporating time-to-event data should be commonplace, improving case-control studies by using richer information about the disease process.

Introduction

Cancer has broad medical importance and a high global health burden, with 19.3 million new cancer cases and almost 10 million cancer deaths occurring in 2020 [1]. Genome-wide association studies (GWAS) aim to attribute risk to regions of the DNA [2] and facilitate polygenic risk score (PRS) calculation [3] to predict subject-specific risk, which may then enable targeted and improved healthcare [4–6]. There is currently evidence for only 450 genomic regions associated with increased risk of 18 common cancers [2], despite recent results showing significant non-zero heritability across a range of cancer occurrences [7]. Current PRS calculated from these findings stratify risk for several cancers, including breast, colon, and prostate cancer, but often add negligible additional predictive information compared to existing non-genomic predictors [8]. Increasing sample size yields increased statistical power for discovery, with extensive recent case-control studies for breast cancer [9], prostate cancer [10, 11], ovarian cancer [12], or testicular cancer [13] showing improved results, but this remains a challenging endeavour. Biobanks provide an additional resource, essential for modern-day medical genetics; however, individuals within these studies have not all reached old age and the number of cancer cases is not high, with recent studies combining biobank cohorts for 18 cancer types [7] to limited effect.

Increased statistical power can also stem from tailored modelling choices, and one factor behind limited predictive performance could be the choice of the genome-wide analysis method. Although most association studies use methods that account for the impact of other genetic regions (fastGWA [14], GMRM [15], BoltLMM [16], REGENIE [17]), it is sometimes still preferred to resort to the basic association testing. In addition, most genome-wide analyses have been performed using a case-control phenotype rather than utilising the cancer diagnosis age as a phenotype, and there is some evidence that analysing data using time-to-event informed methods can have more power for detecting associations [18–21].

Here, we provide empirical evidence using data on 11 common tumours from the UK and Estonian Biobank studies that GWAS discovery and genomic prediction are greatly improved by analysing age-at-diagnosis, compared to a case-control model of association. We extend our recently presented BayesW approach [20], a Bayesian modelling framework that enables joint effect size estimation for time-to-event data, to provide marginal leave-one-chromosome-out mixed-linear age-at-onset adjusted association estimates, in contrast to using Cox mixed model [22] or age-at-onset informed genomic reconstruction of the phenotype [21]. We focus on a re-analysis in the UK Biobank data alone, and we replicate previous findings from large-scale case-control GWAS and find an additional 59 previously unreported independent genomic regions, out of which 16 replicated in independent data (a respective increase of 18% and 6% over current findings). Our novel discoveries provide new insights into underlying cancer pathways, and our model yields a better understanding of the polygenicity and genetic architecture of the 11 tumours. We find substantial SNP-heritability, implying that heritable germline genetic variation plays a vital role in cancer occurrence, with risk attributable to many thousands of underlying genomic regions. Finally, we show that Bayesian modelling strategies that utilise time-to-event data give increased prediction accuracy for all analysed tumours and suggest clinically relevant discrimination rules within the Estonian Biobank study. We argue that it is possible to use existing data more thoughtfully and that a re-analysis of case-control study data exploiting age-at-onset information will lead to novel discoveries and enhanced genomic prediction.

Results

Novel and replicated associations

We analysed data from the UK Biobank on the timing or occurrence of diagnosis of 11 different tumours (see Supplementary Information) using 458,747 individuals of European ancestry and a very weakly LD pruned set of 2,174,071 SNP markers (see Methods and descriptive statistics in Table S1). We present a mixed-linear age-at-onset adjusted association model that mimics the behaviour of recently developed leaving one chromosome out (LOCO) linear mixed models [15, 17] and uses age-at-diagnosis information (GMRM-BayesW, see Methods). We compare the results obtained to a recently proposed more classical case-control information mixed-linear association model that uses the same prior structure for the LOCO adjustment but no age-at-onset information (GMRM-BayesRR-RC [15]). We find that this approach of adjusting the phenotypes with either BayesW or BayesRR-RC predictors results in enhanced statistical power (Figure S1), with GMRM-BayesW resulting in higher power gain as compared to GMRM-BayesRR-RC for most cancer sites. Applying GMRM-BayesW or GMRM-BayesRR-RC to the UK Biobank data, we replicate previously reported findings, with 266 previously identified significant independent trait-marker associations at $p < 5 \cdot 10^{-8}$ (Supplementary data). We also find an additional 59 independent previously unreported variants significant at $p < 5 \cdot 10^{-8}$, of which 16 replicate in independent data of either the Estonian Biobank or previous non-UK Biobank case-control studies (Figure 1a, Table 1, see Methods). The replication analysis demonstrates that the z-scores for 59 previously unreported associations are correlated between the replication and discovery data sets (Fisher's exact test for z-score sign $p = 0.002$, Supplementary figure S3). Furthermore, we observe that the 59 previously undiscovered variants had small but not genome-wide significant p-values in the unadjusted analysis (see Methods), and using GMRM-BayesRR-RC or GMRM-BayesW reduced their p-values below a genome-wide significance threshold (Table S3). We discover novel or replicate previous discoveries slightly better when we account for age-at-onset as compared to the case-control model with 83% and 78% of the markers discovered, respectively, especially for traits that have higher case counts such as breast or prostate cancer (Figure S2).

To map novel associations to functional annotations, we conducted a number of follow-up analyses. SNPs from the previously unreported 16 genomic regions (lead SNPs and those in LD $r^2 > 0.6$, see Methods) lie predominantly in intergenic parts of the genome, with slightly more enrichment than the reference population (enrichment 1.2, $p = 5.0 \cdot 10^{-5}$). We observe substantially higher enrichment in downstream (7.4, $p = 4.3 \cdot 10^{-21}$) and upstream (5.0, $p = 2.4 \cdot 10^{-10}$) regions (the respective values for all significant SNPs are 1.5, $p = 6.9 \cdot 10^{-9}$ and 1.7, $p = 5.4 \cdot 10^{-14}$) (Figure 1b). Majority of SNPs from these novel regions could be linked to regulatory variation. Namely, 11 out of 16 genomic regions contain expression quantitative trait loci (eQTLs) (maximum $p < 1.8 \cdot 10^{-20}$ from FUMA eQTL mapping, see Methods); 6 novel regions have SNPs that fall into RegulomeDB categories [23] that are likely to affect binding, or are linked to expressions of a gene target (Figure 1d). In addition, 5 novel genetic regions include methylation QTLs (mQTLs) associated with other cancers from the Pancan-meQTL database [24]. Moreover, 15 out of 16 novel genomic regions are in open chromatin state in at least 1 of 127 tissue/cell types predicted by ChromHMM [25] (Figure 1d), while for 7 regions, active chromatin state is the most common (Supplementary data).

We confirmed the regulatory effects of novel regions on a wide range of chromatin features using DeepSEA, a deep learning-based model that predicts chromatin effects of sequence variants and priorities regulatory variants [26, 27]. The DeepSEA functionality score median for SNPs in novel genomic regions is 0.18 and one region (index SNP rs804172, associated with prostate cancer) has maximum mean log e-value (MLE) > 2 (Figure 1d), indicating a higher likelihood of regulatory effects than a reference distribution of 1000 Genomes variants. Moreover, 3 novel regions have maximum disease impact score (DIS) > 2 (Figure 1d), highlighting likely disease-associated mutations. We also used the CADD tool that predicts deleterious, functional, and disease causal variants by integrating multiple annotations into one deleteriousness metric [28]. The average CADD score is 4.34, with two novel lead SNPs with CADD score > 12.37, and 9 regions containing SNPs with max CADD score > 12.37 (Figure 1d), a deleteriousness threshold suggested by Kircher et al. [28]. Thus, most novel associations can be attributed to regulatory, intronic, open chromatin functional regions.

To infer whether SNPs from the novel genomic regions impact cancer risk through altering DNA methylation (DNAm) or gene expression levels, we performed Mendelian randomisation (MR) analyses (see Methods). First, we applied single-instrument two-sample summary Mendelian randomisation (SMR) analysis together with heterogeneity in dependent instruments (HEIDI) testing [29] on tissue-specific gene expression data from GTEx (v8) (N = 65-573, European ancestry, [30]) and whole blood-derived eQTLs from eQTLGen (N =31,684, [31]). We could map novel regions to 57 tissue-specific mechanisms involving 29 genes (Figure S5, Supplementary data). For example, *CDC42* expression in muscle skeletal tissue ($\hat{\beta} = 0.033$, $p = 5.2 \cdot 10^{-6}$), colon sigmoid tissue ($\hat{\beta} = 0.039$, $p = 1.0 \cdot 10^{-5}$), and whole blood ($\hat{\beta} = 0.066$, $p = 1.1 \cdot 10^{-11}$) showed a risk-increasing effect on cervical cancer development (Supplementary figure S5). Second, we analysed mQTL data from the GoDMC consortium (N =32,851) [32] using SMR with HEIDI testing (see Methods). The analyses provided evidence for 225 pleiotropic associations between DNAm probes and BCC, cervical and prostate cancers in the 16 novel regions (Bonferroni-correction at $p < 0.05/2380$, HEIDI filtering at $p > 0.05$, Supplementary data). Among these 225 associations, we found that 44 probes overlap with probes linked to 23 other cancers from the Pancan-meQTL database [24]; and 3 of these 44 probes were differentially methylated by means of the same mQTLs (Figure 1c). This colocalisation analysis shows that our novel discoveries, in part, overlap with regions previously found to be methylated in tumour cells, implying that previous methylation differences in tumour cells are driven by germline variation.

We further extended the MR analysis to multivariable MR (MVMR) to jointly estimate whether cancer SNP associations could be mapped to regulatory mechanisms of the scheme DNAm $\rightarrow$ gene expression $\rightarrow$ cancer in a genome-wide screen (see Methods). To maximise statistical power, we used mQTL and eQTL derived from whole blood as provided by the GoDMC and eQTLGen consortia, respectively, and conducted MVMR analyses to quantify mediation between DNAm (exposure E) and cancer traits (outcome Y) through gene expression (mediator M). Among the novel regions, the risk for cervical cancer was found to be increased by increased *CDC42* expression ($M \rightarrow Y$ effect $\widehat{\alpha}_{MY} = 0.040$, $p = 2.4 \cdot 10^{-8}$) through methylation at the cg15582954 probe (chr1:22'470'343; $E \rightarrow Y$ total effect $\widehat{\theta}_T = 0.046$, $p = 3.8 \cdot 10^{-8}$; Supplementary figure S4) corroborating findings from the tissue-specific MR analyses. *CDC42* has been shown to be overexpressed in a number of human cancers and found to be a promising drug target in preclinical studies [33]. The GWAS signal physically locates closer to *WNT4* than *CDC42* (Table 1), however, a putative causal role of

WNT4 in addition to *CDC42* could not be assessed due to a lack of corresponding eQTLs for this gene [31]. Additional DNAm-to-gene expression mechanisms for previously identified cancer regions are listed in Table 2 and include mediation through genes well known to be implicated in cancer such as *PDK1* [34], *LYNX1* [35] and *NEK10* [36].

Finally, we assessed shared molecular functions of these genes and nearest to the novel SNPs genes using FUMA GENE2FUNC [37] software and the KEGG database [38]. The methylation and expression MR analyses identified cancer-specific gene targets whose regulation was changed by SNPs from the novel genomic regions, with a full list of potential novel genes given in Supplementary table S4. The genes were enriched in pathways of linoleic acid metabolism, interferon-gamma signalling, human papillomavirus infection, proteoglycans in cancer, axon guidance and viral carcinogenesis. Besides these pathways, the genes prioritised by FUMA using positional, eQTL, and chromatin mapping showed enrichment in lipids, steroids, and cholesterol metabolism, pathways in cancer, Kaposi sarcoma-associated herpesvirus infection, alcoholic liver disease, chemokine and Rap1 signalling pathway, microRNAs in cancer, as well as interferon signalling and antigen presentations pathways related to MHC-complex. Notable genes driving these discoveries are *KRAS*, *CDC42*, and *WNT4*, part of pan-cancer pathways, and the *FADS* complex associated with metabolism. Moreover, we closer examined the effects of exonic mutations from the novel genomic loci: 7 SNPs map to exons representing 5 nonsynonymous (*HLA-A* (3), *TFAP4*, *ZBED2* genes) and 2 synonymous (*ATG7*, *SLC6A18* genes) substitutions. We mapped 3 of these nonsynonymous mutations to the 3D protein structure of the *HLA-A* complex, where all substitutions fall into alpha-3 domains that form the binding groove that holds a peptide for presentation to CD8+ T-cells (Supplementary figure S11). In summary, our novel findings confirm previous pathways, highlight tumour associated methylation patterns that likely stem from germline variation, and provide additional potential mechanisms through which germline variation can affect cancer risk.

Genetic architecture of 11 cancers

We then aimed at estimating the genetic heritability of the 11 cancers using LD score regression on the marginal associations [39]. When correcting for the discrete nature of the trait (see Methods), the liability-scale heritability estimates were similar or higher and more precise than array-based assays except for non-Hodgkin's lymphoma (Table 3), indicating that heritable genetic variation is a leading risk factor for underlying risk of cancer. The pattern holds even if we use an approach tailored for estimating liability scale heritability for rare traits [40] resulting in slightly more conservative estimates (Table S6). Interestingly, we find that the GMRM-BayesW analysis leads to nominally higher heritability estimates for many cancers than the GMRM-BayesRR-RC estimates, suggesting a better description of genetic architecture when including the timing information in the analysis. The joint Bayesian models for occurrence also enable SNP heritability estimation and comparative inference across cancers of the underlying distribution of genetic effects. The liability scale heritability estimates from the joint Bayesian model are similar to the LD score regression analysis estimates for more prevalent cancers. However, more remarkable differences between the estimates and wider credibility intervals occur for the less prevalent cancers, supporting suggestions [40] that rare traits require extra care as they could be subject to ascertainment bias, sampling bias, and their effective sample size is low. We further used cross-trait LD score regression on the BayesW or BayesRR-RC adjusted marginal associations to estimate the genetic correlation between the traits (see Methods). There is a sizable genetic correlation between melanoma and basal cell carcinoma (BayesW estimate 0.51, 95%CI 0.34-0.68), and we replicate [7] a previous result of negative genetic correlation between endometrial and testicular cancer (BayesW estimate -0.38, 95%CI -0.68 - -0.07) (Supplementary table S5). Interestingly, BayesW-based genetic correlations have a narrower confidence interval than BayesRR-RC based genetic correlation estimates for each significant cancer trait pair.

We find that all traits are highly polygenic, with most of the h^2_{SNP} attributed to SNPs that contribute an average of 0.1% and 0.01% of the group genetic variance for BayesRR-RC and BayesW models, respectively (Figure 2a). We find some differences across cancers, notably melanoma (10%), basal cell carcinoma (13%), breast (8.0%), cervical (5.1%), and prostate (5.4%) cancers; and for age-at-diagnosis of non-Hodgkin's lymphoma (7.0%), bladder (6.0%) and ovarian (5.3%) cancers where at least 5% of the h^2_{SNP} can be attributable to a small number of large effects (mixture 10^{-2}) (Figure 2a). In general, the analysis of

time-to-event phenotypes results in more of the genetic variance assigned to the smallest mixture component (Figure 2a). The result is in line with the number of LD-independent regions required to explain a proportion of the SNP heritability, where time-to-event analysis results in a more polygenic architecture compared to the case-control analysis (Figure 2b). Each curve reaches a plateau with 80-90% of the genetic variance attributable to a small number of genomic regions, and the remaining 10-20% attributable to 10,000 to 20,000 LD-independent regions. The number of remaining regions required to capture all of the association signals varies greatly across cancers, from 13,600 for ovarian and testicular cancer to 22,500 for basal cell carcinoma (Figure 2b). Additionally, we find that the 11 cancers differ in how rare and common variants contribute to the SNP heritability (Figure 2c). We further observe that genetic variance often positively correlates with variants' MAF structure. For example, the largest proportion of genetic variance is consistently attributable to common variants in the fourth MAF quartile for both time-to-event (TTE) or case-control (CC) models on basal cell carcinoma (TTE 66%, CC 68%), melanoma (TTE 32%, CC 36%), breast (TTE 44%, CC 70%), colon (TTE 43%, CC 36%), and prostate cancers (TTE 40%, CC 57%) (Figure 2c). In contrast, testicular cancer, non-Hodgkin's lymphoma and ovarian cancer have 61%, 54%, 63% of the genetic variance explained by the rarest effects in the first MAF quartile according to the case-control model. Thus, our MAF-LD stratified h^2_{SNP} estimation approach suggests: (i) strong differences in the underlying genetic architecture across these 11 cancers, (ii) that only a limited number of genomic regions are required to capture most of the risk for all cancers, and (iii) that mapping further associations will be extremely difficult as a small amount of variance is attributable to a large number of independent regions of the DNA.

Genomic prediction of 11 cancers

Next, we used BayesW SNP marker estimates to predict the occurrence of cancer in 195,432 individuals within the Estonian Biobank data (Figure 3, see Methods). We compared our results to those obtained by a baseline BayesRR-RC model, and also to SNP marker effect estimates obtained by Rashkin et al. [7] which combine the UK Biobank (408,786 European ancestry individuals; 48,961 cancer cases) and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging cohorts (66,526 European ancestry individuals; 16,001 cancer cases) to obtain a larger sample size. Additionally, we also compare our prediction estimates to the results obtained by Kachuri et al. [41] where they combined GWAS summary statistics across many cancer studies to achieve a much larger sample size. We provide a comparison of applying Bayesian models on either self-reported or medical record data (see Methods), showing that medical record data models outperform models using self-reported data. Thus, we resorted to using only medical record data (Figure S10), which illustrates the importance of high data quality and accurate measurement to facilitate phenotype linking across studies.

