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

Sven E. Ojavee\textsuperscript{1,2,\dagger, *}, Ekaterina S. Maksimova\textsuperscript{3,†}, Kristi Låll\textsuperscript{4}, Marie C. Sadler\textsuperscript{1,2,5}, Reedik Mägi\textsuperscript{4}, Zoltan Kutalik\textsuperscript{1,2,5}, Matthew R. Robinson\textsuperscript{3,*}

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

\dagger 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 represent 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 by an average of 20\% compared to a recent summary statistic approach (LDpred-funct). As sample sizes increase, 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 \cite{1}. Genome-wide association studies (GWAS) aim to attribute risk to regions of the DNA \cite{2} and facilitate polygenic risk score (PRS) calculation \cite{3} to predict subject-specific risk, which may then enable targeted and improved healthcare \cite{4}. There is currently evidence for only 450 genomic regions associated with increased risk of 18 common cancers \cite{2}, despite recent results showing significant non-zero heritability across a range of cancer occurrences \cite{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 \cite{8}. Increasing sample size yields increased statistical power for discovery, with extensive recent case-control studies for breast cancer \cite{9}, prostate cancer \cite{10,11}, ovarian cancer \cite{12}, or testicular cancer \cite{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 [2] 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 would lead to novel discoveries and enhanced genomic prediction within our framework.

### 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 Supplementary 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 (Figures S6, S8, and S7, with GMRM-BayesW resulting in higher power gain as compared to GMRM-BayesRR-RC for most cancer sites (Figure S1). 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 1, 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 $p = 0.002$, 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 1d). Majority of SNPs from these novel regions could be linked to regulatory variation. Namely, 11 out of 16 novel lead SNPs are 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 [28] that are likely to affect binding, or are linked to expressions of a gene target (Figure 1d). In addition, 3 novel lead SNPs and 2 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 [29] (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 [20][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 derived from whole blood as provided by the GoDMC and eQTLGen consortia, respectively, and conducted 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 MR 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 $\hat{\delta}_M \cdot Y = 0.040$, $p = 2.4 \cdot 10^{-8}$) through methylation at the cg15582954 probe (chr1:22'470'343; $E \rightarrow Y$ total effect $\hat{\theta}_T = 0.046$, $p = 3.8 \cdot 10^{-8}$; Figure S3), 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 (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 1), 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 credible 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 adjusted 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 adjusted estimate -0.38, 95%CI (-0.68, -0.07)) (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 used power of joint Bayesian modelling to account for MAF and LD structure to further describe genetic architecture of 11 cancers. We find that all traits are highly polygenic, with most of the $h^2_{NP}$ attributed to SNPs that contribute an average of 0.1% and 0.01% of the group genetic variance for joint BayesRR-RC and BayesW models, respectively (Figure 2a). We find some differences across cancers, notably
1. Genomic prediction of 11 cancers

Next, we used joint BayesW SNP marker estimates to predict cancer occurrence within the Estonian Biobank data (Figure 2, see Methods). We compared our results to those obtained by a baseline joint BayesRR-RC model and also to LDpred-funct approach [41] that using summary statistics from fastGWA method [14] for the same variants using precisely the same individuals as for the Bayesian models (see Methods). Furthermore, 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), illustrating the importance of high data quality and accurate measurement to facilitate phenotype linking across studies.

We find that joint 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, 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 (BayesW) for basal cell carcinoma (BayesW: 1.68, BayesRR-RC: 1.66), bladder cancer (BayesW: 1.35, BayesRR-RC: 1.24), colon cancer (BayesW: 1.21, BayesRR-RC: 1.17), melanoma (BayesW: 1.31, BayesRR-RC: 1.27), non-Hodgkin’s lymphoma (BayesW: 1.19, BayesRR-RC: 1.13), prostate cancer (BayesW: 1.87, BayesRR-RC: 1.85), and testicular cancer (BayesW: 1.50, BayesRR-RC: 1.41) (Figure 2a). A similar trend can be observed when using the C-statistic or hazards ratio for comparison (Figure S9, see Methods). The Bayesian approaches generally have better predictive accuracy than the summary statistic approach using LDpred-funct, with an average of 20% increase in odds ratio. LDpred-funct yielded at least nominally smaller odds ratios (of standard deviation PRS difference) for testicular cancer (LDpred-funct: 1.05), prostate cancer (LDpred-funct: 1.41), breast cancer (LDpred-funct: 1.24) and non-Hodgkin’s lymphoma (LDpred-funct: 1.01) with differences odds ratios differing significantly between BayesW and LDpred-funct for bladder, prostate and breast cancers. For the cases of non-Hodgkin’s lymphoma, bladder cancer and testicular cancer, the LDpred-funct score failed to yield a significantly predictive score.

