Large-scale integration of omics and electronic health records to identify potential risk protein biomarkers and therapeutic drugs for cancer prevention

Qing Li[1,2†], Qingyuan Song[3,4,†], Zhishan Chen[1,†], Jungyoon Choi[5], Victor Moreno[6,7,8,9], Jie Ping[1], Wanqing Wen[1], Chao Li[1], Xiang Shu[10], Jun Yan[11], Xiao-ou Shu[1], Qiuyin Cai[1], Jirong Long[1], Jeroen R Huyghe[12], Rish Pai[13], Stephen B Gruber[14], Yaohua Yang[15], Graham Casey[15], Xusheng Wang[16], Adetunji T. Toriola[17], Li Li[18], Bhuminder Singh[19], Ken S Lau[19], Li Zhou[20], Zichen Zhang[21], Chong Wu[21], Ulrike Peters[12,22], Wei Zheng[1], Quan Long[2,23,24,25,26*], Zhijun Yin[3,4*], Xingyi Guo[1,3*]

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA
[2] Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Alberta, Canada
[3] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[4] Department of Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[5] Division of Oncology and Hematology, Department of Internal Medicine, Korea University Ansan Hospital, Korea University College of Medicine, Ansan 15355, Korea
[6] Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain
[7] Colorectal Cancer Group, ONCOBELL Program, Institut de Recerca Biomedica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain
[8] Department of Clinical Sciences, Faculty of Medicine and health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona (UB), L'Hospitalet de Llobregat, Barcelona, Spain
[9] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain
[10] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[11] Physiology and Pharmacology, University of Calgary, Calgary, Alberta, Canada
[12] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[13] Mayo Clinic Arizona, Scottsdale, AZ, USA.
[14] Department of Preventive Medicine & USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[15] Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[16] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA
[17] Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine and Siteman Cancer Center, St. Louis, MO, USA
[18] Department of family medicine, School of Medicine, University of Virginia, Charlottesville, VA, USA
[19] Epithelial Biology Center and Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN, USA
[20] Harvard Medical School, Boston, Massachusetts; Division of General Internal Medicine

and Primary Care, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts.
[21] Department of Biostatistics, the University of Texas MD Anderson, Houston, Texas, USA
[22] Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA
[23] Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[24] The Mathison Centre for Mental Health Research & Education, Hotchkiss Brain Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[25] Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[26] Department of Mathematics and Statistics, Faculty of Science, University of Calgary, Calgary, Alberta, Canada

[†] Author names shared co-first authorship
*Corresponding authors

Corresponding Authors:

Dr. Xingyi Guo
Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine
2525 West End Ave. Suite 330, Nashville, TN 37203
Tel: +1 615-936-3471; Fax: +1 615-332-0502
Email: xingyi.guo@vumc.org

Dr. Zhijun Yin
Department of Biomedical Informatics, Department of Computer Science
Vanderbilt University Medical Center
2525 West End Ave. Suite 1475, Nashville, TN 37203
Tel: +1 615-936-3690; Fax: +1 615-322-0102
Email: zhijun.yin.1@vumc.org

Dr. Quan Long
Department of Biochemistry and Molecular Biology,
University of Calgary,
3330 Hospital Drive NW
Room 1151 (1173), Health Sciences Centre
Calgary, AB T2N 4N1 Canada
Tel: 403-220-5580; Fax: 403-210-9538
Email: quan.long@ucalgary.ca

Keywords: Human cancers, risk variants, GWAS, pQTL, proteomics, EHR, therapeutic drug, Caffeine, Acetazolamide