We find that Bayesian models for individual-level data, especially those utilising age-at-onset information, yield substantially improved genomic prediction for cancer occurrence, and the benefit is amplified as case count increases. Except for a few cancers for which none of the models gives significantly useful predictors, we find that conducting the analysis using an age-at-onset phenotype (BayesW) yields a nominally higher odds ratio (of having one standard deviation higher PRS) than a case-control phenotype (BayesRR-RC) for basal cell carcinoma (BayesW: 1.66, BayesRR-RC: 1.65), bladder cancer (BayesW: 1.36, BayesRR-RC: 1.24), colon cancer (BayesW: 1.19, BayesRR-RC: 1.16), melanoma (BayesW: 1.29, BayesRR-RC: 1.25), non-Hodgkin's lymphoma (BayesW: 1.19, BayesRR-RC: 1.13), prostate cancer (BayesW: 1.86, BayesRR-RC: 1.85), and testicular cancer (BayesW: 1.55, BayesRR-RC: 1.44) (Figure 3a). A similar trend can be observed when using C-statistic or hazards ratio for comparison (Supplementary figure S9, see Methods). At least one of the Bayesian methods consistently outperforms previous analyses with a larger sample size that combine the biobank cohorts of the UK Biobank and Kaiser Permanente studies (UKB-KP) (see Methods). More notably, UKB-KP score has at least nominally smaller odds ratios (of standard deviation PRS difference) for testicular cancer (UKB-KP: 1.30), prostate cancer (UKB-KP: 1.51), breast cancer (UKB-KP: 1.23) and non-Hodgkin's lymphoma (BayesW: 1.19, BayesR: 1.12, UKB-KP: 1.04). For the latter case of non-Hodgkin's lymphoma, the UKB-KP score did not yield a significantly predictive score. The PRSs suggested by Kachuri et al. [41] combining multiple studies from a far more significant number of underlying cancer cases (Total marginal), yield mostly similar results to the BayesW and BayesRR-RC predictors using only UK Biobank data (Figure 3, Supplementary figure S9). For example, for breast cancer, where the previous study had 119,000 total

cancer cases [9] as compared to our 17,000 cases (Supplementary table S1), we achieved similar odds ratios for standard deviation PRS change (BayesRR-RC: 1.59, Total marginal: 1.61). Bayesian scores outperform Total marginal PRS for prostate cancer (Total marginal: 1.77) and cervical cancer (Total marginal: 0.96), and Bayesian scores offer a marginal improvement for melanoma (Total marginal: 1.23) and non-Hodgkin's lymphoma (Total marginal: 1.13). Testicular cancer stands out as both Bayesian scores perform noticeably worse compared to the Total marginal estimate, but here, our analysis resorted to only 886 cases whereas the previous risk score is combining more than 10,000 cases [13, 42] and thus the power in the UK Biobank in likely simply too low.

We observe that the highest 5% PRS quantile discriminates well for the disease occurrence (Figure 3c). Whereas the risk to develop prostate cancer by age 85 is estimated to be 11% (Supplementary table S7) among the top 5% highest PRS individuals, nearly 55% will develop prostate cancer according to the BayesRR-RC model. In comparison, UKB-KP PRS finds that 46% of the top 5% PRS develop prostate cancer. The top 5% polygenic risk score yields a useful discrimination rule for most other cancers as well, notably for breast cancer for which 19% of the top 5% BayesW PRS gets diagnosed with by age 85 (12% in the population, Supplementary table S7) and basal cell carcinoma for which 39% of the top 5% BayesRR-RC PRS gets diagnosed by age 85 (31% in the population, Supplementary table S7). The share of individuals getting a cancer diagnosis before age 50 is disproportionately higher among individuals with top 5% or 10% of the PRS across many cancers and risk score types (Figure 3b). For example, 23% out of all basal cell carcinoma cases and 30% out of all prostate cancer cases have the top 10% highest genetic risk according to BayesW risk score suggesting that the BayesW risk score discriminates well the early onset of prostate cancer or basal cell carcinoma. Our results suggest that the top 5% highest Bayesian polygenic risk scores could serve as a rule to detect individuals who should not only receive earlier communication about their risks but it could also result in a cost-effective screening model for this subset of individuals.

Discussion

Our results demonstrate the advantages of using joint Bayesian modelling and age-at-onset phenotypes for genomic prediction and GWAS discovery, highlighting how these approaches can be used to improve utilisation of existing data. Biobanks are becoming increasingly common worldwide, size and numbers of biobanks are increasing and improved links to electronic health record data enable information to be obtained regarding the age-at-diagnosis. Thus, we expect that our approach of incorporating age-at-onset data in the analysis will become commonplace, improving case-control studies by using richer information about the disease process.

One of the fundamental problems of analysing cancer phenotypes in a case-control fashion is the uncertainty that the control group subjects might later be diagnosed with cancer. Many cancers often become more prevalent only at later ages (Figure 3c), and as biobanks primarily consist of young, healthy individuals, it could distort the inference. That issue can be mitigated, for example, by introducing age thresholds to eliminate younger individuals who have been at risk only for a limited amount of time, or by age-matching individuals. However, this will always be somewhat arbitrary, it would reduce the sample size, and there is no guarantee that older individuals would not develop cancer in their later life. In contrast, time-to-event analysis treats these individuals as right censored, making no additional assumptions about the cancer occurrence in future. Therefore, time-to-event analysis suggests an alternative with a more sound conceptual background to yield more accurate inference. Interestingly, time-to-event adjustment tends to yield higher power than the case-control adjustment once the case count is high enough. Hence, time-to-event analyses could become more statistically powerful for future studies with potentially many added cases than their case-control counterparts at a high fixed case count.

Our PRS results remain limited as our work represents a re-analysis of a single biobank study, with the aim of demonstrating the methodological improvements that can be obtained. However, there is nothing preventing BayesW being run across different studies and posterior mean SNP effects being combined to improve the effectiveness of the PRS, providing predictors with the potential to stratify individuals for screening programs. For example, prostate cancer screening has been found to be only moderately useful for the general population with 17-40% [43, 44] reduction in cancer-specific deaths but as the mortality rate

is low (in USA stage I-III 5-year survival rate >95%, stage IV 5-year survival rate 30% [45]) and as there are potential complications following the treatment, a general screening program has not been commonly implemented. Nevertheless, there are recommendations for stratified risk communication. For example, the American Cancer Society suggests that men with a first-degree relative with prostate cancer before age 65 should be informed about screening and its risks already at age 45, and men with multiple relatives with prostate cancer before 65 should be informed about screening even at age 40 [46]. Moreover, it has been found that even if screening is not cost-effective for men at average risk of prostate cancer, it is still cost-effective for men at very high risk (five times higher risk than the average) [44]. Our results suggest that the top 5% highest Bayesian polygenic risk scores could serve as a rule to detect those who should be screened and whose risk should be communicated.

There are important caveats to this study. Firstly, the discovery and training set of our study is limited to UK Biobank individuals with European ancestry whereas many other recent studies that have rather focused on merging and meta-analysing multiple data sets from various backgrounds. However, we replicate our findings on an independent data set and despite having a smaller discovery set we show that there are discoveries yet to be made on existing data set simply by using enhanced methodology for timing-related traits rather than occurrence-related traits. Secondly, the number of cancer cases is very low for some of our cancers such as testicular, ovarian or endometrial cancer leading to sub-optimal prediction accuracy, while cancers with higher case counts (prostate, breast) yielded good prediction accuracy. Hence, in future these analysis should be replicated with greater case counts for the cancers with smaller case counts. Thirdly, our current analysis combines both prevalent and incident cases to maximise the statistical power. However, it has been shown [7] that effect sizes are in general very similar even if we restrict the analysis only to incident cases. Future time-to-event analyses could also benefit from using information about left truncation by including entry date to the analysis, although the gain might be marginal as long as the phenotypic information is derived from the medical records and the onset happens later in life.

In summary, we have shown random effect models, especially those which utilise time-to-event data, maximise the use of existing data, for h^2_{SNP} estimation, genomic prediction and GWAS discovery of 11 common cancers.

Methods

UK Biobank Data

We restricted our discovery analysis in the UK Biobank to a sample of European-ancestry individuals. To infer ancestry, we used both self-reported ethnic background (UK Biobank field 21000-0) selecting coding 1 and genetic ethnicity (UK Biobank field 22006-0) selecting coding 1. We also took the 488,377 genotyped participants and projected them onto the first two genotypic principal components (PC) calculated from 2,504 individuals of the 1,000 Genomes project with known ancestries. Using the obtained PC loadings, we then assigned each participant to the closest population in the 1000 Genomes data: European, African, East-Asian, South-Asian or Admixed, selecting individuals with PC1 projection < absolute value 4 and PC 2 projection < absolute value 3. Samples were also excluded if in the UK Biobank quality control procedures they (i) were identified as extreme heterozygosity or missing genotype outliers; (ii) had a genetically inferred gender that did not match the self-reported gender; (iii) were identified to have putative sex chromosome aneuploidy; (iv) were excluded from kinship inference; (v) had withdrawn their consent for their data to be used. We used genotype probabilities from version 3 of the imputed autosomal genotype data provided by the UK Biobank to hard-call the genotypes for variants with an imputation quality score above 0.3. The hard-call-threshold was 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p-value for Hardy-Weinberg test larger than 10^{-6}, as determined in the set of unrelated Europeans) we selected those with minor allele frequency (MAF) > 0.0002 and rs identifier, in the set of European-ancestry participants, providing a data set 9,144,511 SNPs. From this we took the overlap with the Estonian Biobank data described below to give a final set of 8,430,446 markers. This provides a set of high quality SNP markers present across both discovery and prediction data sets. For computational convenience when conducting the joint Bayesian analysis we created an additional subset of

markers by removing markers in very high LD, through the selection of the highest MAF marker from any set of markers with LD $R^2 \geq 0.8$ within a 1Mb window. These filters resulted in a data set with 458,747 individuals and 2,174,071 markers.

We used the recorded measures of individuals to generate the phenotypic data sets for 11 most common types of cancer: bladder, breast, cervix, colon, endometrium, ovary, prostate, testis, basal cell carcinoma, melanoma, and non-Hodgkin's lymphoma. Then, we created time-to-event phenotypes using either self-reported data or the linked electronic medical records data. For the medical record data, we used UK Biobank field 40008 to get the earliest age at each cancer diagnosis together with fields 40006 and 40013 to indicate the ICD10 or ICD9 cancer type (Table S2). Individuals without an entry on those fields were considered censored and their time was set to their last known age alive (exact birth date imputed to day 15 of a month as only month and year are known) without a cancer diagnosis. Each individual i was therefore assigned a censoring indicator C_i that was defined $C_i = 1$ if the person had the event before the end of the follow-up period and $C_i = 0$ otherwise. For self-reported time-to-event phenotypes, we created a pair of last known time (averaged between assessments) without an event and censoring indicator C_i. Similarly to the medical record phenotypes, if the event had not happened, then the last time without having the event was defined as the last date of assessment centre or date of death visit minus date of birth. For creating the self-reported phenotypes, we used UK Biobank field 20001 for the presence or absence of certain cancer type and UK Biobank field 20007 for interpolated ages of individuals when the disease was first diagnosed; we excluded from the self-reported phenotype analysis individuals who said that they had cancer, but there was no record of the diagnosis age. In an attempt to further increase power and to account for potential missingness, for each individual who had self-reported data about cancer timing but no medical record data, we used used the self-reported age-at-diagnosis instead of treating the individual as censored. However, this approach only yielded marginal improvements compared to using purely medical record information (Supplementary figure S10). Finally, the case-control-phenotypes corresponding to the time-to-event phenotypes were defined as the censoring indicators C_i.

The analyses were adjusted for the following covariates: sex for sex-unspecific cancers, age in case-control analyses, UK Biobank recruitment center, home location, genotype batch and 20 first principal components (UK Biobank field 22009) to account for the population stratification in a standard way. For the analyses that used age-at-diagnosis as phenotypes we did not include any covariates of age or year of birth because these are directly associated to our phenotypes.

Estonian Biobank Data

For the Estonian Biobank data, 195,432 individuals were genotyped on Illumina Global Screening (GSA) arrays and we imputed the data set to an Estonian reference, created from the whole genome sequence data of 2,244 participants [47]. From 11,130,313 markers with imputation quality score > 0.3, we selected SNPs that overlapped with those selected in the UK Biobank, resulting in a set of 8,430,446 variants out of which 2,174,071 variants were used in the prediction analysis. The 60 previously unreported variants that were found significant in the marginal association analysis of UK Biobank (Table S3) were used in a replication analysis using the same Estonian Biobank individuals.

We created the phenotypes for the Estonian Biobank individuals using the respective medical record information. The occurrence of each of the cancers was defined by using the respective ICD10 codes exactly as it was defined for the UK Biobank medical record phenotypes (Supplementary table S2) by first defining the last known time person did not have a respective diagnosis. Individuals with a respective cancer diagnosis received a censoring indicator $C_i = 1$ and 0 otherwise that then defined the case-control phenotypes adjusted for covariates such as sex for sex-unspecific cancers and age.

Analysis with joint Bayesian models

We estimated the hyperparameters such as genetic variance and prior inclusion probability by grouping markers into MAF-LD bins as recent theory suggests this yields improved estimation [48–50]. We ran the BayesW model on the UK Biobank data with 8 MAF-LD groups that were defined as first splitting markers by MAF quartiles and then splitting each of those MAF quartiles into two LD score blocks (MAF quartiles

are 0.007, 0.020, 0.102; median LD score in each quartile are 2.32, 3.33, 5.69, 9.25, from the lowest MAF quartile respectively). We decided not to split these groups further as the potentially low statistical power of cancer-related phenotypes could lead to many groups with zero genetic variance. Both BayesW and BayesRR-RC models were executed with mixture components (0.0001, 0.001, 0.01, 0.1) for each of the groups, reflecting that the markers allocated into those mixtures explain the magnitude of 0.01%, 0.1%, 1% or 10% of the group-specific genetic variance. We ran the BayesW model using the timing of cancers as the phenotype while treating individuals without cancer as censored, and we ran the BayesRR-RC type model using the occurrence of cancer as the phenotype. In the BayesW analyses we took the covariates into account by estimating the effects of the covariates within the BayesW model while in the BayesRR-RC we regressed out the covariates from the phenotype prior to the analysis.

We specified the hyperparameters for the models such that they would be weakly informative. For BayesW model, the choice of hyperparameters and quadrature points was exactly the same as in [20]; for BayesRR-RC model the choice was exactly the same in [50]. We ran the chains for each of the cancer types twice for 6000 iterations, discarding the first 2000 iterations as a burn-in and using a thinning step of 5, leaving us with a final of 1600 samples of the posterior distribution. As estimation is done in parallel, the run time will depend on the degree of parallelisation. For example, for basal cell carcinoma we used 11 nodes and 12 tasks per node (total 132 tasks) for BayesW and 7 nodes and 12 tasks per node (total 84 tasks) for BayesRR-RC. This resulted in a total run time of 67.5 hours (40.5s per iteration) for BayesW and 79.7 hours (47.8s per iteration) for BayesRR-RC. Although BayesW was faster in the absolute time, adjusted for the number of tasks in the example, BayesW requires 33% more time per iteration than BayesRR-RC. Other choices for parallelisation (for example synchronisation rate) were set the same as described in [20].

Association testing with adjusted marginal models

For association testing, we used binary case-control phenotypes of cancer occurrences as a baseline. From those binary phenotypes we regressed out sex, (if applicable for the cancer), age, UK Biobank assessment centre, coordinates of birth place, genotype chip, and the leading 20 PCs of the SNP data. In addition, when testing each chromosome we regressed out from the phenotype the genetic effects of all the other chromosomes and the genetic effects were calculated using either the BayesRR-RC or BayesW models. That is, for every chromosome k, the phenotype was defined as

$$\tilde{\mathbf{y}}_k = \tilde{\mathbf{y}} - \sum_{l \neq k} \hat{\mathbf{g}}_l, \tag{1}$$

where $\mathbf{g}_l = \mathbf{X}^l_{UK}\bar{\boldsymbol{\beta}}_l$, $\mathbf{X}^l_{UK} : N \times M_l$ matrix of SNPs in the lth chromosome, $\bar{\boldsymbol{\beta}}_l$ is the vector of average effect sizes from joint Bayesian analysis in chromosome l, $\tilde{\mathbf{y}}$ is the binary phenotype that has been adjusted for covariates. Then, at each chromosome k we fitted the linear model for every marker j

$$\tilde{\mathbf{y}}_k = \mathbf{X}^k_{UK_j}\beta_j + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{X}^k_{UK_j}$ is the vector of jth marker values, β_j is the jth SNP effect that we are estimating, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2_e I_N)$ is the error term with an error variance σ^2_e. Therefore, we estimated the effect size and standard error for each of the SNPs and we tested the significance using two-sided Wald test with a null hypothesis of $\beta_j = 0$. For comparison, we also estimated the conventional unadjusted models where we did not adjust the phenotype for the genetic effects of other chromosomes, that is

$$\tilde{\mathbf{y}} = \mathbf{X}_{UK_j}\beta_j + \boldsymbol{\varepsilon}. \tag{3}$$

We used the full overlap of UK Biobank and Estonian Biobank markers giving us a total of 8,433,421 markers to be analysed. We used p-value threshold of $5 \cdot 10^{-8}$ to determine the significance of each marker.