We observe that the highest 5% PRS quantile discriminates well for the disease occurrence (Figure 3). Whereas the risk of developing prostate cancer by age 85 is estimated to be 11% (Table S7) among the top 5% highest PRS individuals, nearly 58% or 61% will develop prostate cancer according to the BayesRR-RC.
or BayesW model, respectively. In comparison, LDpred-funct PRS finds that 43% 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 25% of the top 5% BayesW PRS gets diagnosed by age 85 (12% in the population, Table S7) and basal cell carcinoma for which 44% of the top 5% BayesRR-RC PRS gets diagnosed by age 85 (31% in the population, Table S7). The share of individuals getting a cancer diagnosis before age 50 is disproportionately higher among individuals with the top 5% or 10% of the PRS across many cancers and risk score types (Figure 3). For example, the top 10% highest genetic risk according to the BayesW risk score account for 23% out of all basal cell carcinoma cases and 29% out of all prostate cancer cases, 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 the utilisation of existing data. Biobanks of large sample sizes are becoming increasingly common worldwide and improved links to electronic health record data enable information to be obtained regarding 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 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 3), and as biobanks primarily consist of young, healthy individuals, it could distort the inference. For example, that issue can be mitigated 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, reducing the sample size, and there is no guarantee that older individuals will not develop cancer later in life. In contrast, time-to-event analysis treats these individuals as right censored, making no additional assumptions about the cancer occurrence in future, which suggests an alternative with a more sound conceptual background to yield more accurate inferences. Interestingly, time-to-event adjustment tends to yield higher power than the case-control adjustment once the case count is sufficiently high. Hence, time-to-event analyses could become more statistically powerful than their case-control counterparts as cases accrue.

There are important caveats to this study. First, our PRS results remain limited as our work represents a re-analysis of a single biobank study to demonstrate the methodological improvements that can be obtained. However, nothing prevents BayesW from 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% [42,43] 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% [44]) 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 [45]. 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) [43]. 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.

A second limitation is our study’s restriction to a discovery set of UK Biobank individuals with European ancestry, whereas many other recent studies have instead focused on merging and meta-analysing multiple data sets from various backgrounds. However, our goal was simply to show that discoveries are yet to be
made on the existing data set simply by using an enhanced methodology for timing-related traits rather
than occurrence-related traits, and these improvements should also transfer when meta-analysing multiple
data sets. Third, the low number of cancer cases, such as testicular, ovarian or endometrial cancer, lead to
sub-optimal prediction accuracy, while cancers with higher case counts (prostate, breast) yielded larger-scale
improvements in prediction accuracy. Again, as stated above, these analyses should be repeated in data with
greater case counts for the cancers with smaller case counts. Fourth, our current analysis combines prevalent
and incident cases to maximise the statistical power. However, it has been shown [7] that effect sizes are
generally 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 the entry date in the analysis,
but the gain is likely marginal if the phenotypic information is derived from the medical records and the
onset happens later in life.

In summary, we have shown random effect models which utilise time-to-event data, maximise the use of
existing data for $h_{SNP}^2$ 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 $\leq 0.9$ as missing. From the good quality
markers (with missingness less than 5% and p-value for Hardy-Weinberg test larger than $10^{-9}$, 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
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

\begin{align}
C_i = \begin{cases}
0 & \text{if the person had the event before the end of the follow-up period} \\
1 & \text{otherwise}
\end{cases}
\end{align}
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 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 (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 [46]. We kept only the European ancestry individuals with available information on sex, age-at-recruitment and date of recruitment. 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

Both GMRM-BayesRR-RC and GMRM-BayesW models are based on grouped effects with spike-and-slab mixture priors and provide joint effect estimates [20,47]. Briefly, genetic markers are grouped into annotation-specific sets, e.g. based on MAF or LD, with independent hyper-parameters for the phenotypic variance attributable to each group. This allows estimation of the phenotypic variance attributable to the group-specific effects.