Abstract

Identifying risk protein targets and their therapeutic drugs is crucial for effective cancer prevention. Here, we conduct integrative and fine-mapping analyses of large genome-wide association studies data for breast, colorectal, lung, ovarian, pancreatic, and prostate cancers, and characterize 710 lead variants independently associated with cancer risk. Through mapping protein quantitative trait loci (pQTL) for these variants using plasma proteomics data from over 75,000 participants, we identify 365 proteins associated with cancer risk. Subsequent colocalization analysis identifies 101 proteins, including 74 not reported in previous studies. We further characterize 36 potential druggable proteins for cancers or other disease indications. Analyzing >3.5 million electronic health records, we conducted analyses of emulated trials for 11 drugs across 290 comparisons and identified three drugs significantly associated with reduced colorectal cancer risk: caffeine vs. paroxetine (HR, 0.51; 95% CI, 0.41–0.64), haloperidol vs. prochlorperazine (HR, 0.47; 95% CI, 0.33–0.68), and trazodone hydrochloride vs. paroxetine (HR, 0.49; 95% CI, 0.38–0.63). Conversely, caffeine was associated with increased cancer risk in comparisons with finasteride (colorectal cancer) and fluoxetine (breast cancer). Meta-analysis identified six drugs significantly associated with cancer risk, including acetazolamide, which was associated with reduced colorectal cancer risk (HR, 0.79; 95% CI, 0.72–0.87). This study identifies novel protein biomarkers and candidate drug targets across six major cancer types and highlights several approved drugs with potential chemopreventive effects.

Introduction

Human genetic research has not only advanced our understanding of disease mechanisms but has also significantly contributed to drug discovery and development. Drugs supported by genetic evidence exhibit enhanced therapeutic validity compared to those lacking such support, highlighting the importance of incorporating genetic evidence in drug development initiatives[1; 2]. Common risk variants implicated in diseases can dysregulate nearby gene or protein expression, which can mimic the effects of therapeutic drugs on the targetable proteins. These proteins could serve as potential targets for therapeutic intervention[3]. Thus, concerted efforts for cancer prevention based on proteins influenced by common polymorphisms that modulate cancer risk, are urgently needed[4]. To date, genome-wide association studies (GWAS) have identified several hundred common genetic risk loci for each of three prevalent cancer types: breast, colorectal, and prostate[5-8], and several dozen risk loci have been identified for other cancers, such as cancer of lung, pancreas, and ovarian[9-13]. Previous research, including our work, has identified hundreds of putative cancer susceptible genes potentially regulated by these risk variants, using methods such as expression quantitative trait loci (eQTL) analysis[8-12; 14-20] and transcriptome-wide association studies (TWAS)[7; 19; 21-29]. However, most dysregulated gene expressions have not been thoroughly investigated at the protein level.

To deepen the understanding of causal mechanisms and enhance drug discovery endeavors, it is imperative to explore data from transcriptomic to proteomic studies. Proteins, the ultimate products of mRNA translation, play critical roles in cellular activities and represent promising therapeutic targets, as evidenced by successful drug targeting of enzymes, transporters, ion channels, and receptors[30]. Recent studies include protein quantitative trait loci (pQTL) mapping and Mendelian randomization (MR) analysis by integrating cancer GWAS and blood proteomics data to identify potential risk proteins. However, only a few dozen of cancer risk proteins have been reported, with a false discovery rate < 0.05[31-36]. Most reported proteins have not been directly linked to the GWAS-identified risk variants in common cancer types. Furthermore, research is lacking in integrating multiple population-scale proteomic studies like the recent emerging UK Biobank Pharma Proteomics Project (UKB-PPP)[37], which offers an unprecedented opportunity to establish extensive pQTL databases, accelerating therapeutic drug discovery for therapeutic prevention and intervention in human cancers.

Traditional drug discovery faces numerous challenges, including escalating costs, lengthy timelines, and high failure rates[38]. Drug repurposing presents a promising strategy

by identifying new applications for existing drugs, leveraging their well-documented characteristics[39]. With the widespread adoption of modern electronic health record (EHR) systems, vast amounts of real-world patient data are available to augment pre-clinical outcomes and facilitate drug repurposing screening. Recently, drug repurposing using EHRs has successfully discovered repurposing hypotheses for preventing Alzheimer's Disease[40], reducing cancer mortality[41; 42], treating COVID-19[43; 44], and coronary artery disease[45]. However, for therapeutic drugs that have been used for a long term to treat disease indications with evidence of affecting the expression of cancer risk proteins, their potential association with the risk of human cancers remains largely unclear. Some of these drugs may be linked to an increased cancer risk due to long-neglected side effects.