We applied the following steps on the association results to filter out independent and potentially previously undiscovered markers. Firstly, we LD clumped the results such that the index SNPs would have a p-value below $5 \cdot 10^{-8}$ and SNPs could be added to a clump if they were 1Mb from the index SNP, they were correlated with $r^2 > 0.05$ and they were nominally significant ($P < 0.05$). Next, we used COJO method [51] from

GCTA software [52] to find clumps with independent signals by conducting stepwise selection of index SNPs in 1Mb window and we considered SNPs independent if they had a p-value below $5 \cdot 10^{-8}$ in the joint model. To determine novelty, we first removed all markers that were significantly associated in the unadjusted model (eq. 3). We then removed all the markers that had a correlation of $r^2 > 0.1$ with a marker that had been previously found associated with a cancer of interest using GWAS Catalog (published until July 2021) and LDtrait tool with the British in England and Scotland population. We then again used COJO to condition the remaining markers on the previously identified associations for each cancer of interest and SNPs that did not fall below $5 \cdot 10^{-8}$ in the joint model were eliminated. For the remaining SNPs we conducted an additional literature review using Phenoscanner database [53, 54] to find any previous associations with variants of interest or variants in LD. The remaining candidates of novel associations were concatenated across BayesW or BayesRR-RC adjusted analyses and then included in the replication analysis using the largest available studies conducted for each specific cancer type. We were able to use studies with around 80,000 prostate cancer cases [11], 120,000 breast cancer cases [9]. We also checked findings for replication in the Estonian Biobank. Replication was defined as Benjamini-Hochberg corrected p-value being lower than 0.05 and the direction of the effect size same in both the original analysis and the replication analysis.

Liability scale heritability and genetic correlation

We used the summary statistics from the marginal association analysis in LD score regression [39] to calculate the observed scale heritability. We used the LD scores from the 1000 Genomes European data `https://alkesgroup.broadinstitute.org/LDSCORE/` and the summary statistics were taken from either BayesW-adjusted or BayesRR-RC-adjusted association analysis. The conversion of the heritability to the liability scale was done using the formula by Lee et al. [55] (Table 3) and using the risks from SEER 2016-2018 (Table S7) [56] using the risks of having cancer diagnoses between ages 0 to 85 for non-hispanic white people providing similarity with the study population (European ancestry, UK Biobank, oldest person age 86). We further provide an alternative liability scale transformation [40] designed for rare traits. Using the alternative rare trait liability scale transformation we also present the heritability estimates from the joint Bayesian case-control model (Table S6). In addition, we used cross-trait LD score regression [57] to calculate the genetic correlations using results from BayesW or BayesRR-RC adjusted analyses, again using LD scores from the 1000 Genomes European data.

Discovery follow-up analyses

We conducted a number of follow-up analyses using the mixed-linear age-at-onset adjusted association model (GMRM-BayesW) and the baseline mixed-linear age-at-onset adjusted association model summary statistics.

Functional annotation analysis

We used FUMA (Functional Mapping and Annotation) [37] platform to functionally characterise novel replicated variants and prioritise genes. We defined a threshold for independent significant novel SNPs and corresponding novel genetic regions as LD $r^2 = 0.6$ on the reference panel UKB/release2b. When performing gene mapping, we used 10kb maximum distance for positional mapping, all available tissue types and maximum p-value threshold of $5 \cdot 10^{-8}$ for eQTL mapping, and builtin chromatin interaction data with $1 \cdot 10^{-6}$ FDR threshold for chromatin interaction mapping.

We annotated SNPs by function and identified nearest genes using ANNOVAR [58], performing a two-sided Fisher's exact test for quantification of functional classes enrichment. We obtained RegulomeDB categorical score [23], eQTL information, 15-core chromatin state [25], and CADD deleteriousness score [28] from SNP2GENE analysis of FUMA. We considered markers as deleterious if their CADD score exceeded the 12.37 threshold suggested by Kircher et al. (2014) [28] and chromatin state as open/active for SNPs with 15-core chromatin score $\leq$7 predicted by ChromHMM [25] based on 5 chromatin marks for 127 epigenomes. We annotated the SNPs with DeepSEA, a deep learning-based model that predicts chromatin effects of sequence variants and priorities regulatory variants, and we report here the functional significance score [26], maximum mean log e-value (MLE), and disease impact score (DIS) [27]. We calculated all of the mentioned parameters

for each significant independent novel SNP as well as minimum/maximum/common values within the novel genetic regions.

We visualised the location of the exonic SNPs on a 3D protein structure by identifying amino acid substitutions in the gnomAD database [59], aligning proteins using Protein BLAST [60], and visualising with iCn3D Structure Viewer [61].

Mendelian randomisation analysis

We applied two-sample summary-data-based Mendelian Randomisation (SMR) and heterogeneity in dependent instruments (HEIDI) method [29] to assess genetic colocalisation between the 16 novel regions identified by our mixed-linear model BayesW analyses and DNA methylation (DNAm) as well as tissue-specific gene expression. DNAm levels were instrumented by mQTL data derived from whole blood provided by the GoDMC consortium (N =32,851, [32]). Tissue-specific eQTL data came from the GTEx consortium (v8) for 49 tissue types (N = 65-573, European ancestry, [30]). Additionally, we used whole blood-derived eQTLs from the eQTLGen consortium (N =31,684, [31]). We selected *cis*-m/eQTLs (< 1Mb of associated probe, $p < 1 \cdot 10^{-6}$) from the novel regions (if available in the investigated tissue) and conducted SMR-HEIDI tests using the top m/eQTL for each DNAm/gene probe, respectively. Given that most tissue-specific gene expression levels cannot be instrumented by multiple independent instrumental variables (IVs), the SMR-HEIDI approach was chosen over other MR methods that need multiple IVs to test for robust causal associations [62]. SMR outputs were also filtered based on a stringent HEIDI threshold of $p > 0.05$ [29]. Significance of gene-cancer SMR pleiotropic associations was defined at a Bonferroni-corrected threshold accounting for 1,825 ($p < 2.7 \cdot 10^{-5}$) tests summing across different tissues and instruments.

We further conducted multivariable MR (MVMR) to dissect significant DNAm-to-cancer causal effects (θ_T) into direct (θ_D) and indirect effects through transcript levels following the methodology outlined in [63]. While the previous SMR analysis was restricted to the 16 novel regions, the MVMR framework was applied on genome-wide GMRM-BayesW estimates across the 11 cancers. MVMR necessitates multiple IVs and we based the analysis on mQTLs and eQTLs from the GoDMC and eQTLGen consortium, respectively.

First, we conducted univariable inverse-variance weighting MR (MR-IVW) analyses for every exposure E DNAm probe with at least 5 near-independent instrumental variables ($r^2 < 0.05$, $p < 1 \cdot 10^{-6}$, < 1 Mb from DNAm probe, ∼50,000 DNAm probes) accounting for correlated instruments [64] to obtain DNAm-to-trait total causal effect estimates ($E \rightarrow$ outcome Y effect ($\widehat{\theta_T}$)). DNAm-trait pairs with an associated Bonferroni-corrected significant causal effect ($p_T < 0.05/50{,}000 = 1 \cdot 10^{-6}$) were retained and distance-pruned (> 1 Mb) based on p_T to be independent of each other.

Second, MVMR analyses were performed to estimate the direct effect $\hat{\theta}_D$ by including transcript mediators (M) and their associated genetic instruments ($p < 1 \cdot 10^{-6}$). Transcripts were required to be in *cis* (< 500kb away from the DNAm probe) and causally associated to the DNAm probe. This latter condition was verified by calculating univariable MR-IVW effects from the DNAm probe on the transcript to estimate a causal effect $\widehat{\alpha}_{EM}$ and associated p-value p_{EM} (significance was defined at $p_{EM} < 0.01$). Analogously to the total causal effect estimation, direct effects $\widehat{\theta}_D$ were then calculated in an MVMR regression accounting for correlation among IVs [65]. The mediation proportion ($\widehat{MP}$) was then estimated as $1 - \widehat{\theta}_D/\widehat{\theta}_T$. Causal effect estimates from the transcript mediator on the outcome trait ($\widehat{\alpha}_{MY}$) were obtained from univariable MR-IVW analyses.

To ascertain that MVMR estimates did not suffer from heterogeneity, which could point towards horizontal pleiotropy, we computed heterogeneity tests based on Cochran's Q-statistic [66]. Homogeneity within the genetic instrument set was assured at $p_{HET} > 0.01$.

We compared the significantly associated methylation probes from the novel regions with the Pancan-meQTL database [24] which provides meQTLs across 23 cancer types from The Cancer Genome Atlas. For this comparison, we used the probes identified by SMR (2,380 tests). The significance threshold was adjusted for the total number of tests ($p < 0.05/2{,}380 = 2.1 \cdot 10^{-5}$).

For pathway analyses, we derived a list of genes whose differential expression was found to be associated with cancer risk based on the eQTL and mQTL MR analyses. For this, we combined results from the following analyses: single-instrument SMR on GTEx (v8) (1680 tests), eQTLGen (145 tests), and GoDMC (2380 tests) datasets and multi-instrument MR-IVW on eQTLGen (61 tests) and GoDMC (998 tests) datasets. The

overall significance threshold on combined data was Bonferroni-corrected $p < 0.05/5264 = 9.5 \cdot 10^{-6}$. We used both the cis- and trans-meQTL datasets across 23 cancer types from the Pancan-meQTL database to identify methylation-related genes whose expression is affected by the SNPs from the novel regions. Finally, we extended the resulting list of genes by adding the nearest to the novel SNPs genes identified by ANNOVAR and protein coding genes from FUMA positional, eQTL, and chromatin interaction mapping. We then tested pathway enrichment with FUMA GENE2FUNC software [37] and KEGG Mapper [38].

Genomic prediction in the Estonian Biobank

The predictors based on BayesW or BayesRR-RC models into Estonian Biobank $\hat{\mathbf{g}}$ were obtained by multiplying the standardised genotype matrix with the average SNP effect across iterations

$$\hat{\mathbf{g}} = \mathbf{X}_{Est}\hat{\boldsymbol{\beta}}\mathbf{1} = \mathbf{X}_{Est}\bar{\boldsymbol{\beta}}, \tag{4}$$

where $\mathbf{X}_{Est}$ is $N_{Est} \times M$ matrix of standardised Estonian genotypes (each column is standardised using the mean and the standard deviation of the Estonian data), $\hat{\boldsymbol{\beta}}$ is the $M \times I$ matrix containing the posterior distributions for M marker effect sizes across I iterations, $\bar{\boldsymbol{\beta}}$ is the average SNP effect. We calculated the average predictor from BayesW and BayesRR-RC models using for each cancer using 1600 iterations (see Analysis with joint Bayesian models). We compared our Bayesian risk scores with the ones provided by Kachuri et al. [41] (Total marginal) and the ones provided by Rashkin et al. [7] (UK Biobank - Kaiser Permanente). Total marginal estimate combines summary statistics from several studies (total 543 SNPS were used in total for 11 cancer scores, each score having an individual subset of SNPs) and they calculate the PRS using inverse variance weights that showed the best predictive performance. The PRS was thus calculated as the sum of standardised SNP dosage values in Estonian Biobank time inverse variance weighted SNP effects. Using the same idea of inverse variance weights, we combined 314 SNPs with weights from [7] to also get PRS for each Estonian Biobank individual.

We evaluated the performance of the 4 types of genetic predictors for each cancer phenotype by comparing them to the true phenotype case-control status using logistic regression and true phenotype timing using Cox proportional hazards (PH) model. The 4 predictors were compared using the top 5% PRS with the rest, the top 10% PRS with the rest and the comparing the effect of one standard deviation increase in PRS. From the logistic regression we calculated odds ratios for the top 5%, top 10% and scaled change effect. From the Cox PH model we calculated hazards ratios and Harrell's C-statistics [67] for the top 5%, top 10% and scaled change effect. In addition to the predictor, gender (if applicable) and age-at-entry were included in the logistic regression and Cox PH model that was calculating the hazards ratio. Harrell's C-statistic was calculated from the Cox PH model where the true phenotype was regressed only one the predictor. The results of odds ratios, hazards ratios and Harrell's C-statistics are shown in Figure 3a and Supplementary figure S9. We further used the top 5% and top 10% PRS individuals to see what percentage of them develop a cancer (Figure 3b).

Across all the cancers and 4 predictive scores we calculated the respective cumulative incidence curves for the top 5% highest PRS individuals (Figure 3c) adjusting the analysis for the competing risks. The calculation was done using R package cmprsk [68, 69].

Author contributions

SEO and MRR conceived and designed the study. MRR, SEO, ESM designed the study with contributions from KL and MCS. SEO, ESM, KL, MCS and MRR contributed to the analysis. SEO, ESM, MCS and MRR wrote the paper. KL, ZK, RM provided study oversight, contributed data and ran computer code for the analysis. All authors approved the final manuscript prior to submission

Author competing interests

The authors declare no competing interests.

Data availability

This project uses UK Biobank data under project 35520. UK Biobank genotypic and phenotypic data is available through a formal request at (`http://www.ukbiobank.ac.uk`). The UK Biobank has ethics approval from the North West Multi-centre Research Ethics Committee (MREC). For access to be granted to the Estonian Biobank genotypic and corresponding phenotypic data, a preliminary application must be presented to the oversight committee, who must first approve the project, ethics permission must then be obtained from the Estonian Committee on Bioethics and Human Research, and finally a full project must be submitted and approved by the Estonian Biobank. This project was granted ethics approval by the Estonian Committee on Bioethics and Human Research (`https://genomics.ut.ee/en/biobank.ee/data-access`).

Code availability

The BayesW model was executed with the software Hydra, with full open source code available at `https://github.com/medical-genomics-group/hydra` [70]. Summary MR-HEIDI tests were conducted using the SMR software (version 1.03) [29]. The multivariable MR analyses were carried out with SMR-IVW extension software `https://github.com/masadler/smrivw`. plink version 1.9 is available at `https://www.cog-genomics.org/plink/`. R version 4.0.3 is available at `https://www.r-project.org/`.

Acknowledgements

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria and the University of Lausanne.

K.L. and R.M. were supported by the Estonian Research Council grant PRG687.

Estonian Biobank computations were performed in the High Performance Computing Centre, University of Tartu.

We would like to acknowledge the participants and investigators of UK Biobank and Estonian Biobank studies.

SNP	Cancer	Chr	Position	Nearest gene	Eff/ Oth	MAF	Effect size	Model	p-value
rs8154	Basal cell carcinoma	3	11596302	ATG7	C/T	0.32	-0.013	BW,BRR-RC	3.31×10^{-8}
rs140813608	Basal cell carcinoma	3	181981570	RP11-338L18.1*	A/G	0.018	0.045	BW,BRR-RC	4.00×10^{-8}
rs174570	Basal cell carcinoma	11	61597212	FADS2	T/C	0.128	-0.021	BW,BRR-RC	1.16×10^{-10}
rs113537654	Basal cell carcinoma	16	68939047	TANGO6	C/T	0.179	-0.016	BW	4.09×10^{-8}
rs35661976	Breast cancer	5	1243699	SLC6A18	T/C	0.02	0.062	BW	1.71×10^{-9}
rs10842546	Breast cancer	12	25525513	RNU4-67P*	A/G	0.49	-0.017	BW	1.28×10^{-8}
rs4654783	Cervical cancer	1	22439520	WNT4*	T/C	0.299	0.022	BW,BRR-RC	2.56×10^{-12}
rs111842003	Ovarian cancer	3	111306354	CD96	A/T	0.005	0.12	BRR-RC	6.12×10^{-9}
rs12142017	Prostate cancer	1	154980341	ZBTB7B	C/T	0.075	-0.033	BW	1.83×10^{-8}
rs68003823	Prostate cancer	4	74325956	AFP*	G/A	0.137	-0.026	BW	1.53×10^{-8}
rs2523761	Prostate cancer	6	29818726	MICF*	G/A	0.202	-0.024	BW,BRR-RC	9.14×10^{-10}
rs3829734	Prostate cancer	12	48379810	COL2A1	A/G	0.317	0.022	BW,BRR-RC	1.07×10^{-10}
rs9543212	Prostate cancer	13	73624100	KLF5*	T/C	0.304	0.019	BW	3.25×10^{-8}
rs804172	Prostate cancer	16	4349751	GLIS2*	C/G	0.477	-0.018	BW,BRR-RC	5.90×10^{-9}
rs35539606	Prostate cancer	18	76709587	CTD-2382H12.1*	T/C	0.152	0.026	BW,BRR-RC	2.89×10^{-9}
rs73140002	Prostate cancer	20	52175811	RP4-724E16.2	C/T	0.031	0.059	BW,BRR-RC	7.35×10^{-11}

Table 1. Novel and replicated discoveries using mixed-linear association models that either adjust for age-at-onset (GMRM-BayesW), or not (GMRM-BayesRR-RC). Mixed-linear association model results were LD clumped such that the index SNPs would have a p-value below $5 \cdot 10^{-8}$ and SNPs could be added to a clump if they were 1Mb from the index SNP, they were correlated with $r^2 > 0.05$ and they were nominally significant ($p < 0.05$). We then used the COJO method from the GCTA software (see Methods) to find clumps with independent signals by conducting stepwise selection of index SNPs in 1Mb window and we considered SNPs independent if they had a p-value below $5 \cdot 10^{-8}$ in the joint model. To determine novelty, we first removed all markers that were significantly associated in the unadjusted model (eq. 3). We then removed all the markers that had a correlation of $r^2 > 0.1$ with a marker that had been previously found associated with a cancer of interest using GWAS Catalog (published until July 2021) and LDtrait tool with the British in England and Scotland population. We then again used COJO to condition the remaining markers on the previously identified associations for each cancer of interest and SNPs that did not fall below $5 \cdot 10^{-8}$ in the joint model were eliminated. For the remaining SNPs, we conducted an additional literature review using Phenoscanner database (see Methods) to find any previous associations with variants of interest or variants in LD. The remaining candidates of novel associations were concatenated across GMRM-BayesW or GMRM-BayesRR-RC adjusted analyses and then replicated in the largest available studies conducted for each specific cancer type and the Estonian Biobank. Replication was defined as Benjamini-Hochberg corrected p-value being lower than 0.05 and the direction of the effect size same in both the original analysis and the replication analysis. The column Nearest gene is mapped from the SNP using ANNOVAR software, * in that column denotes intergenic regions.