Assuming $N$ individuals, $M$ genetic markers, and $\Phi$ groups, the GMRM-BayesRR-RC model of an observed phenotype vector $y$ is:

$$y = 1\mu + \sum_{\varphi=1}^{\Phi} X_{\varphi}\beta_{\varphi} + \varepsilon,$$

where $X_{\varphi}$ is a standardised genotype matrix containing SNPs allocated to group $\varphi$, $\mu$ is an intercept, $\beta_{\varphi}$ is the vector of SNP effects in group $\varphi$ and $\varepsilon$ is a vector of Gaussian residuals such that every element in the vector would be distributed as $\varepsilon_i \sim N(0, \sigma_{\varepsilon}^2)$. For each group, we assume that $\beta_{\varphi}$ are distributed according to a mixture of Gaussian components with mixture specific proportions $\pi_{\varphi}$ and mixture variances $\sigma_{\varphi}^2 = \sigma_{\varphi1}^2, \ldots, \sigma_{\varphi L}^2$ and a Dirac delta at zero which induces sparsity:

$$\beta_{\varphi j} \sim \pi_{\varphi 0}\delta_0 + \pi_{\varphi 1}N(0, \sigma_{\varphi 1}^2) + \ldots + \pi_{\varphi L}N(0, \sigma_{\varphi L}^2),$$

![Equation 2](https://via.placeholder.com/150)

---

**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 [46]. We kept only the European ancestry individuals with available information on sex, age-at-recruitment and date of recruitment. 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**

Both GMRM-BayesRR-RC and GMRM-BayesW models are based on grouped effects with spike-and-slab mixture priors and provide joint effect estimates [20,47]. Briefly, genetic markers are grouped into annotation-specific sets, e.g. based on MAF or LD, with independent hyper-parameters for the phenotypic variance attributable to each group. This allows estimation of the phenotypic variance attributable to the group-specific effects.

Assuming $N$ individuals, $M$ genetic markers, and $\Phi$ groups, the GMRM-BayesRR-RC model of an observed phenotype vector $y$ is:

$$y = 1\mu + \sum_{\varphi=1}^{\Phi} X_{\varphi}\beta_{\varphi} + \varepsilon,$$

where $X_{\varphi}$ is a standardised genotype matrix containing SNPs allocated to group $\varphi$, $\mu$ is an intercept, $\beta_{\varphi}$ is the vector of SNP effects in group $\varphi$ and $\varepsilon$ is a vector of Gaussian residuals such that every element in the vector would be distributed as $\varepsilon_i \sim N(0, \sigma_{\varepsilon}^2)$. For each group, we assume that $\beta_{\varphi}$ are distributed according to a mixture of Gaussian components with mixture specific proportions $\pi_{\varphi}$ and mixture variances $\sigma_{\varphi}^2 = \sigma_{\varphi1}^2, \ldots, \sigma_{\varphi L}^2$ and a Dirac delta at zero which induces sparsity:

$$\beta_{\varphi j} \sim \pi_{\varphi 0}\delta_0 + \pi_{\varphi 1}N(0, \sigma_{\varphi 1}^2) + \ldots + \pi_{\varphi L}N(0, \sigma_{\varphi L}^2),$$

![Equation 2](https://via.placeholder.com/150)
where $L$ is the number of mixture components.

The GMRM-BayesW follows similar prior and grouping formulation. However, here we assume for an individual $i$ that the age-at-onset of a disease $y_i$ has Weibull distribution, with a reparametrisation of the model to represent the mean and the variance of the logarithm of the phenotype as

$$E(\log(y_i)|\mu, \beta, \delta, \alpha) = \mu + \sum_{\varphi=1}^{\Phi} X_{\varphi i} \beta_{\varphi} + Z_{i} \delta, \quad (3)$$

$$\text{Var}(\log(y_i)|\mu, \beta, \delta, \alpha) = \frac{\pi^2}{6\alpha^2} \quad (4)$$

where $Z_i$ are additional covariates (such as sex or genetic principal components), $\delta$ are the additional covariate effect estimates and $\alpha$ is the Weibull shape parameter.

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 \cite{47, 49}. 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 \cite{20}; for BayesRR-RC model the choice was exactly the same in \cite{47}. 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 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 \cite{20} and note that these timings were recorded prior to the recent speed-ups in our BayesRR-RC model \cite{15}.