In this work, we integrate large GWAS data for breast, colorectal, lung, ovarian, pancreatic, and prostate cancers and population-scale proteomics data from over 75,000 participants combined from Atherosclerosis Risk in Communities study (ARIC)[46], deCODE genetics[47], and UKB-PPP to identify risk proteins associated with each cancer. We further characterized therapeutic drugs based on druggable risk proteins targeted by approved drugs or undergoing clinical trials for cancer treatment or other indications. We further evaluate the effect of cancer risk for those drugs approved for the indications, using over 3.5 million EHR database at Vanderbilt University Medical Center (VUMC). Findings from this study offer novel insights into therapeutic drugs targeting risk proteins for cancer prevention and intervention.

Material and Methods

Characterization of lead variants in six types of cancer

We performed a comprehensive analysis to characterize lead genetic variants associated with breast, colorectal, lung, ovary, pancreas and prostate (**Figure 1A**). For breast cancer, we included strong independent association signals at $P < 1 \times 10^{-6}$ from a fine-mapping study[48] and risk variants from another GWAS[6]. We combined the reported lead variants from these two studies after removing those variants in LD ($r^2 < 0.1$ in European populations) (**Figure S1**). Next, we further included additional lead variants from SuSiE fine-mapping analysis on GWAS (N = 247,173)[49], with fine-mapping windows of 500 kilobases (kb) and allowed a maximum of five causal variants. LD reference was based on the British-ancestry UK Biobank samples (N = 337,000)[50]. We identified a credible set of causal variants with a 95% posterior inclusion probability (95% PIP) for each independent risk signal and a lead variant was represented by the variant with the minimum *P*. We included additional lead variants from our SuSiE analysis with LD $r^2 < 0.1$ in European populations

with the above set of lead variants for those with independent risk-associated signals at GWAS $P < 5 \times 10^{-8}$ and located in GWAS loci with independent risk-associated signals at $P < 1 \times 10^{-6}$ in European populations (**Figure S1**).

For colorectal cancer, we analyzed lead variants from our recent fine-mapping study[51] based on the GWAS data from 254,791 participants in both European and Asian populations. We characterized lead variants with independent risk-associated signals at minimal $P < 1 \times 10^{-6}$ in European populations, from the analysis based on GWAS from trans-ancestry and European populations, respectively (**Figure S1**). For prostate cancer, we first identified lead variants with independent risk-associated signals at GWAS $P < 5 \times 10^{-8}$ from our SuSiE fine-mapping analysis on GWAS summary statistics (N = 140,306). We also included additional GWAS-identified risk variants with $P < 1 \times 10^{-6}$ in European populations from the previous trans-ancestry GWAS[8] and $r^2 < 0.1$ with any lead variants from the above set in the fine-mapping analysis (**Figure S1**). Similarly, we used the above strategy to characterize lead variants from fine-mapping analysis for ovarian cancer (N = 63,347) and pancreatic cancer (N = 21,536). We next included additional risk variants that are missed in the above set of lead variants from previous GWAS for ovarian[11] and pancreatic cancer[10]. For lung cancer, we included risk variants at $P < 1 \times 10^{-6}$ in European populations from the trans-ancestry GWAS (N=85,716)[9]. Further details on lead variant identification are summarized in **Figure S1** and details on GWAS resources can be found in **Table S1** and **Supplemental Methods I.**