Cancer	Probe	Gene mediator	Chr	Total effect (SE)	Gene → Trait effect (SE)	DNAm → Gene effect (SE)	Mediation proportion
Basal cell carcinoma	cg21169611	SMC2	9	0.03 (0.005)	0.31 (0.059)	0.05 (0.010)	0.41
Bladder cancer	cg13446199	LYNX1	8	0.01 (0.002)	-0.13 (0.021)	-0.03 (0.010)	0.53
Breast cancer	cg03895047	NEK10	3	0.03 (0.005)	-0.04 (0.012)	-0.34 (0.051)	0.31
Cervical cancer	cg15582954	CDC42	1	0.05 (0.008)	0.70 (0.172)	0.04 (0.007)	0.80
Cervical cancer	cg19083407	PAX8	2	-0.02 (0.003)	-0.54 (0.001)	0.04 (0.006)	0.84
Prostate cancer	cg20240347	PIK3C2B	1	-0.02 (0.003)	0.10 (0.019)	-0.03 (0.018)	0.28
Prostate cancer	cg06639874	COL6A3	2	-0.04 (0.006)	0.11 (0.016)	-0.04 (0.035)	0.73
Prostate cancer	cg03779937	PDK1	2	0.04 (0.007)	0.28 (0.055)	0.06 (0.014)	0.35
Prostate cancer	cg09757087	TMEM204	16	0.02 (0.004)	-0.98 (0.043)	-0.02 (0.003)	0.56
Prostate cancer	cg01799818	FAM57A	17	0.03 (0.005)	-0.09 (0.017)	-0.20 (0.046)	0.25

Table 2. Mediation of DNAm effect on cancer through gene expression. Estimates from multivariable Mendelian randomisation (MR, see Methods), which was used to quantify mediation through gene expression for significant DNAm-to-cancer MR total effects. This provides a list of genes through which germline variation alters altering methylation, which changes gene expression, and in turn influences cancer risk. The mediation proportion quantifies the proportion of mediated causal effect (DNAm → Gene → cancer) relative to the total effect (DNAm → cancer).

	BayesW	BayesRR-RC	Array-based estimate[a]	Family-based estimate[b]
Basal cell carcinoma	0.25 (0.19-0.31)	0.25 (0.19-0.31)	0.17 (0.07-0.27)[c]	0.43 (0.26-0.59)
Bladder cancer	0.08 (0.01-0.16)	0.09 (0.01-0.17)	0.08 (0.04–0.12)	0.30 (0.00–0.67)
Breast cancer	0.20 (0.15-0.24)	0.16 (0.13-0.20)	0.10 (0.08–0.13)	0.31 (0.11–0.51)
Cervical cancer	0.05 (0.03-0.07)	0.05 (0.03-0.06)	0.07 (0.02–0.12)	0.13 (0.06–0.15) [d]
Colon cancer	0.08 (0.04-0.12)	0.08 (0.04-0.12)	0.07 (0.04–0.10)	0.15 (0.00–0.45)
Endometrial cancer	0.14 (0.06-0.22)	0.08 (0.00-0.16)	0.13 (0.07–0.18)	0.27 (0.11–0.43)
Melanoma	0.11 (0.06-0.16)	0.11 (0.06-0.16)	0.08 (0.04–0.11)	0.58 (0.43–0.73)
Non-Hodgkin's lymphoma	0.03 (0.00-0.09)	0.03 (0.00-0.10)	0.13 (0.03–0.23)	0.10 (0.08–0.10)[d]
Ovarian cancer	0.05 (0.00-0.13)	0.02 (0.00-0.10)	0.07 (0.01–0.13)	0.39 (0.23–0.55)
Prostate cancer	0.30 (0.23-0.37)	0.25 (0.19-0.31)	0.16 (0.13–0.20)	0.57 (0.51–0.63)
Testicular cancer	0.38 (0.19-0.58)	0.29 (0.11-0.47)	0.26 (0.15–0.38)	0.37 (0.00–0.93)

Table 3. SNP-heritability estimates. Estimates with 95% CI from LD Score regression, using mixed linear association model estimates adjusting for age-at-onset (GMRM-BayesW) or not (GMRM-BayesRR-RC), as compared with previous array or family based estimates. a - estimate from Rashkin et al. [7]; b - estimate from Mucci et al. [71]; c - estimate from Kilgour et al. [72]; d - estimates from Czene et al. [73].

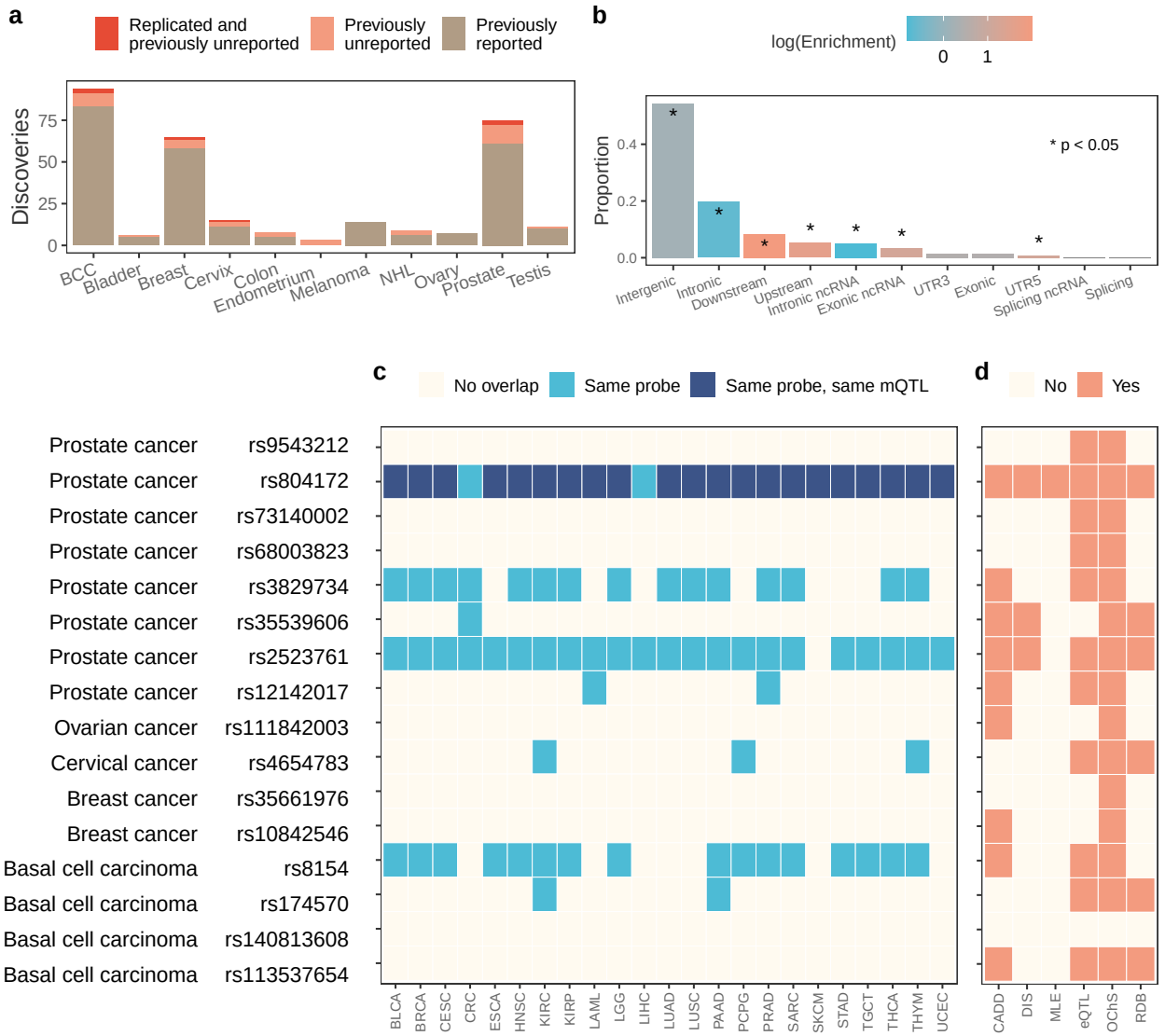

Figure 1. SNP discovery and properties of replicated novel discoveries. (a) Count of previously reported discoveries, previously unreported discoveries and previously unreported replicated discoveries using a mixed-linear age-at-onset adjusted association model (GMRM-BayesW) and mixed-linear association model (GMRM-BayesRR-RC). (b) Proportion of SNPs from the novel genetic regions per functional consequences on genes annotated using ANNOVAR; enrichment of functional consequences of SNPs are tested using a two-sided Fisher's exact test. (c) Overlap of methylation probes affected by SNPs from the novel genetic regions with probes that are differentially methylated across 23 cancers from the Pancan-meQTL database; the probes that are differentially methylated by means of the same SNP are marked dark blue (BLCA - Bladder Urothelial Carcinoma, BRCA - Breast invasive carcinoma, CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma, CRC - Colon adenocarcinoma + Rectum adenocarcinoma, ESCA - Esophageal carcinoma, HNSC - Head and Neck squamous cell carcinoma, KIRC - Kidney renal clear cell carcinoma, KIRP - Kidney renal papillary cell carcinoma, LAML - Acute Myeloid Leukemia, LGG - Lower Grade Glioma, LIHC - Liver hepatocellular carcinoma, LUAD - Lung adenocarcinoma, LUSC - Lung squamous cell carcinoma, PAAD - Pancreatic adenocarcinoma, PCPG - Pheochromocytoma and Paraganglioma, PRAD - Prostate adenocarcinoma, SARC - Sarcoma, SKCM - Skin Cutaneous Melanoma, STAD - Stomach adenocarcinoma, TGCT - Testicular Germ Cell Tumors, THCA - Thyroid carcinoma, THYM - Thymoma, UCEC - Uterine Corpus Endometrial Carcinoma). (d) Properties of novel replicated genetic regions. CADD - maximum CADD score of the region is above 12.37, DIS - maximum DeepSEA disease impact score (DIS) of the genetic region is above 2, MLE - maximum DeepSEA mean log e-value (MLE) of the region is above 2, eQTL - an SNP from the the genetic region is an eQTL with p-value $< 5 \cdot 10^{-8}$, OChS - open/active chromatin state (minimum 15-core chromatin score of the lead SNP is less or equal than 7), RDB - minimum RegulomeDB category of the genetic region is 1 or 2.

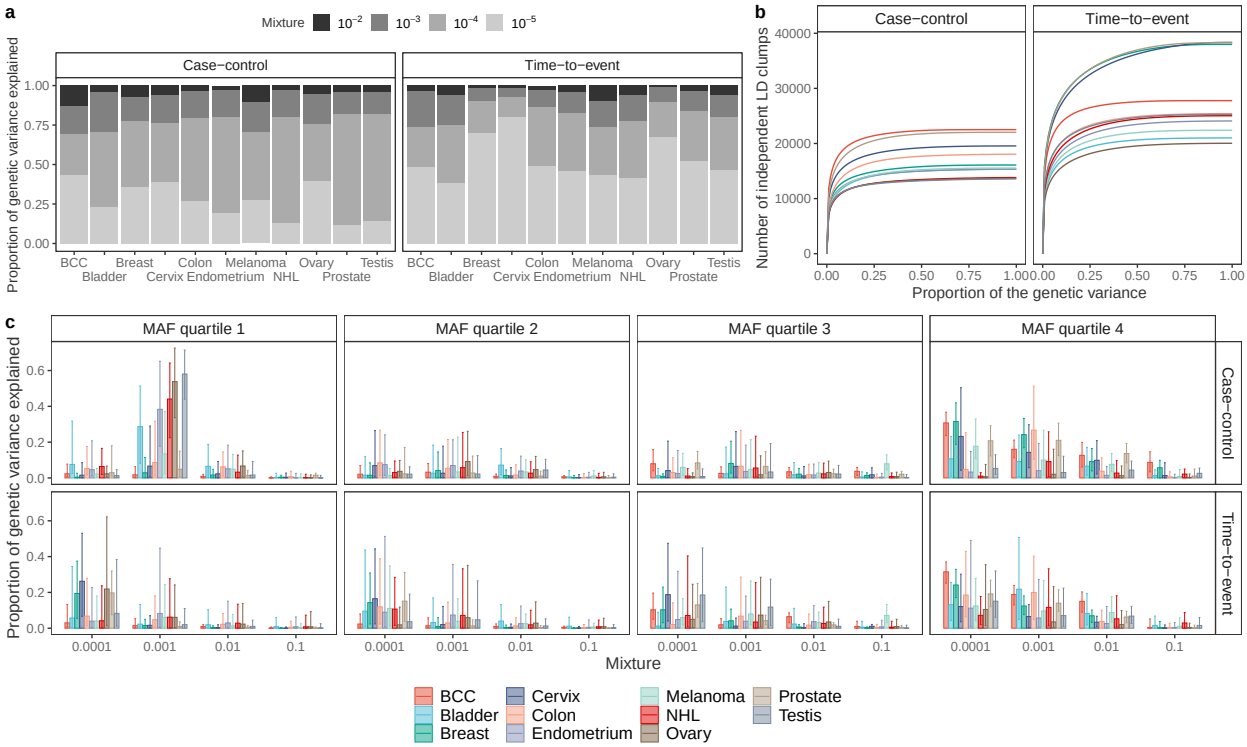

Figure 2. Genetic architecture and polygenicity of 11 cancers. (a) Mean proportion of genetic variance explained by each of the mixtures components using either case-control or age-at-onset phenotype. We find evidence that age-at-onset is highly polygenic with most of the genetic variance attributable to SNPs contributed by markers in the 10^{-4} mixture group, while the majority of the case-control phenotype genetic variance is explained by the markers from the 10^{-3} mixture group. (b) Number of LD-independent regions (see Methods) needed to explain total genetic variance. The contributions of LD-independent regions were sorted ascendingly such that the smallest contributing regions were added first. (c) Median proportion of genetic variance explained by each mixture class and MAF quartile combination, with 95% CI. For both case-control and age-at-onset models, most of the genetic variance is attributable to the small effect common variants (MAF quartile 4), however rare variants from the first MAF quartile contribute significantly to the variance for bladder, endometrial, ovarian, testicular cancers, non-Hodgkin's lymphoma for the BayesR model. BCC indicates basal cell carcinoma.

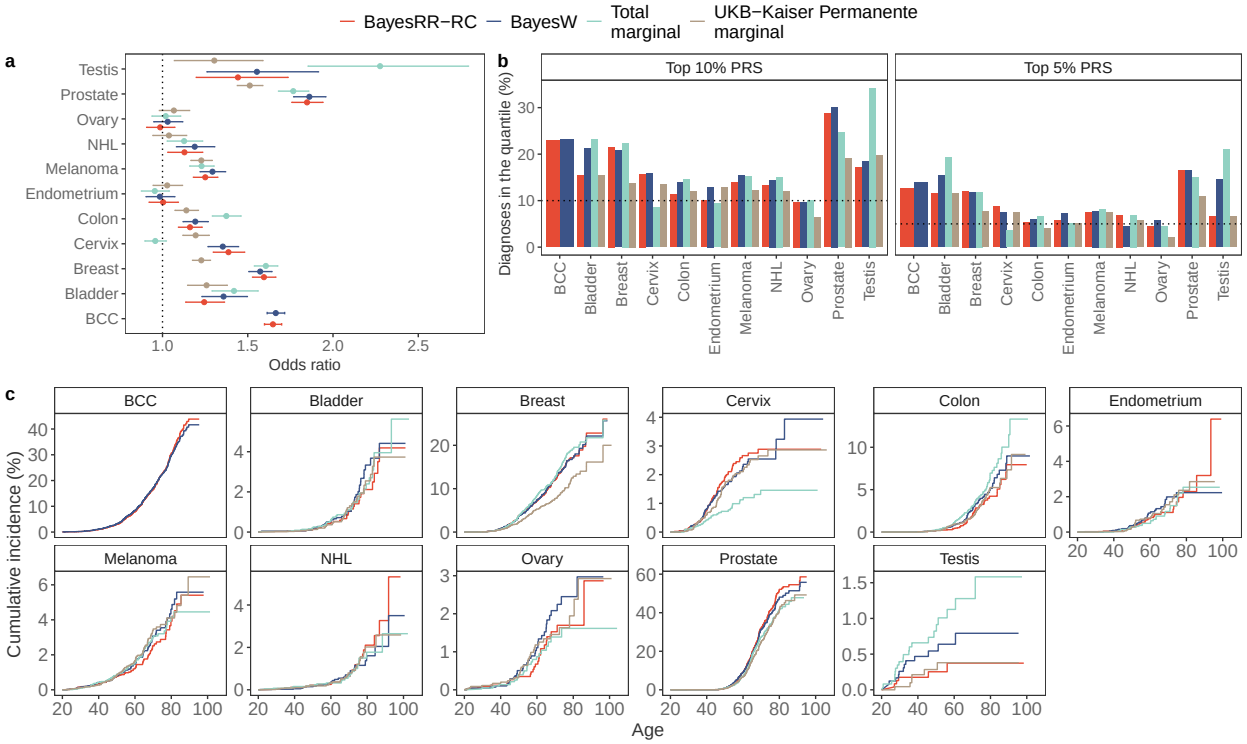

Figure 3. Predictive validation of different polygenic risk scores (PRS) in the Estonian Biobank data. (a) Odds ratio for diagnosis of a tumour given one standard deviation increase in PRS, with 95% confidence intervals. (b) Percent of individuals diagnosed with cancer before age 50 having a top 10% or top 5% highest PRS; (c) cumulative incidence curves adjusted for competing risk for individuals with the top 5% highest PRS. The number of Estonian Biobank individuals used in the validation was N =195,432. BayesRR-RC and BayesW estimates were obtained by running the corresponding models on UK Biobank using either case-control or age-at-onset data. Total marginal estimates were obtained by using the marginal estimates that were concatenated from different GWA studies by Kachuri et al [41]. UKB-Kaiser Permanente estimates were obtained from the meta-analysis that combined analyes of UK Biobank and Kaiser Permanente cohorts [7]. We see that although BayesW and BayesRR-RC have the smallest sample sizes along with the smallest numbers of cases, the predictors uniformly perform better than a marginal analysis conducted on a slightly larger data set (UKB-Kaiser Permanente), and with a few exceptions it achieves similar or better predictive accuracy compared to the total marginal estimates that use effectively up to 10 times more tumour cases than BayesW and BayesRR-RC analyses. For all cancers except breast, cervical, endometrial and ovarian cancer, BayesW predictor gives a nominally higher odds ratios compared to BayesRR-RC predictor.