### 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

$$\hat{y}_{k} = \hat{y} - \sum_{l \neq k} \hat{g}_{l}, \quad (5)$$

where $g_{l} = X_{U,K}^{l} \hat{\beta}_{l}, \ X_{U,K}^{l} : N \times M_l$ matrix of SNPs in the $l$th chromosome, $\hat{\beta}_{l}$ is the vector of average effect sizes from joint Bayesian analysis in chromosome $l$, $\hat{y}$ is the binary phenotype that has been adjusted for
covariates. Then, at each chromosome $k$ we fitted the linear model for every marker $j$

$$\bar{y}_k = X_{UKj}^k \beta_j + \varepsilon,$$  \hspace{1cm} (6)

where $X_{UKj}^k$ is the vector of $j$th marker values, $\beta_j$ is the $j$th SNP effect that we are estimating, $\varepsilon \sim N(0, \sigma^2 I_N)$ is the error term with an error variance $\sigma^2$. 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

$$\bar{y} = X_{UKj} \beta_j + \varepsilon.$$  \hspace{1cm} (7)

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 \cite{COJO} from GCTA software \cite{GCTA} 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. (7)). 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 Phenoscaner database \cite{Phenoscaner} 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 \cite{ProstateCancer}, 120,000 breast cancer cases \cite{BreastCancer}. 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 \cite{LDSCORE} 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. \cite{Lee2011} (Table 3) and using the risks from SEER 2016-2018 (Table S7) \cite{SEER} 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 \cite{AlternativeLiability} 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 \cite{CrossTraitLD} 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\(^{57}\), 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\(^{58}\), aligning proteins using Protein BLAST\(^{59}\), and visualising with iCn3D Structure Viewer\(^{60}\).

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. Specifically, we used the full overlap of the UK and Estonian Biobanks (8,430,446 markers) as input, but later restricted the identified associations to those related to markers from the 16 novel regions. 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 (\(< 1\) Mb 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\(^{61}\). 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 \(1,825 \left( p < 2.7 \cdot 10^{-5}\right)\) tests summing across different tissues and instruments.

We further conducted multivariable MR (MVMR) to dissect significant DNAm-to-cancer causal effects (\(\theta_T\)) into direct (\(\theta_D\)) and indirect effects through transcript levels following the methodology outlined in\(^{62}\). While the previous SMR analysis was restricted to the 16 novel regions, the MVMR framework was applied on genome-wide GMRM-BayesW estimates (8,430,446 markers) 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, \(\sim 50,000\) DNAm probes) accounting for correlated instruments\(^{63}\) to obtain DNAm-to-trait total causal effect estimates (\(E \rightarrow \text{outcome} \ Y \ \text{effect} \ (\hat{\theta}_T)\)). 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 $\hat{\alpha}_{EM}$ and associated p-value $p_{EM}$ (significance was defined at $p_{EM} < 0.01$). Analogously to the total causal effect estimation, direct effects $\hat{\theta}_D$ were then calculated in an MVMR regression accounting for correlation among IVs [64]. The mediation proportion ($\hat{MP}$) was then estimated as $1 - \hat{\theta}_D/\hat{\theta}_T$. Causal effect estimates from the transcript mediator on the outcome trait ($\hat{\alpha}_{MY}$) 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 [65]. 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 $\hat{g}$ were obtained by multiplying the standardised genotype matrix with the average SNP effect across iterations

$$\hat{g} = \hat{X}_{Est} \hat{\beta} = \hat{X}_{Est} \hat{\bar{\beta}}, \quad (8)$$

where $X_{Est}$ 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), $\hat{\beta}$ is the $M \times I$ matrix containing the posterior distributions for $M$ marker effect sizes across $I$ iterations, $\hat{\bar{\beta}}$ is the average SNP effect. We calculated the average predictor from BayesW and BayesRR-RC models for each cancer using 1600 iterations (see Analysis with joint Bayesian models). To facilitate comparison between marginal and joint estimates, we used the same training data set (same 2,174,071 variants and 458,747 individuals) from UK Biobank to estimate the marginal summary statistics for the 11 cancers. The summary statistics were estimated using fastGWA [14], taking into account the relatedness in the data. Then, we used these summary statistics in LDpred-funct method [41]. As annotations for LDpred-funct, we used 8 MAF-LD bins similarly to the Bayesian analyses; we used an LD radius of 1Mb; heritability estimates were taken from LDscore regression, using the fastGWA summary statistics. The weights from LDpred-funct were then applied to the Estonian Biobank individuals to create the respective genomic predictors that were compared with the BayesW and BayesRR-RC results.