Identification of putative target proteins for lead variants

To identify potential cancer risk proteins, we mapped GWAS lead variants to *cis*-pQTLs (+/- 500Kb region of a gene) results from three studies among European populations: UK Biobank Pharma Proteomics Project[37], Atherosclerosis Risk in Communities study[44] and deCODE genetics[45] (**Figure 1A**). To increase the power of pQTLs, we combined *cis*-pQTLs from the ARIC and the deCODE (both assayed through SOMAscan® platform) via a a fixed-effects meta-analysis using META[52]. *Cis*-pQTLs from the UKB-PPP (assayed through Olink platform) were independently analyzed. In a few cases where the lead variant did not overlap with any *cis*-pQTLs, we substituted it with the correlated variant exhibiting the strongest association signal. Putative cancer risk protein was defined based on pQTL significance at a Bonferroni threshold of $P < 0.05$ (nominal $P = 2.3 \times 10^{-5}$, corresponding to 2,164 variant-protein tests for UKB-PPP; nominal $P = 3.8 \times 10^{-5}$, corresponding to 1,322 variant-protein tests for ARIC+deCODE).

Colocalization analyses between pQTL and GWAS signals

To identify cancer risk proteins, we conducted colocalization analysis using two approaches: Bayesian method *coloc* [53] and summary data-based Mendelian Randomization (SMR)[54]. For the SMR approach, a followed HEIDI test is performed on significant SMR results to determine if the colocalized signals can be explained by one single causal variant or by multiple causal variants in the locus. For each protein, SNPs with $P < 0.5$ from GWAS, MAF > 0.01, and within 50 kb of the lead variant were included. To estimate the posterior probability (PP) of colocalization, we utilized the default priors and coloc.abf function. In our study, we particularly focused on the assumption that one genetic variant is simultaneously associated with both two traits, which was quantified by PP.H4. We considered a protein to host one shared causal variant from GWAS and pQTLs if its *coloc* PP.H4 > 0.5. Additionally, we also performed SMR+HEIDI analysis for significant *cis*-pQTL with default parameter settings. Specifically, significant SMR+HEIDI results were defined as a tested locus with Bonferroni-adjusted SMR $P < 0.05$ and HEIDI $P \geq 0.05$ (no obvious evidence of heterogeneity of estimated effects or linkage). The above analyses were only conducted in European populations, which were available for both GWAS and proteomics in European populations.

Functional genomic analyses

For our identified cancer risk proteins, we examined their xQTLs, including eQTLs, sQTLs, and apaQTL using the resource from the GTEx (version 8). We collected eQTLs and sQTLs from six normal tissues and whole blood from GTEx studies, and we collected apaQTLs from Li's work[55]. A nominal P value < 0.05 for at least one xQTL in either tissue or blood samples was considered supportive of the pQTL results.

We identified putative regulatory variants in strong linkage disequilibrium (LD) ($r^2 >$ 0.8 in European population) for lead variants with significant colocalization between GWAS and *cis*-pQTL signals. Using the HaploReg tool[56], we annotated these variants with a variety of epigenetic annotations, including regulatory chromatin states based on DNAse and histone ChIP-Seq from Roadmap Epigenomics Project, histone marks for promoter and enhancer, binding sites of transcription factors, and gene annotation from the GENCODE and RefSeq. We denoted variants as "Proximal" if they overlapped with these functional annotations near the closest target gene. We analyzed a variety of chromatin-chromatin interaction data, from 4D genome[57], FANTOM5[58], EnhancerAtlas[59], and super-enhancer[60]. We examined the overlap between putative regulatory variants and enhancer elements in corresponding cell lines or tissues of these six cancer types. We further determined enhancer-promoter loops after combining these data with ChIP-seq data of the histone modification H3K27ac (an active enhancer mark). We focused on interacted loops in which a

fragment overlapped an H3K27ac peak (enhancer-like elements). In contrast, the other fragment overlapped the promoter of a gene (defined as a region of upstream 2kb and downstream 100bp around transcript start site). We denoted variants as “Distal” if they overlapped with these chromatin-chromatin variants.