Supplementary material

Supplementary tables

	All			Females			Males		
	Cases (%)	Mean (sd)	Median (Range)	Cases (%)	Mean (sd)	Median (Range)	Cases (%)	Mean (sd)	Median (Range)
Basal cell carcinoma	26758 (5.8%)	60.6 (9.42)	62.1 (7.5-80.6)	13277 (5.3%)	59.5 (9.73)	61 (7.5-80.6)	13481 (6.4%)	61.8 (8.96)	63.3 (22.2-79.7)
Bladder cancer	2470 (0.5%)	61.3 (9.38)	62.9 (19.1-77.8)	608 (0.2%)	60.8 (9.50)	62.1 (24.7-76.9)	1862 (0.9%)	61.5 (9.34)	63.1 (19.1-77.8)
Breast cancer	16972 (6.8%)	55.7 (9.20)	55.6 (18.8-80.9)	16972 (6.8%)	55.7 (9.20)	55.6 (18.8-80.9)			
Cervical cancer	8680 (3.5%)	38.0 (9.27)	36.8 (13.5-76.6)	8680 (3.5%)	38.0 (9.27)	36.8 (13.5-76.6)			
Colon cancer	4463 (1.0%)	60.9 (9.13)	62.3 (11.2-78.8)	2009 (0.8%)	60.2 (9.53)	61.4 (11.2-78.0)	2454 (1.2%)	61.5 (8.76)	62.7 (14.0-78.8)
Endometrial cancer	2227 (0.9%)	56.5 (11.22)	58.3 (13.0-76.9)	2227 (0.9%)	56.5 (11.22)	58.3 (13.0-76.9)			
Melanoma	5778 (1.3%)	54.1 (12.16)	55.8 (0.5-77.5)	3243 (1.3%)	52.1 (12.36)	53.3 (1.5-77.5)	2535 (1.2%)	56.6 (11.41)	58.5 (0.5-77.5)
Non-Hodgkin's lymphoma	2298 (0.5%)	58.7 (11.36)	61.0 (3.5-79.1)	1026 (0.4%)	58.8 (10.97)	60.9 (5.3-78.1)	1272 (0.6%)	58.7 (11.67)	61.0 (3.5-79.1)
Ovarian cancer	1573 (0.6%)	54.6 (12.38)	56.0 (10.1-79.6)	1573 (0.6%)	54.6 (12.38)	56.0 (10.1-79.6)			
Prostate cancer	9824 (4.7%)	64.6 (5.87)	65.2 (22.9-80.2)				9824 (4.7%)	64.6 (5.87)	65.2 (22.9-80.2)
Testicular cancer	886 (0.4%)	40.5 (11.08)	40.0 (0.5-76.0)				886 (0.4%)	40.5 (11.08)	40.0 (0.5-76.0)

Table S1. UK Biobank data composition for the cancer cases and their timings used within the study.

	ICD10 code	ICD9 code
Basal cell carcinoma	C44.0, C44.1, C44.2, C44.3, C44.4, C44.5, C44.6, C44.7, C44.8, C44.9, D04.0, D04.1, D04.2, D04.3, D04.4, D04.5, D04.6, D04.7, D04.8, D04.9	173.0, 173.1, 173.2, 173.3, 173.4, 173.5, 173.6, 173.7, 173.8, 173.9, 232.1, 232.2, 232.3, 232.4, 232.5, 232.6, 232.7, 232.8, 232.9
Bladder cancer	C67.0, C67.1, C67.2, C67.3, C67.4, C67.5, C67.6, C67.6, C67.7, C67.8, C67.9, D09.0	188.0, 188.2, 188.4, 188.6, 188.8, 188.9, 233.7
Breast cancer	C50.0, C50.1, C50.2, C50.3, C50.4, C50.5, C50.6, C50.7, C50.8, C50.9, D05.0, D05.1, D05.7, D05.9	174.0, 174.1, 174.2, 174.3, 174.4, 174.5, 174.6, 174.7, 174.8, 174.9, 233.0
Cervical cancer	C53.0, C53.1, C53.8, C53.9, D06.0, D06.1, D06.7, D06.9	180.0, 180.1, 180.8, 180.9, 233.1
Colon cancer	C18.0, C18.1, C18.2, C18.3, C18.4, C18.5, C18.6, C18.7, C18.8, C18.9, D01.0	153.0, 153.1, 153.2, 153.3, 153.4, 153.5, 153.6, 153.7, 153.8, 153.9, 230.3
Endometrial cancer	C54.1, D07.0	182.0
Melanoma	C43.0, C43.1, C43.2, C43.3, C43.4, C43.5, C43.6, C43.7, C43.8, C43.9	172.0, 172.1, 172.2, 172.3, 172.4, 172.5, 172.6, 172.7, 172.8, 172.9
Non-Hodgkin's lymphoma	C82.0, C82.1, C82.2, C82.7, C82.9, C83.0, C83.1, C83.2, C83.3, C83.4, C83.5, C83.6, C83.7, C83.8, C83.9, C85.0, C85.1, C85.7, C85.9	202.8
Ovarian cancer	C56	183.0
Prostate cancer	C61, D07.5	185
Testicular cancer	C62.0, C62.1, C62.9	186.9

Table S2. Cancer-specific ICD10 and ICD9 codes used to select cases from the UK and Estonian biobank studies. For each of the tumour types, the corresponding ICD10 and ICD9 codes are presented that were used to define cancer occurrence.

Site	Chr	SNP	Eff/ Oth	p (BayesW)	p (BayesRR-RC)	p (Unadj.)	$\hat{\beta}$ (BayesW)	$\hat{\beta}$ (BayesRR-RC)	$\hat{\beta}$ (Unadj.)
Basal cell carcinoma	1	rs501823	T/A	2.76×10^{-8}	2.25×10^{-8}	3.80×10^{-7}	-0.016	-0.016	-0.015
Basal cell carcinoma	1	rs5011752	G/A	2.89×10^{-8}	4.23×10^{-8}	1.15×10^{-6}	0.013	0.013	0.011
Basal cell carcinoma	2	rs11686655	G/A	2.26×10^{-8}	2.47×10^{-8}	2.83×10^{-7}	-0.039	-0.039	-0.036
Basal cell carcinoma	3	rs8154	C/T	3.31×10^{-8}	1.81×10^{-8}	8.81×10^{-8}	-0.013	-0.013	-0.012
Basal cell carcinoma	3	rs140813608	A/G	4.00×10^{-8}	3.88×10^{-8}	6.09×10^{-7}	0.045	0.045	0.041
Basal cell carcinoma	3	rs17514215	G/T	2.57×10^{-8}	3.23×10^{-8}	1.77×10^{-7}	0.016	0.016	0.015
Basal cell carcinoma	5	rs36137978	C/A	3.44×10^{-9}	5.29×10^{-9}	3.78×10^{-7}	0.015	0.015	0.013
Basal cell carcinoma	10	rs3122367	C/G	4.42×10^{-8}	6.43×10^{-8}	3.65×10^{-7}	-0.013	-0.013	-0.012
Basal cell carcinoma	11	rs174570	T/C	1.16×10^{-10}	2.17×10^{-10}	5.39×10^{-8}	-0.021	-0.02	-0.018
Basal cell carcinoma	12	rs3819817	C/T	1.44×10^{-8}	7.41×10^{-9}	1.18×10^{-7}	0.012	0.013	0.012
Basal cell carcinoma	14	rs12892576	G/A	4.39×10^{-9}	6.44×10^{-9}	1.62×10^{-7}	0.019	0.019	0.017
Basal cell carcinoma	15	rs8023809	A/G	4.96×10^{-9}	4.70×10^{-9}	4.97×10^{-7}	0.015	0.015	0.013
Basal cell carcinoma	16	rs55752638	A/T	6.98×10^{-10}	6.90×10^{-10}	1.65×10^{-7}	0.017	0.017	0.015
Basal cell carcinoma	16	rs113537654	C/T	4.09×10^{-8}	7.03×10^{-8}	2.44×10^{-6}	-0.016	-0.015	-0.013
Basal cell carcinoma	16	rs7206699	C/T	6.74×10^{-10}	1.06×10^{-9}	7.46×10^{-8}	0.013	0.013	0.012
Bladder cancer	11	rs147529765	A/G	1.07×10^{-7}	2.71×10^{-8}	1.53×10^{-7}	0.063	0.065	0.062
Bladder cancer	15	rs16969577	T/C	4.10×10^{-8}	1.24×10^{-8}	8.64×10^{-8}	0.072	0.074	0.07
Breast cancer	5	rs35661976	T/C	1.71×10^{-9}	7.92×10^{-8}	1.94×10^{-6}	0.062	0.055	0.049
Breast cancer	11	rs143173464	C/T	4.67×10^{-8}	1.56×10^{-7}	3.89×10^{-7}	0.067	0.064	0.062
Breast cancer	12	rs10842546	A/G	1.28×10^{-8}	9.92×10^{-8}	3.85×10^{-6}	-0.017	-0.016	-0.013
Breast cancer	16	rs7500465	A/G	1.40×10^{-9}	2.46×10^{-8}	5.26×10^{-7}	0.018	0.016	0.015
Cervical cancer	1	rs4654783	T/C	2.56×10^{-12}	3.21×10^{-8}	3.79×10^{-6}	0.022	0.017	0.015
Cervical cancer	11	rs182301918	G/A	7.46×10^{-9}	1.53×10^{-7}	3.14×10^{-6}	0.124	0.112	0.1
Colon cancer	2	rs181425761	A/G	4.81×10^{-9}	1.68×10^{-8}	1.08×10^{-7}	0.107	0.103	0.097
Colon cancer	10	rs149750027	A/G	6.49×10^{-8}	4.49×10^{-8}	2.97×10^{-7}	0.098	0.099	0.093
Endometrial cancer	1	rs114357987	T/C	7.72×10^{-10}	1.22×10^{-7}	4.67×10^{-6}	-0.053	-0.046	-0.04
Endometrial cancer	8	rs185211261	C/T	3.29×10^{-5}	1.16×10^{-8}	1.62×10^{-6}	0.117	0.161	0.136
Endometrial cancer	10	rs183400920	T/C	2.56×10^{-8}	5.21×10^{-7}	2.93×10^{-6}	0.151	0.136	0.127
Endometrial cancer	16	rs146032590	G/C	2.54×10^{-9}	7.43×10^{-9}	1.50×10^{-5}	0.083	0.08	0.06
Non-Hodgkin's lymphoma	2	rs144739970	C/T	3.74×10^{-8}	1.15×10^{-7}	1.80×10^{-7}	0.071	0.068	0.067
Non-Hodgkin's lymphoma	3	rs73113445	A/C	4.59×10^{-8}	2.72×10^{-8}	1.91×10^{-6}	0.029	0.029	0.025
Non-Hodgkin's lymphoma	8	rs181582553	A/G	2.71×10^{-7}	2.16×10^{-8}	1.01×10^{-6}	0.084	0.092	0.08
Ovarian cancer	1	rs147207960	C/T	1.02×10^{-8}	5.80×10^{-8}	2.49×10^{-6}	0.12	0.114	0.099
Ovarian cancer	3	rs111842003	A/T	1.62×10^{-5}	6.12×10^{-9}	5.45×10^{-7}	0.089	0.12	0.104
Ovarian cancer	17	rs7210734	C/T	2.72×10^{-8}	4.10×10^{-10}	5.18×10^{-8}	0.074	0.083	0.072
Ovarian cancer	20	rs77014550	T/C	1.04×10^{-7}	6.33×10^{-9}	4.49×10^{-7}	0.11	0.121	0.105
Prostate cancer	1	rs12142017	C/T	1.83×10^{-8}	2.69×10^{-7}	3.40×10^{-5}	-0.033	-0.03	-0.025
Prostate cancer	4	rs68003823	G/A	1.53×10^{-8}	3.08×10^{-7}	2.48×10^{-6}	-0.026	-0.023	-0.022
Prostate cancer	5	rs62359313	A/G	3.32×10^{-9}	6.21×10^{-8}	6.79×10^{-6}	-0.02	-0.019	-0.016
Prostate cancer	6	rs2523761	G/A	9.14×10^{-10}	3.79×10^{-8}	2.50×10^{-6}	-0.024	-0.021	-0.018
Prostate cancer	6	rs6910025	T/C	9.35×10^{-9}	6.99×10^{-7}	4.78×10^{-6}	-0.028	-0.024	-0.022
Prostate cancer	6	rs9397090	A/G	1.50×10^{-8}	1.65×10^{-7}	5.36×10^{-6}	-0.026	-0.024	-0.021
Prostate cancer	8	rs77965869	T/C	2.07×10^{-9}	4.52×10^{-7}	1.12×10^{-5}	-0.125	-0.105	-0.092
Prostate cancer	8	rs78653149	A/G	6.63×10^{-9}	1.05×10^{-7}	2.13×10^{-6}	0.03	0.028	0.025
Prostate cancer	10	rs12774441	T/G	2.02×10^{-9}	2.83×10^{-8}	3.58×10^{-5}	-0.025	-0.023	-0.017
Prostate cancer	12	rs3829734	A/G	1.07×10^{-10}	4.37×10^{-8}	1.23×10^{-5}	0.022	0.019	0.015
Prostate cancer	13	rs9543212	T/C	3.25×10^{-8}	8.23×10^{-8}	4.75×10^{-7}	0.019	0.018	0.017
Prostate cancer	16	rs804172	C/G	5.90×10^{-9}	7.15×10^{-8}	7.07×10^{-6}	-0.018	-0.017	-0.014
Prostate cancer	16	rs9935422	T/C	2.82×10^{-10}	2.47×10^{-8}	1.35×10^{-6}	-0.034	-0.03	-0.026
Prostate cancer	18	rs35539606	T/C	2.89×10^{-9}	2.68×10^{-8}	5.02×10^{-7}	0.026	0.024	0.022
Prostate cancer	20	rs73140002	C/T	7.35×10^{-11}	2.91×10^{-9}	5.18×10^{-8}	0.059	0.054	0.049
Prostate cancer	21	rs74503316	T/C	2.01×10^{-8}	1.62×10^{-7}	1.64×10^{-5}	0.02	0.019	0.016
Testicular cancer	2	rs115509835	A/G	1.31×10^{-6}	1.56×10^{-8}	3.83×10^{-7}	0.15	0.176	0.158
Testicular cancer	3	rs137897706	A/G	2.01×10^{-5}	4.16×10^{-8}	1.58×10^{-7}	0.106	0.136	0.13
Testicular cancer	5	rs192275753	C/T	1.19×10^{-6}	2.56×10^{-8}	6.01×10^{-7}	0.126	0.145	0.13
Testicular cancer	9	rs10960870	A/G	1.01×10^{-8}	1.99×10^{-8}	4.37×10^{-7}	0.064	0.063	0.057
Testicular cancer	11	rs368211890	A/C	3.08×10^{-6}	3.47×10^{-8}	2.74×10^{-5}	0.146	0.173	0.132
Testicular cancer	20	rs118074710	G/A	8.45×10^{-7}	2.52×10^{-8}	1.27×10^{-5}	0.13	0.147	0.115
Testicular cancer	22	rs183880425	T/G	5.71×10^{-8}	2.76×10^{-8}	1.06×10^{-7}	0.122	0.125	0.119

Table S3. Previously unreported discoveries from GMRM-BayesRR-RC or GMRM-BayesW analyses in comparison with results from an unadjusted marginal association analysis. We observe that for the 59 previously unreported variants, the p-value in the unadjusted association analysis is borderline significant ($5 \cdot 10^{-8} < p < 10^{-4}$) and by using the BayesW or BayesRR-RC adjustments, we arrive at statistically significant test statistics.