We evaluated the performance of the three genetic predictors for each cancer phenotype by comparing them to the actual phenotype case-control status using logistic regression and true phenotype timing using Cox proportional hazards (PH) model. The three predictors were compared using the top 5% PRS with the rest, the top 10% PRS with the rest, and 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 hazard ratios and Harrel’s C-statistics [66] 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 calculated the hazard ratio. The Cox PH model calculated...
Harrell's C-statistic, where the true phenotype was regressed only on 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 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 \([67,68]\).

**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](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](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](https://github.com/medical-genomics-group/hydra). 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](https://github.com/masadler/smrivw), plink version 1.9 is available at [https://www.cog-genomics.org/plink/](https://www.cog-genomics.org/plink/). R version 4.0.3 is available at [https://www.r-project.org/](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. This project uses UK Biobank data under project number 35520.
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 \times 10^{-8} and SNPs could be added to a clutch if they were 1Mb from the index SNP, they were correlated 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 r^2 > 0.1 were excluded in the joint model. To determine novelty, we first removed all markers that were significantly associated in the unadjusted model (eq. \( p < 0.05 \)) and they were nominally significant (\( p < 0.05 \)) 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 \times 10^{-8} in the joint model were eliminated. For the remaining SNPs, we conducted an additional literature review using Phenoscan 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.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Cancer</th>
<th>Chr</th>
<th>Position</th>
<th>Nearest gene</th>
<th>Oth</th>
<th>MAF</th>
<th>Effect size</th>
<th>Model</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs73140002</td>
<td>Prostate cancer</td>
<td>20</td>
<td>52175811</td>
<td>RP4-724E16.2</td>
<td>C/T</td>
<td>0.031</td>
<td>-0.024</td>
<td>BW,BRR-RC</td>
<td>0.045</td>
</tr>
<tr>
<td>rs35559666</td>
<td>Prostate cancer</td>
<td>11</td>
<td>76079587</td>
<td>CTD-2392H12.1*</td>
<td>T/C</td>
<td>0.152</td>
<td>0.026</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs10842546</td>
<td>Breast cancer</td>
<td>10</td>
<td>25525513</td>
<td>RNU4-67P*</td>
<td>A/G</td>
<td>0.152</td>
<td>0.026</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs9543212</td>
<td>Prostate cancer</td>
<td>13</td>
<td>73624100</td>
<td>KLF5*</td>
<td>T/C</td>
<td>0.477</td>
<td>-0.018</td>
<td>BW</td>
<td>0.019</td>
</tr>
<tr>
<td>rs68003823</td>
<td>Prostate cancer</td>
<td>4</td>
<td>74325956</td>
<td>GLIS2*</td>
<td>G/A</td>
<td>0.317</td>
<td>-0.016</td>
<td>BW</td>
<td>0.022</td>
</tr>
<tr>
<td>rs2523761</td>
<td>Prostate cancer</td>
<td>1</td>
<td>15498047</td>
<td>ZBTB7B*</td>
<td>C/T</td>
<td>0.005</td>
<td>0.012</td>
<td>BRR-RC</td>
<td>0.015</td>
</tr>
<tr>
<td>rs3829734</td>
<td>Prostate cancer</td>
<td>18</td>
<td>43497513</td>
<td>AFIP*</td>
<td>G/A</td>
<td>0.202</td>
<td>-0.024</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs9543212</td>
<td>Prostate cancer</td>
<td>13</td>
<td>73624100</td>
<td>KLF5*</td>
<td>T/C</td>
<td>0.364</td>
<td>0.019</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs804172</td>
<td>Prostate cancer</td>
<td>16</td>
<td>43497513</td>
<td>GLIS2*</td>
<td>C/G</td>
<td>0.477</td>
<td>-0.018</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs12142017</td>
<td>Prostate cancer</td>
<td>1</td>
<td>15498047</td>
<td>ZBTB7B*</td>
<td>C/T</td>
<td>0.005</td>
<td>0.012</td>
<td>BRR-RC</td>
<td>0.015</td>
</tr>
<tr>
<td>rs10842546</td>
<td>Breast cancer</td>
<td>12</td>
<td>25525513</td>
<td>RNU4-67P*</td>
<td>A/G</td>
<td>0.152</td>
<td>0.026</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs174570</td>
<td>Basal cell carcinoma</td>
<td>11</td>
<td>61597212</td>
<td>FADS2*</td>
<td>T/C</td>
<td>0.128</td>
<td>-0.021</td>
<td>BW,BRR-RC</td>
<td>0.016</td>
</tr>
<tr>
<td>rs113537654</td>
<td>Basal cell carcinoma</td>
<td>16</td>
<td>68939647</td>
<td>TANGO6</td>
<td>C/T</td>
<td>0.179</td>
<td>-0.016</td>
<td>BW</td>
<td>0.022</td>
</tr>
<tr>
<td>rs35661976</td>
<td>Breast cancer</td>
<td>5</td>
<td>12436999</td>
<td>SLC6A18</td>
<td>T/C</td>
<td>0.02</td>
<td>0.062</td>
<td>BW</td>
<td>0.711</td>
</tr>
<tr>
<td>rs10842546</td>
<td>Breast cancer</td>
<td>12</td>
<td>25525513</td>
<td>RNU4-67P*</td>
<td>A/G</td>
<td>0.49</td>
<td>-0.017</td>
<td>BW</td>
<td>1.28 \times 10^{-8}</td>
</tr>
<tr>
<td>rs4654783</td>
<td>Cervical cancer</td>
<td>1</td>
<td>22439520</td>
<td>WNT4*</td>
<td>T/C</td>
<td>0.299</td>
<td>0.022</td>
<td>BW,BRR-RC</td>
<td>2.56 \times 10^{-12}</td>
</tr>
<tr>
<td>rs111842003</td>
<td>Ovarian cancer</td>
<td>3</td>
<td>113086354</td>
<td>CD96</td>
<td>A/T</td>
<td>0.005</td>
<td>0.12</td>
<td>BRR-RC</td>
<td>6.12 \times 10^{-9}</td>
</tr>
<tr>
<td>rs12142017</td>
<td>Prostate cancer</td>
<td>1</td>
<td>15498047</td>
<td>ZBTB7B*</td>
<td>C/T</td>
<td>0.075</td>
<td>-0.033</td>
<td>BW</td>
<td>1.83 \times 10^{-8}</td>
</tr>
<tr>
<td>rs68003823</td>
<td>Prostate cancer</td>
<td>18</td>
<td>76709587</td>
<td>CTD-2392H12.1*</td>
<td>T/C</td>
<td>0.152</td>
<td>0.026</td>
<td>BW,BRR-RC</td>
<td>2.89 \times 10^{-9}</td>
</tr>
<tr>
<td>rs35539666</td>
<td>Prostate cancer</td>
<td>18</td>
<td>76709587</td>
<td>CTD-2392H12.1*</td>
<td>T/C</td>
<td>0.152</td>
<td>0.026</td>
<td>BW,BRR-RC</td>
<td>2.89 \times 10^{-9}</td>
</tr>
</tbody>
</table>