For our identified cancer risk proteins, we assessed the statistical significance of their differential protein expression between tumor and normal tissue in breast, colorectal, lung, ovarian, and pancreatic cancer samples using data from CPTAC, accessed through the UALCAN website[61; 62]. Similarly, we analyzed their differential gene expression between tumor and normal tissue using data from TCGA, also through the UALCAN website.

Identification of focal drugs and formulation of drug-treated patient groups

We assessed the therapeutic relevance of cancer risk proteins by mapping them to known drug–target pairs using four major databases: DrugBank[63], ChEMBL[64], the Therapeutic Target Database (TTD)[65] and Open Targets[66]. This approach identified multiple drugs targeting cancer risk proteins, which we refer to as focal drugs (**Figure 1B**). To evaluate their potential impact on cancer development, we conducted comparative analyses between focal drugs and corresponding control drugs. To minimize potential confounding factors associated with a drug prescription, we selected control drugs that belong to the same second-level Anatomical Therapeutic Chemical classification category (ATC-L2) as the focal drug. For each focal drug, we constructed emulated clinical trials, each comprising a treated group (patients prescribed the focal drug) and a control group (patients prescribed a control drug). One focal drug may have multiple trials, depending on the number of potential control drugs belonging to the same ATC-L2 category. After defining focal-control drugs pairs, we selected patients from the Synthetic Derivative (SD), a de-identified electronic health record database of 3.5 million individuals at Vanderbilt University Medical Center (VUMC)[67]. Using strict eligibility criteria (**Supplementary Methods II**), we ensured each trial included at least 500 patients in both the treated and control groups[40].

Calculation of overall hazard ratio for cancer risk drug on cancer development risk

To rigorously evaluate the impact of focal drugs on cancer development, we applied the Inverse Probability of Treatment Weighting (IPTW) framework[68] to reduce confounding and ensure balanced baseline characteristics between the treated and control groups (**Supplementary Methods III, IV**). Within each eligible trial emulation, we used the weighted Cox proportional hazards model to estimate the hazard ratio (HR) for cancer risk, comparing patients exposed to the focal drug versus those exposed to the control drug over a ten-year follow-up period (**Figure 1C**). For each focal drug, we also performed a random-effects

meta-analysis across all balanced trials to estimate its overall association with cancer risk (**Supplementary Methods V**). Finally, to assess the robustness of our findings, we conducted sensitivity analyses using a five-year follow-up window and an alternative control drug selection based on third-level Anatomical Therapeutic Chemical (ATC-L3) classifications, which specifies the pharmacological or chemical properties within the therapeutic subgroup defined at ATC-L2.

Results

Characterizing lead variants for breast, ovarian, prostate, colorectal, lung, and pancreas cancers

For breast cancer, we included 196 lead variants with independent association signals at loci from a previous fine-mapping study[48] and additional 32 genetic variants from another GWAS[6]. Our additional fine-mapping analysis using SuSiE[49] on Breast Cancer Association Consortium GWAS (N = 247,173) resulted in five novel lead variants. After integrating the previous results with new fine-mapping efforts, we identified 227 lead variants with independently associated with cancer risk at each locus. We applied similar strategies to identify lead variants for colorectal[51] cancer and other cancer types, leveraging both previously published GWASs and our fine-mapping analyses (**Material and Methods**). In total, we identified 710 lead variants, including 227 for breast cancer, 213 for colorectal cancer, 213 for prostate cancer, 26 for lung cancer, 13 for ovarian cancer and 18 for pancreatic cancer (**Table S2**).