SNP	Function	Site	eQTL (eQTLGen) MR-IVW	eQTL (GTEx v8) SMR	eQTL (eQTLGen) SMR	mQTL (GoDMC) SMR	mQTL (GoDMC) MR-IVW	Nearest gene (bp)
rs10842546	intergenic	Breast cancer						KRAS(121643), RNU4-67P(31709)
rs113537654	intronic	Basal cell carcinoma						TANGO6
rs12142017	intronic	Prostate cancer				DCST1, DCST2		ZBTB7B
rs140813608	intergenic	Basal cell carcinoma						RP11-416O18.2(121645), RP11-338L18.1(101365)
rs174570	intronic	Basal cell carcinoma				RPLP0P2	FADS1, FADS2, FADS3, RAB3IL1	FADS2
rs2523761	intergenic	Prostate cancer		HCP5B, HLA-A, HLA-V, TRIM31 ZFP57	DDX39BP2, HCG4P3, TRIM10, ZNRD1	ZKSCAN3, ZNF192, ZNF193, ZNF323, ZNF389, ZSCAN12 ZSCAN16, ZSCAN23	ZSCAN12	HLA-G(19824), MICF(1414)
rs35539606	intergenic	Prostate cancer					SALL3	CTD-2382H12.1(22146), RP11-849I19.1(26968)
rs35661976	exonic	Breast cancer						SLC6A18
rs3829734	intronic	Prostate cancer		PFKM		COL2A1, H1FNT, PFKM, RND1	COL2A1, PFKM	COL2A1
rs4654783	intergenic	Cervical cancer	CDC42	CDC42	CDC42	WNT4		RP1-224A6.9(11667), WNT4(4278)
rs68003823	intergenic	Prostate cancer		AFP				AFP(4065), AFM(21444)
rs73140002	ncRNA intronic	Prostate cancer		RP4-724E16.2, ZNF217	RP4-724E16.2, ZNF217			RP4-724E16.2
rs804172	intergenic	Prostate cancer				TFAP4	TFAP4	TFAP4(26675), GLIS2(15011)
rs8154	exonic	Basal cell carcinoma		ATG7		ATG7, VGLL4	ATG7, HRH1 VGLL4	ATG7
rs9543212	intergenic	Prostate cancer						PSMD10P3(13114), KLF5(5014)
rs111842003	intronic	Ovarian cancer						CD96

Table S4. Mapping the 16 novel associations to genes. We combined results from the following analyses: single-instrument SMR on GTEx (v8) (1680 tests), eQTLGen (145 tests), and GoDMC (2380 tests) datasets and multi-instrument MR-IVW on eQTLGen (61 tests) and GoDMC (998 tests) datasets. The overall significance threshold on combined data was Bonferroni-corrected $p < 0.05/5264 = 9.5 \cdot 10^{-6}$. The genes affected by differentiated methylation were taken from the Pancan-meQTL database. Finally, we extended the resulting list of genes by appending the nearest to the novel SNPs genes and protein coding genes from FUMA positional, eQTL, and chromatin interaction mapping.

		BayesW (95% CI)	BayesRR-RC (95% CI)
Basal cell carcinoma	Cervical cancer	0.20 (0.06, 0.35)	0.18 (0.03, 0.33)
Basal cell carcinoma	Melanoma	0.51 (0.34, 0.68)	0.51 (0.34, 0.67)
Breast cancer	Endometrial cancer	0.23 (0.03, 0.44)	0.33 (0.03, 0.64)
Colon cancer	Cervical cancer	0.33 (0.06, 0.60)	0.39 (0.11, 0.68)
Colon cancer	Endometrial cancer	0.46 (0.09, 0.83)	0.45 (-0.05, 0.94)
Testicular cancer	Endometrial cancer	-0.38 (-0.68, -0.07)	-0.57 (-1.00, -0.02)

Table S5. Statistically significant cross-trait genetic correlations from LD score regression analysis. We calculate the genetic correlations between cancers with cross-trait LD score regression [57] applying it on the results from BayesW or BayesRR-RC adjusted analyses. LD scores for the analysis were taken from the 1000 Genomes European data. Both BayesRR-RC and BayesW based significant genetic correlations agree on the magnitude of the estimates but BayesW based estimates result in narrower confidence interval than the BayesRR-RC based estimates.

	Alternative LDSC bW-adjusted	Full Bayesian BayesRR-RC
Basal cell carcinoma	0.15 (0.10-0.20)	0.22 (0.20-0.24)
Bladder cancer	0.06 (0.00-0.11)	0.29 (0.19-0.42)
Breast cancer	0.16 (0.11-0.21)	0.20 (0.17-0.23)
Cervical cancer	0.08 (0.05-0.12)	0.13 (0.08-0.20)
Colon cancer	0.07 (0.03-0.10)	0.19 (0.12-0.28)
Endometrial cancer	0.10 (0.03-0.16)	0.34 (0.23-0.47)
Melanoma	0.09 (0.04-0.14)	0.18 (0.13-0.24)
Non-Hodgkin's lymphoma	0.02 (0.00-0.07)	0.32 (0.23-0.43)
Ovarian cancer	0.05 (0.00-0.13)	0.49 (0.37-0.61)
Prostate cancer	0.22 (0.16-0.29)	0.33 (0.28-0.40)
Testicular cancer	0.36 (0.15-0.56)	0.81 (0.73-0.87)

Table S6. Alternative liability scale heritability estimates with 95% CI. We use the observed scale from LDSC estimates (summary statistics from BayesW-adjusted analysis) and the heritability estimates from the full Bayesian model (BayesRR-RC). Transformation of the observed scale heritabilities are done with a more conservative approach (Ojavee et al. [40]) better suited for rare diseases.

	Risk (age 0-85)
Basal cell carcinoma	0.3050
Bladder cancer	0.0224
Breast cancer	0.1248
Cervical cancer	0.0054
Colon cancer	0.0232
Endometrial cancer	0.0300
Melanoma	0.0269
Non-Hodgkin's lymphoma	0.0161
Ovarian cancer	0.0093
Prostate cancer	0.1117
Testicular cancer	0.0051

Table S7. Cancer risk from birth to age 85, SEER estimate 2016-2018 To ensure that the lifetime risk estimates were similar to the study population (European ancestry, UK Biobank, oldest individual age 86) we used the estimates from SEER of non-hispanic white of getting diagnosed between ages (0-85). The explorer is accessible from `https://seer.cancer.gov/explorer/`. The explorer had a joint estimate for colorectal cancer that we transformed to the risk of colon cancer using the proportion of colon cancer cases among colorectal cases (70.3% , `https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html`, accessed 24.01.2022). For basal cell carcinoma we used lifetime risk estimate from Miller et al. [74].

Supplementary figures

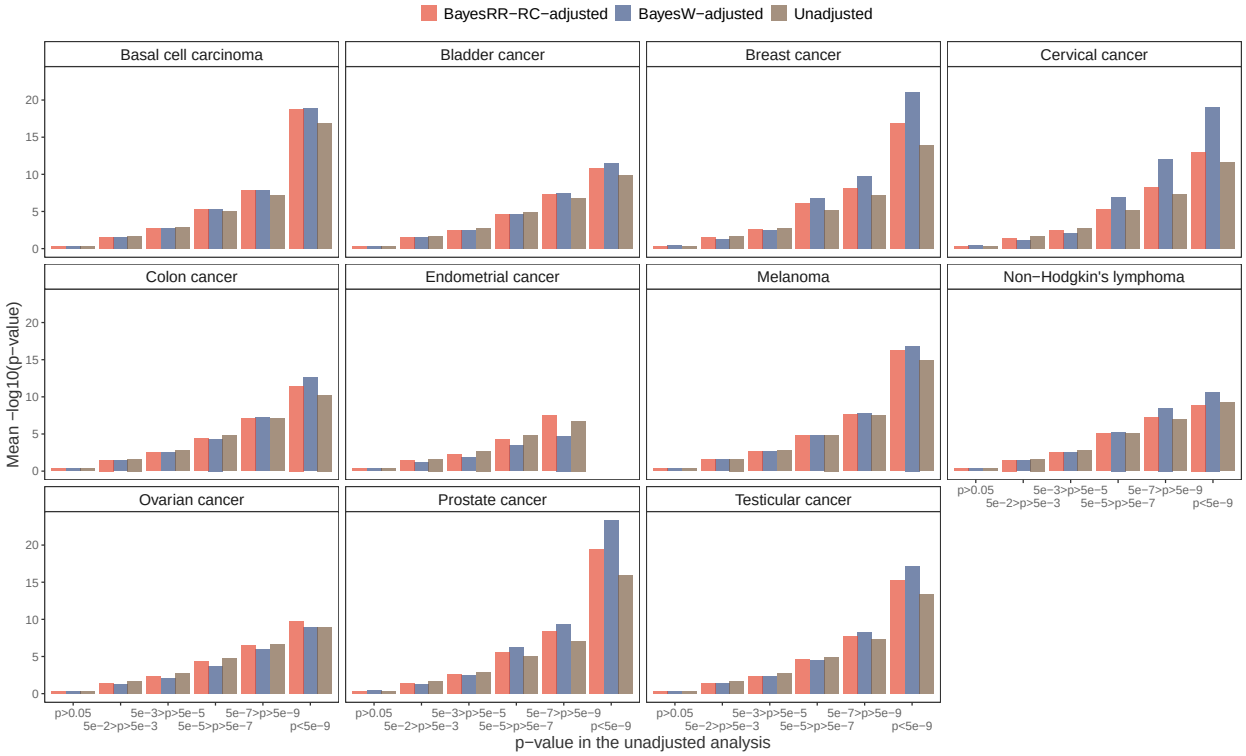

Figure S1. Mean -log10 p-value from the marginal association analysis adjusted with either BayesRR-RC, BayesW or without adjustment. We observe that in general BayesW or BayesRR-RC LOCO adjustments result in decreased p-values suggesting an increase in statistical power. Furthermore, for most traits, BayesW adjustment results in even lower p-values compared to the ones resulting from BayesRR-RC adjustment.

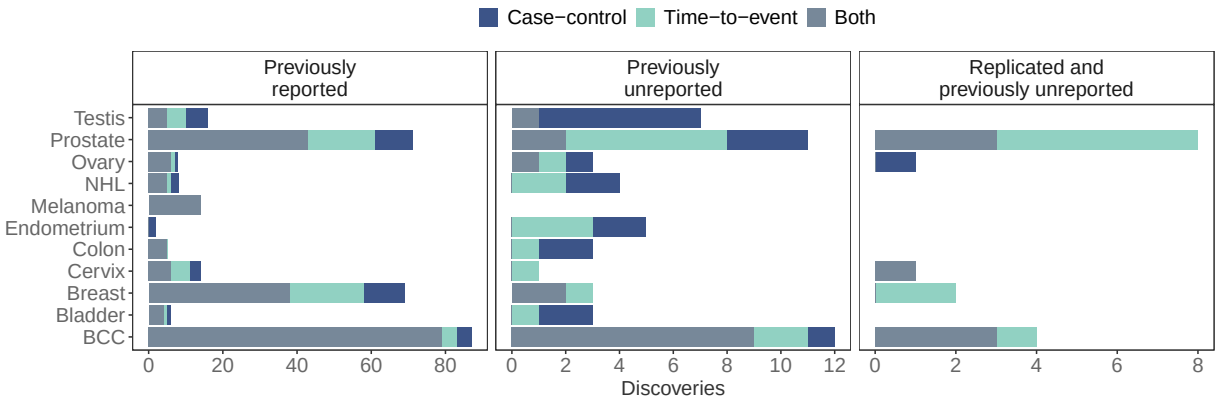

Figure S2. Classification of previously reported, unreplicated previously unreported and replicated previously unreported discoveries by each cancer type. Among previously reported variants both case-control and age-at-onset adjustments most commonly discover same variants in our analysis. Time-to-event adjustment suggests to give more associations than case-control adjustment and that mostly for cancers with high case count (prostate cancer, breast cancer, basal cell carcinoma). The replication rate is slightly better for variants that had been discovered using time-to-event adjustment.

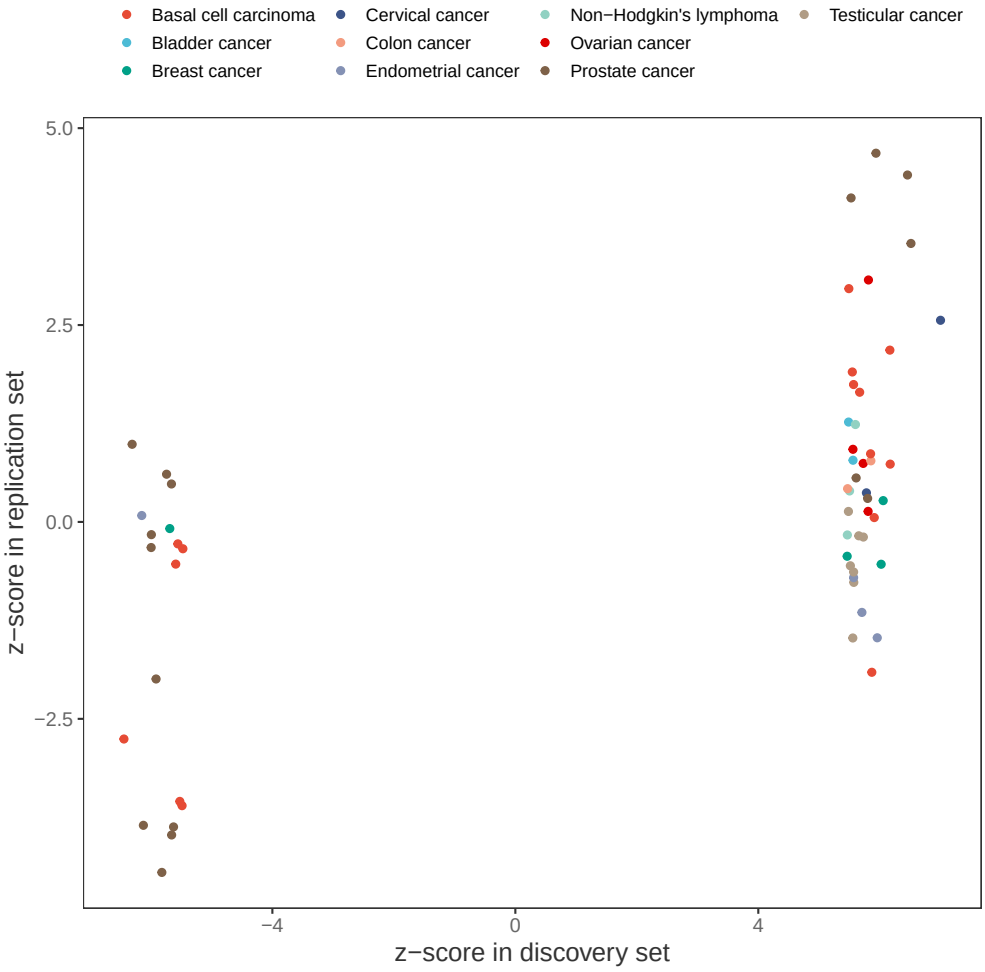

Figure S3. z**-scores for the 59 previously unreported associations in the discovery and replication data sets.** The discovery data set results are based on the UK Biobank analysis whereas replication for most traits was done in the Estonian Biobank with the exception of prostate and breast cancers for which we used previously published results combining larger sample sizes. Fisher's exact test testing the independence of the signs of the replication/discovery z-scores resulted in a $p = 0.0016$ indicating that the z-scores coming from replication or discovery sets are dependent.

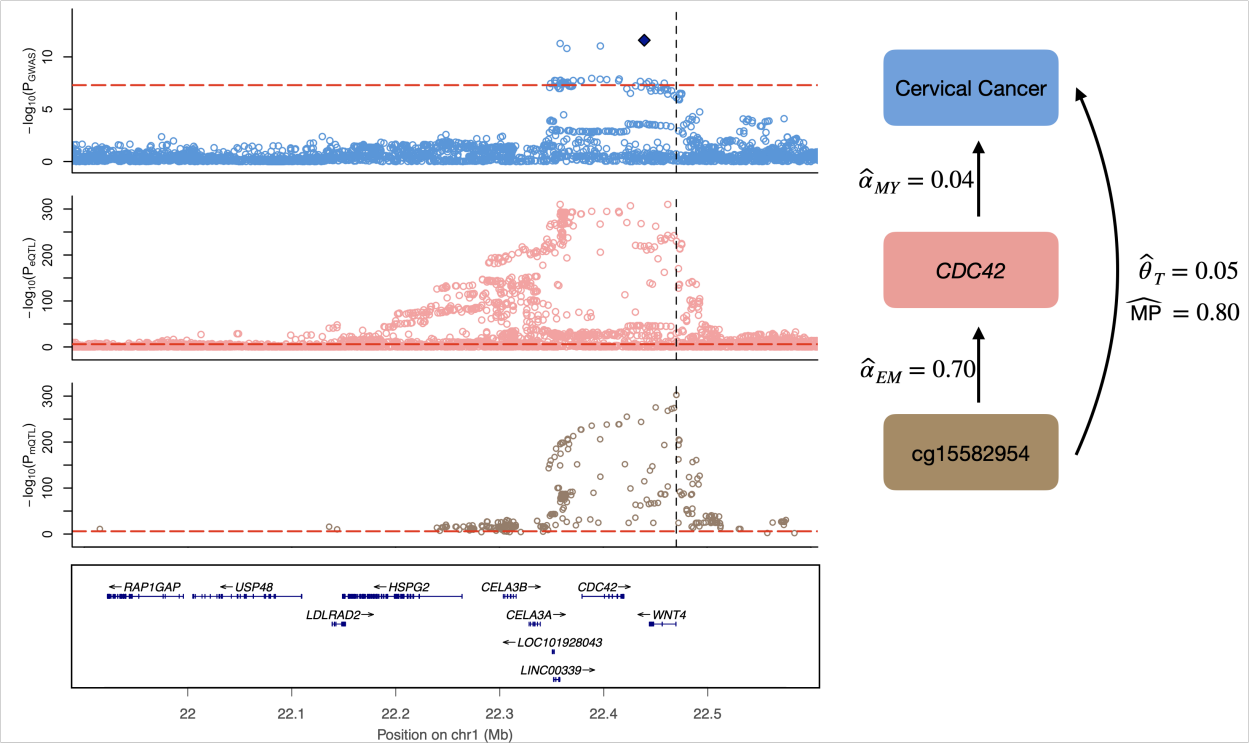

Figure S4. Possible regulatory mechanism between DNAm at cg15582954, *CDC42* gene expression and cervical cancer. The first three rows display the genetic associations (-$\log_{10}$(p-values)) from top to bottom with cervical cancer (blue), *CDC42* transcript levels (pink) and DNAm probe cg15582954 (brown), respectively. The solid diamond in the GWAS locus plot represents the top SNP in LD with the novel GWAS discovery. Red dashed horizontal lines indicate the significance thresholds of the respective SNP associations and the vertical black dashed line represents the DNAm probe position. The bottom row illustrates the positions and strand direction of the genes in the locus. On the right side, a schematic of the regulatory mechanism with calculated MR effects and mediation proportion (MP) is shown.