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 in whole blood 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).
<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>BayesW</th>
<th>BayesRR-RC</th>
<th>Array-based  estimate&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Family-based estimate&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td>0.25 (0.19-0.31)</td>
<td>0.25 (0.19-0.31)</td>
<td>0.17 (0.07-0.27)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.43 (0.26-0.59)</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>0.08 (0.01-0.16)</td>
<td>0.09 (0.01-0.17)</td>
<td>0.08 (0.04-0.12)</td>
<td>0.30 (0.00-0.67)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>0.20 (0.15-0.24)</td>
<td>0.16 (0.13-0.20)</td>
<td>0.10 (0.08-0.13)</td>
<td>0.31 (0.11-0.51)</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>0.05 (0.03-0.07)</td>
<td>0.05 (0.03-0.06)</td>
<td>0.07 (0.02-0.12)</td>
<td>0.13 (0.06-0.15)&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>0.08 (0.04-0.12)</td>
<td>0.08 (0.04-0.12)</td>
<td>0.07 (0.04-0.10)</td>
<td>0.15 (0.00-0.45)</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>0.14 (0.06-0.22)</td>
<td>0.08 (0.00-0.16)</td>
<td>0.13 (0.07-0.18)</td>
<td>0.27 (0.11-0.43)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.11 (0.06-0.16)</td>
<td>0.11 (0.06-0.16)</td>
<td>0.08 (0.04-0.11)</td>
<td>0.58 (0.43-0.73)</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>0.03 (0.00-0.09)</td>
<td>0.03 (0.00-0.10)</td>
<td>0.13 (0.03-0.23)</td>
<td>0.10 (0.08-0.10)&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>0.05 (0.00-0.13)</td>
<td>0.02 (0.00-0.10)</td>
<td>0.07 (0.01-0.13)</td>
<td>0.39 (0.23-0.55)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>0.30 (0.23-0.37)</td>
<td>0.25 (0.19-0.31)</td>
<td>0.16 (0.13-0.20)</td>
<td>0.57 (0.51-0.63)</td>
</tr>
<tr>
<td>Testicular cancer</td>
<td>0.38 (0.19-0.58)</td>
<td>0.29 (0.11-0.47)</td>
<td>0.26 (0.15-0.38)</td>
<td>0.37 (0.00-0.93)</td>
</tr>
</tbody>
</table>