Identifying cancer risk proteins from pQTLs mapping and colocalization analyses

We mapped the 710 lead variants to *cis*-pQTLs to identify cancer risk proteins. At a Bonferroni-corrected $P < 0.05$, we identified a total of 459 pQTL association signals (corresponding 365 proteins after combined proteins unique for each cancer) for 222 lead variants across six cancer types, including 74 for breast, 127 for colorectal, 37 for lung, 5 for ovarian, 9 for pancreatic, and 113 for prostate cancer (**Figure 2; Table S3**). Notably, 312 of the identified proteins (85.4% of 365) among these cancer types have not been reported in previous proteomics-based MR studies[32-36; 69; 70] (**Table S4**). Furthermore, through analysis of the identified proteins commonly observed in multiple cancers, we found that 60 proteins were commonly observed in at least two of these six cancers (**Figure S2**). In particular, we observed that several well-known cancer-related proteins, such as HLA-A and HLA-E, were linked to lead variants located in major histocompatibility complex (MHC) in breast, colorectal and lung cancers, highlighting the potential role of these proteins in cancer pleiotropy and shared cancer risk mechanisms (**Figure 2**).

A further colocalization analysis identified 101 proteins after combined proteins unique for each cancer that showed strong evidence supported by either colocalization or SMR+HEIDI analysis (**Material and Methods**). Specifically, we identified 23 proteins for breast, 38 proteins for colorectal cancer, 7 proteins for lung, 2 for ovarian, 2 for pancreatic and 29 for prostate cancer, respectively (**Figure 3A, 3B**, **Table S5**). Of these, 74 proteins (73.2% of 101) have not been previously linked to cancer risk (**Table S6**). Of note, 71 proteins were only assayed by either SOMAscan® (n=32) or Olink platform (n=39). For the remaining 22 significant proteins commonly assayed, all showed a pQTL significance signal with a minimal nominal $P < 1 \times 10^{-5}$ in both ARIC+deCODE and UKB-PPP ($r = 0.66$, $P = 2\times10^{-4}$; **Figure 3C**). In particular, seven proteins were highlighted as cancer-driver proteins[71; 72] and Cancer Gene Census (CGC)[73], including ALDH2, HLA-A and SUB1 for breast cancer, ALDH2 and HLA-A for colorectal cancer, NT5C2 for lung cancer, and NT5C2, RNF43, TYRO3 and USP28 for prostate cancer.

Cancer risk proteins supported by functional genomics analyses

Of the identified 101 proteins among the six cancers, we examined whether they are supported by functional genomics analyses. Specifically, we first evaluated xQTL (i.e., eQTLs, alternative splicing - sQTLs, and alternative polyadenylation - apaQTLs) results in their respective target tissues and whole blood samples (**Material and Methods**). We found 63 proteins that were supported by at least one xQTLs at a nominal $P < 0.05$, including12 for breast (52% of 23), 22 for colorectal (57% of 38), 5 for lung (71% of 7), 2 for ovarian, 2 for pancreatic, and 20 (68% of 29) for prostate cancer (**Table S6**). Second, we used functional genomic data generated in their cancer-related tissues/cells (i.e., promoter and enhancer) to characterize putative functional variants that are in strong LD ($r^2 > 0.8$ in the European population) with the lead variants (**Material and Methods**) Our results showed that 17 genes were likely regulated by the closest putative regulatory variants with either promoter and/or enhancer activities (**Table S7**). We further investigated the potential distal regulatory effects of putative functional variants on these genes by analyzing chromatin interaction data (**Material and Methods**). We found that 39 genes were regulated distally by putative functional variants through long-term promoter-enhancer interactions (**Table S8**). Lastly, we examined differential protein expression between normal and tumor tissues available for breast, colon, lung and pancreatic cancers using data from Clinical Proteomic Tumor Analysis Consortium (CPTAC). We showed evidence of the 18 identified proteins with consistent association directions supported by significantly differential expression at a nominal $P < 0.05$, including 3 for breast cancer, 10 for colorectal cancer, 4 for lung cancer and 1 for pancreatic cancer. Similarly, we showed evidence of the 40 identified proteins

supported by significantly differential mRNA expression using data from The Cancer Genome Atlas Program (TCGA), including 6 for breast cancer, 15 for colorectal cancer, 3 for lung cancer, 1 for pancreatic cancer and 15 for prostate cancer. Taken together, our analysis provided additional evidence that most of the identified proteins are partially or wholly supported by functional genomics analyses (**Table S9)**.