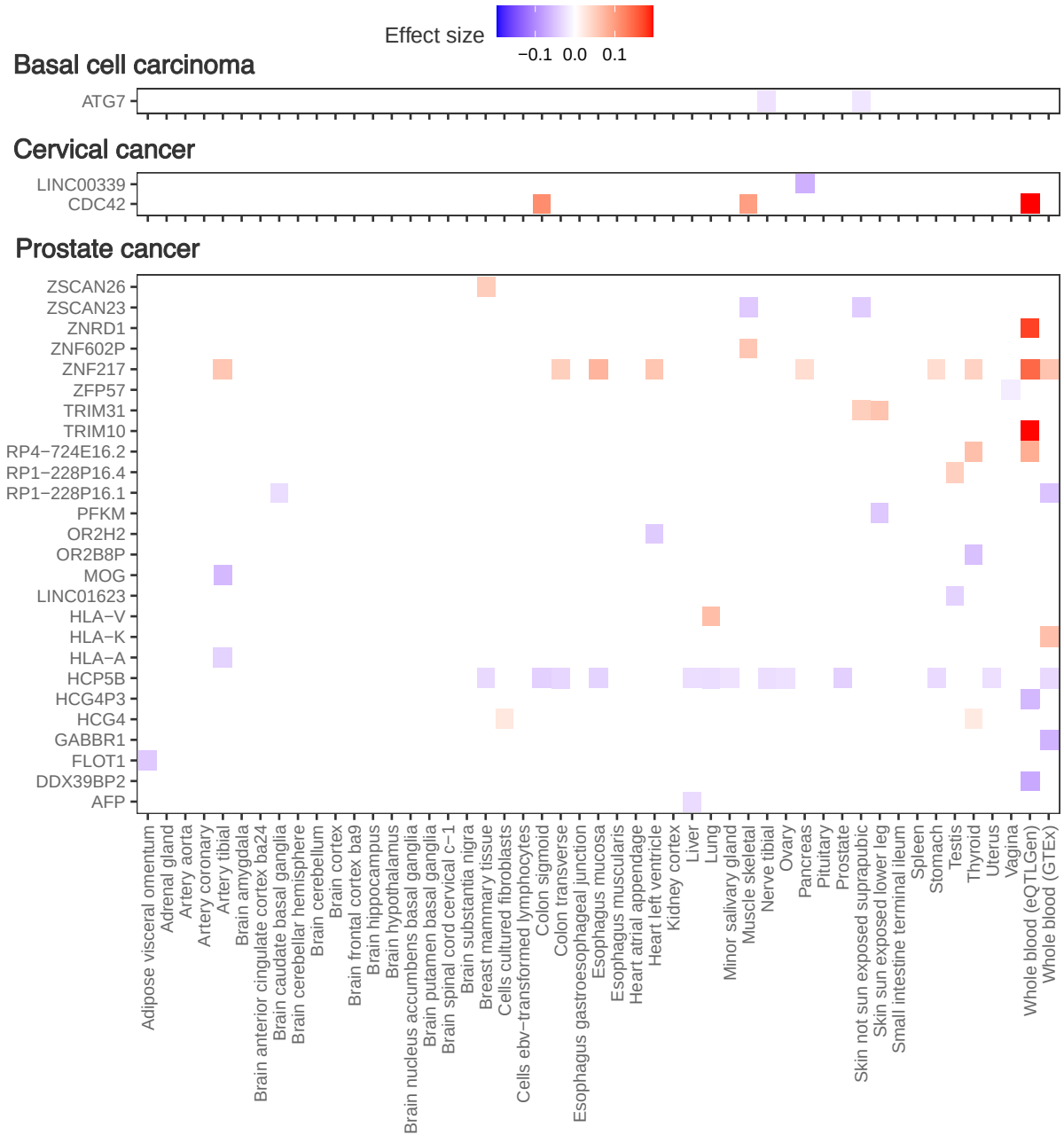

Figure S5. Tissue-specific genetic colocalisation analysis with SMR. We conducted SMR-HEIDI tests using novel SNPs or the SNPs in LD with the novel SNPs as instruments. Plotted are the significant Bonferroni-corrected associations that survived HEIDI filtering.

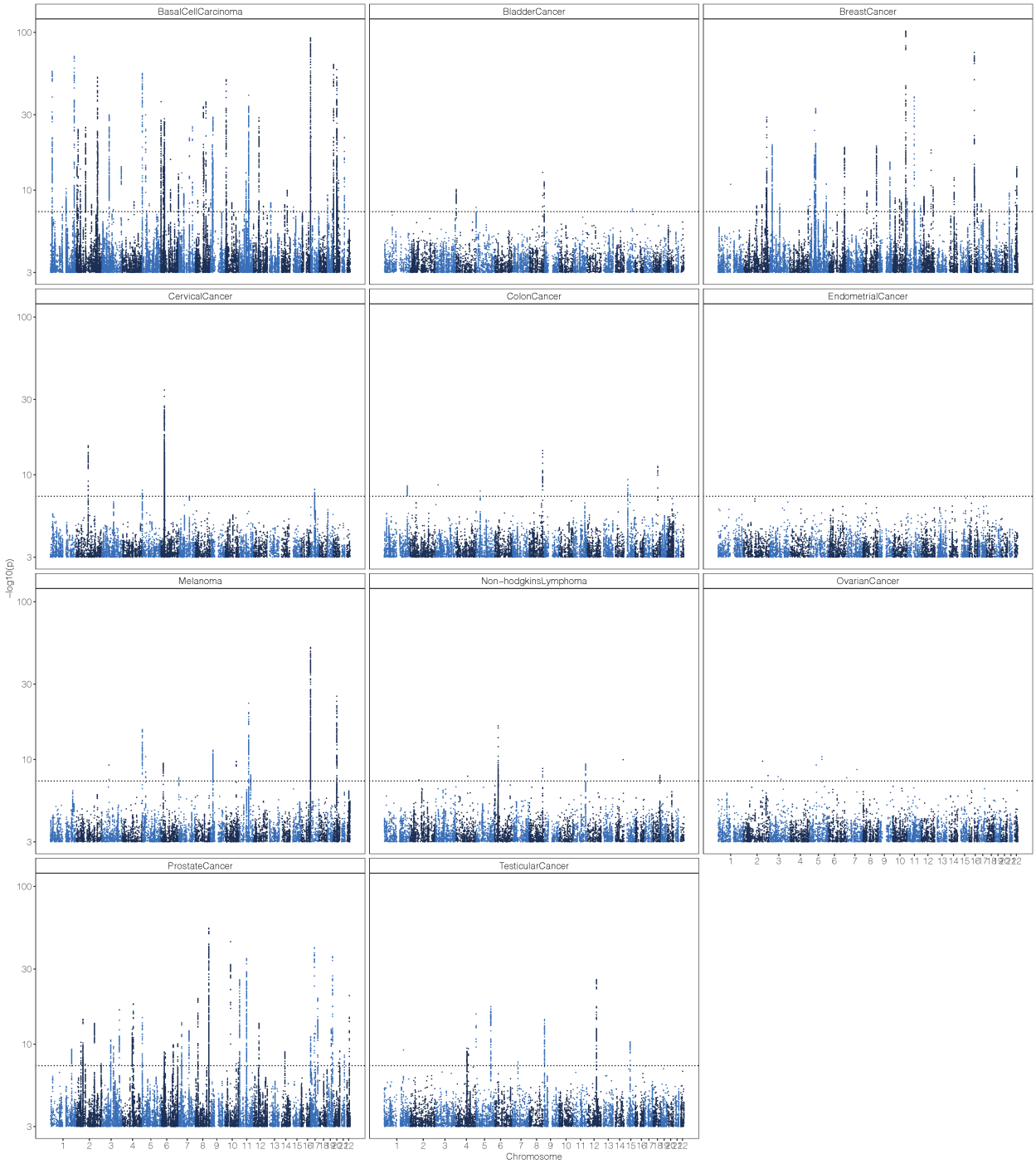

Figure S6. Results from case-control association analysis of 11 tumours, unadjusted for predictors in other chromosomes. The significance of each SNP was tested using binary (case-control) phenotype indicating cancer occurrence that was adjusted only for covariates. Number of markers analysed was M =8,430,446, the number of individuals and cases for each specific cancer are shown in the Supplementary information. For each of the markers a two-sided Wald test was carried out with a null hypothesis of a marker having no effect on the adjusted phenotype. We present the $-\log_{10}$(p-value), the dotted line indicates a significance threshold of $p = 5 \cdot 10^{-8}$.

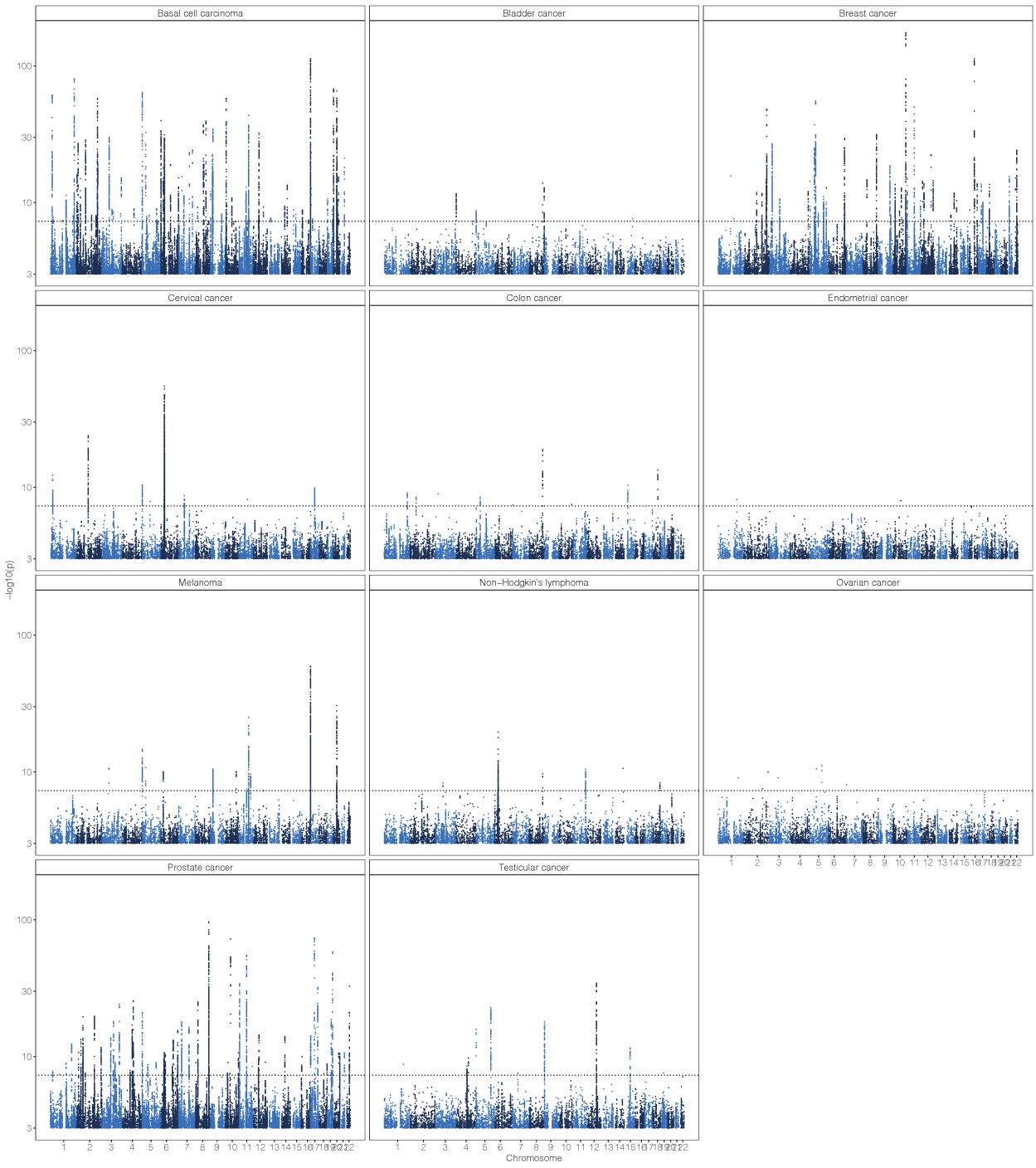

Figure S7. Results from case-control association analysis of 11 tumours, adjusted for BayesW predictors in other chromosomes. The significance of each SNP was tested using binary (case-control) phenotype indicating cancer occurrence that was adjusted for covariates and BayesW genetic predictor using the effects from all of the other chromosome (leave out one chromosome). Number of markers analysed was M =8,430,446, the number of individuals and cases for each specific cancer are shown in the Supplementary information. For each of the markers a two-sided Wald test was carried out with a null hypothesis of a marker having no effect on the adjusted phenotype. We present the $-\log_{10}$(p-value), the dotted line indicates a significance threshold of $p = 5 \cdot 10^{-8}$.

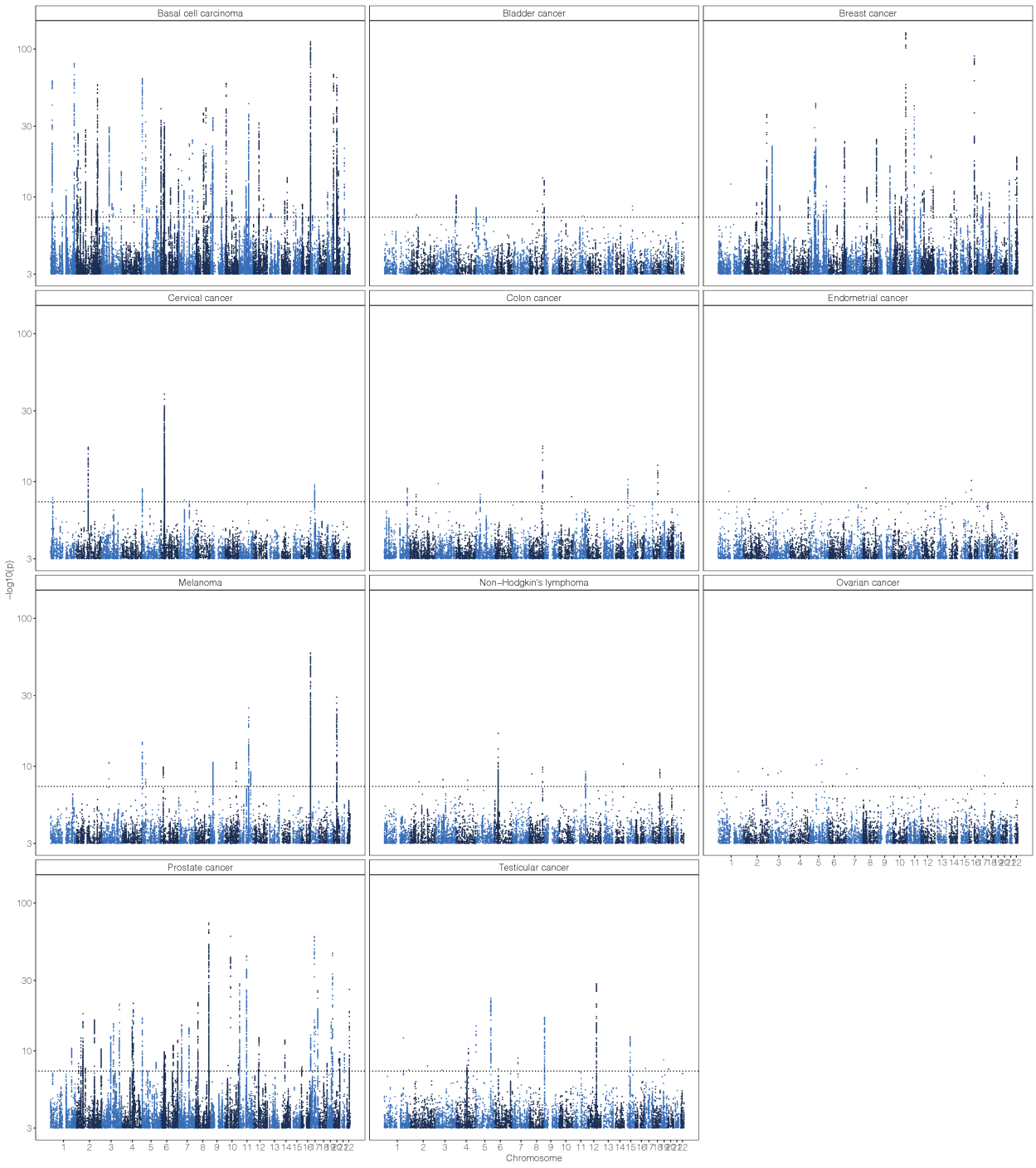

Figure S8. Results from case-control association analysis of 11 tumours, adjusted for BayesRR-RC predictors in other chromosomes. The significance of each SNP was tested using binary (case-control) phenotype indicating cancer occurrence that was adjusted for covariates and BayesRR-RC genetic predictor using the effects from all of the other chromosome (leave out one chromosome). Number of markers analysed was M =8,430,446, the number of individuals and cases for each specific cancer are shown in the Supplementary information. For each of the markers a two-sided Wald test was carried out with a null hypothesis of a marker having no effect on the adjusted phenotype. We present the $-\log_{10}$(p-value), the dotted line indicates a significance threshold of $p = 5 \cdot 10^{-8}$.