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. <sup>a</sup> - estimate from Rashkin et al. <sup>7</sup>; <sup>b</sup> - estimate from Mucci et al. <sup>70</sup>; <sup>c</sup> - estimate from Kilgour et al. <sup>71</sup>; <sup>d</sup> - estimates from Czene et al. <sup>72</sup>.
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 · 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. The LDpred-funct used the summary statistics that were calculated using the same UK Biobank individuals and variants as for BayesRR-RC or BayesW using fastGWA method; then the summary statistics were used in the LDpred-funct method [41] (see Methods). BayesRR-RC and BayesW tend to have more accurate predictions than the summary statistic approach. For all cancers except breast, cervical, endometrial and ovarian cancer, BayesW predictor gives a nominally higher odds ratios compared to BayesRR-RC predictor.
Table S1. UK Biobank data composition for the cancer cases and their timings used within the study.
<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>ICD10 code</th>
<th>ICD9 code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell</td>
<td>C44.0, C44.1, C44.2, C44.3, C44.4, C44.5, C44.6, 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</td>
<td>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, 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</td>
</tr>
<tr>
<td>Bladder</td>
<td>C67.0, C67.1, C67.2, C67.3, C67.4, C67.5, 188.0, 188.2, 188.4, 188.6, 188.8, 188.9, 233.7</td>
<td>C67.6, C67.6, C67.7, C67.8, C67.9, D09.0, C67.6, C67.7, C67.8, C67.9, D09.0</td>
</tr>
<tr>
<td>Breast</td>
<td>C50.0, C50.1, C50.2, C50.3, C50.4, 174.0, 174.1, 174.2, 174.3, 174.4, 174.5, 174.6, 174.7, 174.8, 174.9, 233.0</td>
<td>C50.5, C50.6, C50.7, C50.8, C50.9, D05.0, D05.1, D05.7, D05.9, C50.5, C50.6, C50.7, C50.8, C50.9, D05.0, D05.1, D05.7, D05.9</td>
</tr>
<tr>
<td>Cervical</td>
<td>C53.0, C53.1, C53.8, C53.9, 180.0, 180.1, 180.8, 180.9, 233.1</td>
<td>C53.3, C53.4, C53.5, C53.6, C53.7, D06.0, D06.1, D06.7, D06.9, C53.3, C53.4, C53.5, C53.6, C53.7, D06.0, D06.1, D06.7, D06.9</td>
</tr>
<tr>
<td>Colon</td>
<td>C18.0, C18.1, C18.2, C18.3, C18.4, C18.5, 153.0, 153.1, 153.2, 153.3, 153.4, 153.5, 153.6, 153.7, 153.8, 153.9, 230.3</td>
<td>C18.6, C18.7, C18.8, C18.9, D01.0, C18.6, C18.7, C18.8, C18.9, D01.0</td>
</tr>
<tr>
<td>Endometrial</td>
<td>C54.1, D07.0, 182.0</td>
<td>C54.1, D07.0, 182.0</td>
</tr>
<tr>
<td>Melanoma</td>
<td>C43.0, C43.1, C43.2, C43.3, C43.4, 172.0, 172.1, 172.2, 172.3, 172.4, 172.5, 172.6, 172.7, 172.8, 172.9</td>
<td>C43.5, C43.6, C43.7, C43.8, C43.9, C43.5, C43.6, C43.7, C43.8, C43.9</td>
</tr>
<tr>
<td>Non-Hodgkin's</td>
<td>C82.0, C82.1, C82.2, C82.7, C82.9, C83.0, C83.1, 202.8</td>
<td>C83.2, C83.3, C83.4, C83.5, C83.6, C83.7, C83.8, C83.9, C85.0, C85.1, C85.7, C85.9</td>
</tr>
<tr>
<td>lymphoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovarian</td>
<td>C56, 183.0</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>C61, D07.5, 185</td>
<td></td>
</tr>
<tr>
<td>Testicular</td>
<td>C62.0, C62.1, C62.9, 186.9</td>
<td></td>
</tr>
</tbody>
</table>