Identifying druggable proteins

Using data from DrugBank[74], ChEMBL[75], the Therapeutic Target Database[76] (TTD) and OpenTargets[77], we comprehensively annotated our proteins as therapeutic targets of approved or clinical-stage drugs. Of the 101 proteins among the six cancers, we identified 36 druggable proteins potentially targeted by 404 approved drugs or undergoing clinical trials for cancer treatment or other indications (**Figure 4**, **Table S10**). Specifically, we found 19 proteins targeted by 133 drugs either approved or under clinical trials to treat cancers (**Figure 5, Table S11**). Our results also provide evidence that the remaining draggable proteins are targeted by 197 drugs used for treating indications other than cancer (**Figure S3**).

Evaluating associations of drugs approved for indications with cancer risk

We next evaluated the effect on cancer risk of therapeutic drugs that have been used long-term to treat non-cancer indications based on real-world EHRs from the VUMC Synthetic Derivative (SD) database. Following these stringent criteria descried in **Material and Methods**, we formed 290 balanced trials from 11 treated drugs, associated with seven proteins from three cancers (**Figure 6A**), each with more than 10 eligible control drugs. Specifically, we identified three drugs for breast cancer: Caffeine targeting ALDH2, and Haloperidol and Trazodone Hydrochloride with both targeting HLA-A; several drugs for colorectal cancer: Caffeine and Theophylline with both targeting ALDH2; Acetazolamide and Captopril, with both targeting TF; Haloperidol and Trazodone Hydrochloride, with both targeting HLA-A; and Vilazodone targeting AGER; and one drug for prostate cancer: Sirolimus targeting TYRO3 (**Figure 6B**). All patients analyzed had substantial medical records during the ten-year follow-up period, with a median follow-up of ~7 years (85 months) and an interquartile range (IQR) of 6-8.5 years (72-102 months). Additionally, across the 290 trials, a median of 99.3% of patients were either cancer-free at the end of the follow-up period (median: 58%) or were lost to follow-up, while the remaining 0.7% were diagnosed with cancer (**Table S13**).

For breast cancer, we applied the weighted Cox proportional hazards model to calculate the hazard ratio (HR) for each trial within the ten-year follow-up period for analyzed three drugs

(Caffeine, Haloperidol, and Trazodone Hydrochloride) across 73 eligible treated-control drug trials. Particularly, at a Bonferroni-corrected $P < 0.05$ ($P < 0.05/290$), we found that Caffeine was associated with an increased risk when compared to Fluoxetine (HR = 1.43, 95% CI = [1.19, 1.71]) (**Table S13**). Furthermore, we conducted a random-effects meta-analysis for each of these three drugs. At a Bonferroni-corrected $P < 0.05$ ($P < 0.05/11$), we found that Caffeine was associated with an increased risk (HR = 1.15, 95% CI = [1.09, 1.20]), while Haloperidol was associated with a decreased risk (HR = 0.86, 95% CI = [0.78, 0.95]) (**Table S14, Figure 6C**).

For colorectal cancer, we analyzed seven drugs (Acetazolamide, Caffeine, Captopril, Haloperidol, Theophylline, Trazodone Hydrochloride, and Vilazodone) across 185 eligible treated-control trials. At a Bonferroni-corrected $P < 0.05$, we found that Caffeine was associated with a decreased risk when compared to Paroxetine (HR=0.51, 95% CI = [0.41, 0.64]) and Magnesium Sulfate (HR=0.36, 95% CI = [0.27, 0.48]). Conversely, Caffeine, when compared to the control drug Finasteride (HR=2.32, 95% CI = [1.67, 3.23]), was associated with an increased risk. We also identified that Haloperidol and Trazodone Hydrochloride were associated with an increased risk when compared to the Prochlorperazine (HR = 0.47, 95% CI = [0.33, 0.68]) and Paroxetine (HR = 0.49, 95% CI = [0.38, 0.63]), respectively (**Table S13**). Further meta-analyses showed that Acetazolamide (HR = 0.79, 95% CI = [0.72, 0.87]) was associated with a decreased risk, while Captopril (HR = 2.15, 95% CI = [1.81, 2.57]) and Vilazodone (HR = 1.73, 95% CI = [1.45, 2.05]) were associated with an increased risk (**Table S14, Figure 6C**).