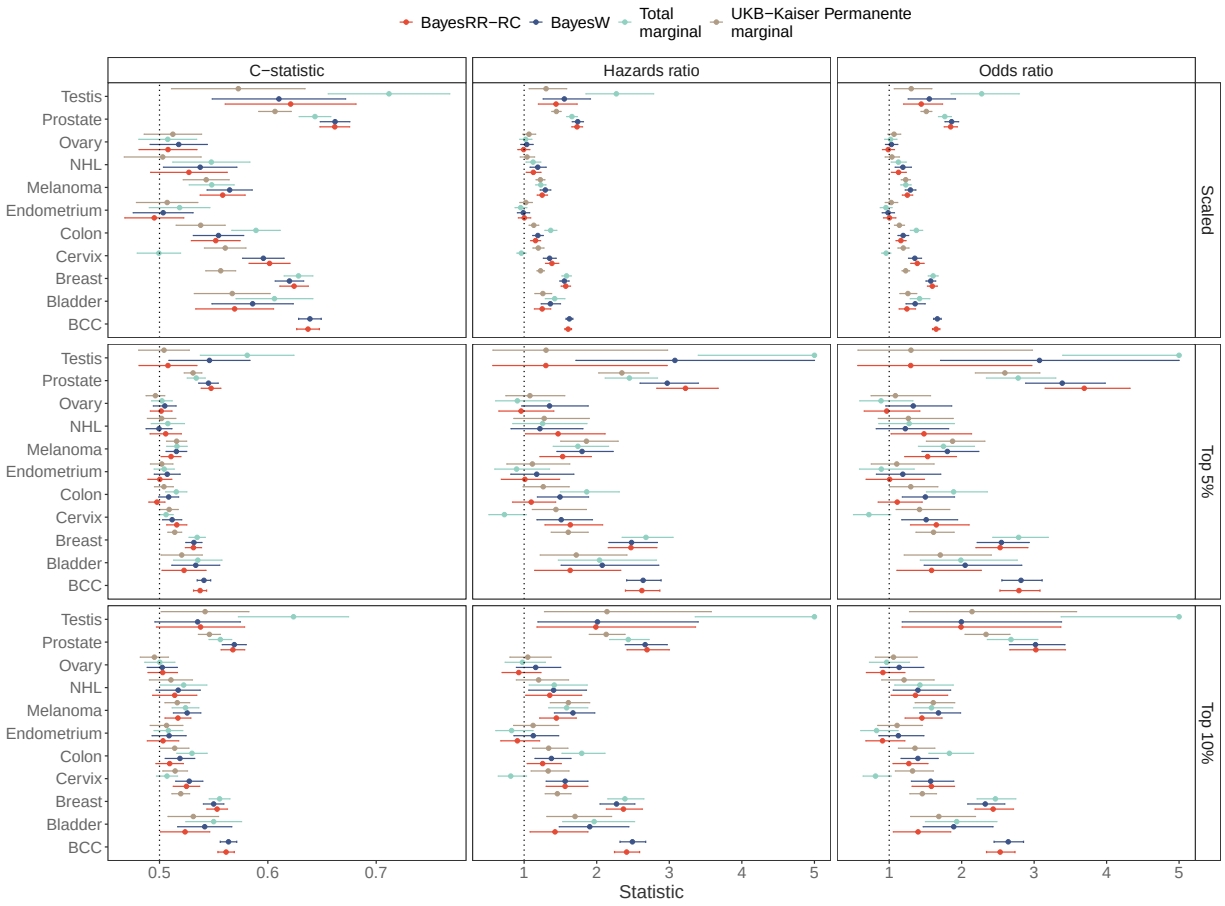

Figure S9. Predictive validation of different PRS on Estonian Biobank data using Harrell's C-statistic, hazards ratio or odds ratio with 95% CI. The statistics were calculated by finding the impact of one standard deviation increase in the PRS (Scaled), by finding the impact of belonging to top 5% quantile of the PRS or by finding the impact of belonging to the top 10% quantile of the PRS on the likelihood of having cancer. Harrel's C-statistic was calculated from Cox proportional hazards model without covariates, odds ratio was calculated from a logistic model using sex and age-at-entry as covariates, hazards ratio was calculated from Cox proportional hazards model using sex and age-at-entry as covariates.

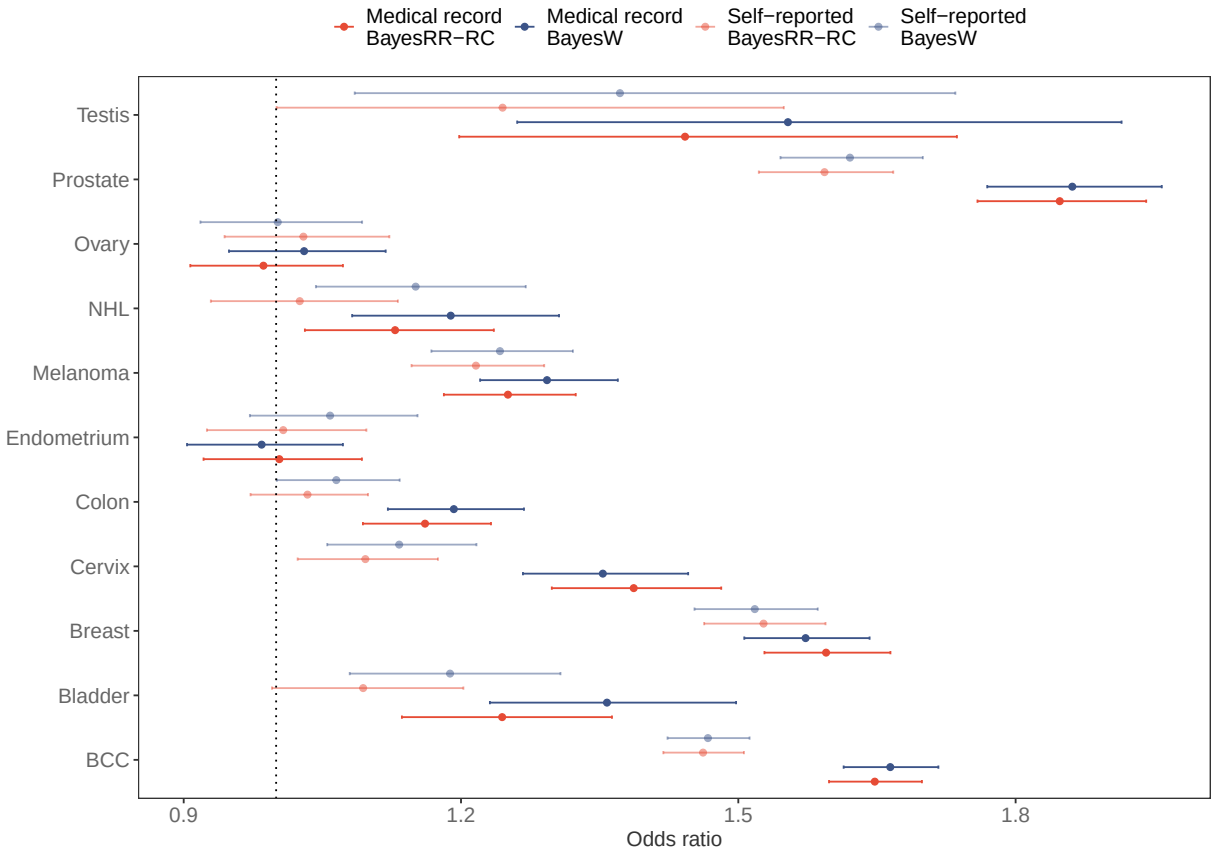

Figure S10. Prediction in Estonian Biobank using either medical record or self-reported phenotypic data in BayesW or BayesRR-RC models. The polygenic risk scores that are using medical record data rather than self-reported data tend to be more predictive across all cancers. The odds ratios were calculated by finding the impact of one standard deviation increase in PRS in a logistic model using sex and age-at-entry as covariates.

Figure S11. HLA-A protein (NCBI identifier $NP\_002107.3$) aligned to Chain F of HLA class I histocompatibility antigen, A-3 alpha chain (PDB: $6ENY\_F$). Three nonsynonymous mutations in the HLA-A complex ($p.Arg68Lys, p.Val91Met, p.Ala174Val$) are marked with yellow colour: all substitutions fall into alpha-3 domains that form the binding groove that holds a peptide for presentation to CD8+ T-cells.

References

1. Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).

2. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer* **17**, 692–704 (2017).

3. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581–590 (2018).

4. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics* **104**, 21–34 (2019).

5. Callender, T. *et al.* Polygenic risk-tailored screening for prostate cancer: A benefit–harm and cost-effectiveness modelling study. *PLOS Medicine* **16**, 1–13 (2019).

6. Pashayan, N. *et al.* Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *British Journal of Cancer* **104**, 1656–1663 (2011).

7. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nature Communications* **11**, 4423 (2020).

8. Zhang, Y. D. *et al.* Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications* **11**, 3353 (2020).

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

10. Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature genetics* **53**, 65–75 (2021).

11. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics* **50**, 928–936 (2018).

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

13. Litchfield, K. *et al.* Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor. *Nature genetics* **49**, 1133–1140 (2017).

14. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51**, 1749–1755 (2019).

15. Orliac, E. J. *et al.* Improving gwas discovery and genomic prediction accuracy in biobank data. *bioRxiv* (2021). `https://www.biorxiv.org/content/early/2021/11/08/2021.08.12.456099.full.pdf`.

16. Loh, P.-R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284–290 (2015).

17. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097–1103 (2021).

18. Staley, J. R. *et al.* A comparison of cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics* **25**, 854–862 (2017).

19. Syed, H., Jorgensen, A. L. & Morris, A. P. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics* **17**, 907–915 (2016). PMID: 27248145.

20. Ojavee, S. E. *et al.* Genomic architecture and prediction of censored time-to-event phenotypes with a bayesian genome-wide analysis. *Nature Communications* **12**, 2337 (2021).

21. Pedersen, E. M. *et al.* Accounting for age of onset and family history improves power in genome-wide association studies. *The American Journal of Human Genetics* (2022).

22. He, L. & Kulminski, A. M. Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. *Genetics* **215**, 41–58 (2020).

23. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using regulomedb. *Genome research* **22**, 1790–1797 (2012).

24. Gong, J. *et al.* Pancan-meqtl: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. *Nucleic acids research* **47**, D1066–D1072 (2019).

25. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with chromhmm. *Nature protocols* **12**, 2478–2492 (2017).

26. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* **12**, 931–934 (2015).

27. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics 2018 50:8* **50**, 1171–1179 (2018).

28. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310–315 (2014).

29. Zhu, Z. *et al.* Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics* **48**, 481–487 (2016).

30. CONSORTIUM, T. G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

31. Võsa, U. *et al.* Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics* 1–11 (2021).

32. Min, J. L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature genetics* **53**, 1311–1321 (2021).

33. Maldonado, M. d. M. & Dharmawardhane, S. Targeting Rac and Cdc42 GTPases in Cancer. *Cancer Research* **78**, 3101–3111 (2018).

34. Gagliardi, P. A., Puliafito, A. & Primo, L. PDK1: At the crossroad of cancer signaling pathways. In *Seminars in cancer biology*, vol. 48, 27–35 (Elsevier, 2018).

35. Upadhyay, G. Emerging role of lymphocyte antigen-6 family of genes in cancer and immune cells. *Frontiers in Immunology* **10**, 819 (2019).

36. Pavan, I. C. B. *et al.* On broken ne(c)ks and broken DNA: The role of human NEKs in the DNA damage response. *Cells* **10** (2021).

37. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with fuma. *Nature communications* **8**, 1–11 (2017).

38. Kanehisa, M., Sato, Y. & Kawashima, M. KEGG mapping tools for uncovering hidden features in biological data. *Protein science : a publication of the Protein Society* **31**, 47–53 (2022).

39. Bulik-Sullivan, B. K. *et al.* Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291–295 (2015).

40. Ojavee, S. E., Kutalik, Z. & Robinson, M. R. Liability-scale heritability estimation for biobank studies of low prevalence disease. *medRxiv* (2022). `https://www.medrxiv.org/content/early/2022/02/04/2022.02.02.22270229.full.pdf`.

41. Kachuri, L. *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nature Communications* **11**, 6084 (2020).

42. Wang, Z. *et al.* Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor. *Nature genetics* **49**, 1141–1147 (2017).

43. Keller, A. *et al.* A cost-utility analysis of prostate cancer screening in australia. *Applied health economics and health policy* **15**, 95–111 (2017).

44. Martin, A. J., Lord, S. J., Verry, H. E., Stockler, M. R. & Emery, J. D. Risk assessment to guide prostate cancer screening decisions: a cost-effectiveness analysis. *Medical Journal of Australia* **198**, 546–550 (2013).

45. Howlader, N. *et al.* Seer cancer statistics review, 1975–2013. *Bethesda, MD: National Cancer Institute* **19** (2016).

46. Wolf, A. M. *et al.* American cancer society guideline for the early detection of prostate cancer: update 2010. *CA: a cancer journal for clinicians* **60**, 70–98 (2010).

47. Tasa, T. *et al.* Genetic variation in the estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics* **27**, 442–454 (2019).

48. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nature Genetics* **52**, 458–462 (2020).

49. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics* **51**, 1244–1251 (2019).

50. Patxot, M. *et al.* Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *Nature Communications* **12**, 1–16 (2021).

51. Yang, J. *et al.* Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**, 369–S3 (2012). PMC3593158[pmcid].

52. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76–82 (2011). PMC3014363[pmcid].

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

54. Kamat, M. A. *et al.* PhenoScanner v2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

55. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88**, 294–305 (2011).

56. Surveillance Research Program, National Cancer Institute. SEER*Explorer: An interactive website for SEER cancer statistics. URL `https://seer.cancer.gov/explorer`. Accessed: 2022-01-24.

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

58. Wang, K., Li, M. & Hakonarson, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164–e164 (2010).

59. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

60. Madden, T. The blast sequence analysis tool. *The NCBI handbook* (2003).

61. Wang, J. *et al.* icn3d, a web-based 3d viewer for sharing 1d/2d/3d representations of biomolecular structures. *Bioinformatics* **36**, 131–135 (2020).

62. Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Davey Smith, G. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nature communications* **11**, 1–11 (2020).

63. Sadler, M. C., Auwerx, C., Porcu, E. & Kutalik, Z. Quantifying mediation between omics layers and complex traits. *bioRxiv* (2021). `https://www.biorxiv.org/content/early/2021/10/01/2021.09.29.462396.full.pdf`.

64. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology* **37**, 658–665 (2013).

65. Burgess, S. *et al.* Dissecting causal pathways using Mendelian randomization with summarized genetic data: application to age at menarche and risk of breast cancer. *Genetics* **207**, 481–487 (2017).

66. Greco M, F. D., Minelli, C., Sheehan, N. A. & Thompson, J. R. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in medicine* **34**, 2926–2940 (2015).

67. Harrell Jr, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15**, 361–387 (1996).

68. Gray, B. *cmprsk: Subdistribution Analysis of Competing Risks* (2014). URL `https://CRAN.R-project.org/package=cmprsk`. R package version 2.2-7.

69. Gray, R. J. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics* 1141–1154 (1988).

70. Robinson, M. hydra (version v1.0). *Zenodo* (2021). URL `http://doi.org/10.5281/zenodo.4555238`.

71. Mucci, L. A. *et al.* Familial risk and heritability of cancer among twins in nordic countries. *Jama* **315**, 68–76 (2016).

72. Kilgour, J. M., Jia, J. L. & Sarin, K. Y. Review of the molecular genetics of basal cell carcinoma; inherited susceptibility, somatic mutations, and targeted therapeutics. *Cancers* **13**, 3870 (2021).

73. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *International journal of cancer* **99**, 260–266 (2002).

74. Miller, D. L. & Weinstock, M. A. Nonmelanoma skin cancer in the united states: incidence. *Journal of the American Academy of Dermatology* **30**, 774–778 (1994).