**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.
We observe that for the 59 previously unreported variants, the p-value in the unadjusted association analysis is borderline significant ($5 \times 10^{-6}$ < p < $10^{-5}$) and by using the BayesW or BayesRR-RC adjustments, we arrive at statistically significant test statistics.
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.
<table>
<thead>
<tr>
<th>Trait Combination</th>
<th>BayesW (95% CI)</th>
<th>BayesRR-RC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma - Cervical</td>
<td>0.20 (0.06, 0.35)</td>
<td>0.18 (0.03, 0.33)</td>
</tr>
<tr>
<td>Basal cell carcinoma - Melanoma</td>
<td>0.51 (0.34, 0.68)</td>
<td>0.51 (0.34, 0.67)</td>
</tr>
<tr>
<td>Breast cancer - Endometrial cancer</td>
<td>0.23 (0.03, 0.44)</td>
<td>0.33 (0.03, 0.64)</td>
</tr>
<tr>
<td>Colon cancer - Cervical cancer</td>
<td>0.33 (0.06, 0.60)</td>
<td>0.39 (0.11, 0.68)</td>
</tr>
<tr>
<td>Colon cancer - Endometrial cancer</td>
<td>0.46 (0.09, 0.83)</td>
<td>0.45 (-0.05, 0.94)</td>
</tr>
<tr>
<td>Testicular cancer - Endometrial cancer</td>
<td>-0.38 (-0.68,-0.07)</td>
<td>-0.57 (-1.00, -0.02)</td>
</tr>
</tbody>
</table>

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 [56] 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.
<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Alternative LDSC bW-adjusted</th>
<th>Full Bayesian BayesRR-RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td>0.15 (0.10-0.20)</td>
<td>0.22 (0.20-0.24)</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>0.06 (0.00-0.11)</td>
<td>0.29 (0.19-0.42)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>0.16 (0.11-0.21)</td>
<td>0.20 (0.17-0.23)</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>0.08 (0.05-0.12)</td>
<td>0.13 (0.08-0.20)</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>0.07 (0.03-0.10)</td>
<td>0.19 (0.12-0.28)</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>0.10 (0.03-0.16)</td>
<td>0.34 (0.23-0.47)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.09 (0.04-0.14)</td>
<td>0.18 (0.13-0.24)</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>0.02 (0.00-0.07)</td>
<td>0.32 (0.23-0.43)</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>0.05 (0.00-0.13)</td>
<td>0.49 (0.37-0.61)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>0.22 (0.16-0.29)</td>
<td>0.33 (0.28-0.40)</td>
</tr>
<tr>
<td>Testicular cancer</td>
<td>0.36 (0.15-0.56)</td>
<td>0.81 (0.73-0.87)</td>
</tr>
</tbody>
</table>

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.
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/](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](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. [73].

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Risk (age 0-85)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td>0.3050</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>0.0224</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>0.1248</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>0.0054</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>0.0232</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>0.0300</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.0269</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>0.0161</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>0.0093</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>0.1117</td>
</tr>
<tr>
<td>Testicular cancer</td>
<td>0.0051</td>
</tr>
</tbody>
</table>
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 $-\log_{10}(p\text{-value})$, the dotted line indicates a significance threshold of $p = 5 \cdot 10^{-8}$. 

31
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 $-\log_{10}(p\text{-value})$, the dotted line indicates a significance threshold of $p = 5 \cdot 10^{-8}$. 

---

CC-BY-NC 4.0 International license
It is made available under a CC-BY-NC 4.0 International license.

medRxiv preprint doi: https://doi.org/10.1101/2022.03.25.22272955; this version posted July 19, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
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