For prostate cancer, we analyzed one drug (Sirolimus) across 32 eligible treated-control trials. At a Bonferroni-corrected $P < 0.05$, none of the trials comparing Sirolimus to its control drugs showed a significant association with cancer risk. However, our meta-analysis revealed that Sirolimus was significantly associated with an increased risk in prostate cancer (HR = 1.37, 95% CI = [1.2, 1.55]) (**Table S14, Figure 6C**).

In addition, we conducted analyses using the weighted Cox proportional hazards model within the five-year follow-up period for the significant treated-control trials identified above, demonstrating consistent associations across all six trials, at a nominal $P < 0.05$ (**Table S15**). Similarly, with meta-analyses based on the results from the five-year follow-up period, we showed five of the identified six drugs (breast cancer: Caffeine; colorectal cancer: Acetazolamide, Captopril and Vilazodone; prostate cancer: Sirolimus) showed consistent associations at a nominal $P < 0.05$ (**Table S14**). Additionally, when meta-analyzing trial emulations at the ATC-L3 level, a narrower pharmacological or therapeutic subgroup, three

drugs (colorectal cancer: Captopril and Vilazodone; prostate cancer: Sirolimus) maintained consistent direction of association with cancer risk, as compared to ATC-L2, in both the ten-year and five-year follow-up periods (nominal $P < 0.05$) (**Table S16**).

Discussion

In this study, we conducted a comprehensive investigation of cancer risk proteins by integrating lead variants and pQTLs for six common cancer types using large-scale GWAS and population-based proteomics data. Through pQTL mapping and subsequent colocalization analysis, we identified 101 risk proteins across the six cancer types, with over three-quarters of them not previously linked to cancer susceptibility. Moreover, most of the proteins we identified are supported by functional genomics analyses. Our findings not only significantly expand the pool of known cancer risk proteins but also offer new insights into the biology and susceptibility of common cancers.

Through analysis of drug-protein interaction databases, we identified 36 druggable proteins potentially targeted by 404 therapeutic drugs. Among these, 30 drugs have already received approval for cancer treatment, while 73 are currently undergoing clinical trials for cancer treatment. These findings offer genetic evidence supporting the effectiveness of certain drugs and suggest potential opportunities for repurposing them to treat additional cancers that share common risk proteins. However, it's crucial to acknowledge that while the cancer risk proteins identified in our study hold promise as therapeutic targets for cancer treatment, drugs may also have adverse effects, potentially exacerbating cancer development through these targets (i.e., depending on their inhibitory or promotive effects)[78]. Additionally, our analysis characterized 197 drugs used for indications other than cancer, which may influence cancer risk due to their interactions with cancer-risk proteins. Overall, our findings have the potential to accelerate therapeutic drug discovery for the prevention and intervention of human cancers.

Our results showed that Acetazolamide exhibits a notable effect in preventing colorectal cancer development, through targeting the colorectal cancer risk protein, transferrin (TF). Previous literature suggested that TF protein may function similarly to an inhibitor of CA (e.g., ICA protein), as ICA belongs to the same TF protein superfamily[79; 80]. Thus, our genetic discovery provides strong evidence supporting acetazolamide as a promising drug for colorectal cancer prevention, potentially influencing a TF-regulated CA etiological pathway. In line with our findings, prior studies demonstrated Acetazolamide's

role in inhibiting cell viability, migration, and colony formation ability of colorectal cancer cells[81], as well as its ability to suppress the development of intestinal polyps in Min Mice[82]. Interestingly, we also identified another drug, Captopril, which potentially targets the cancer risk protein TF and is associated with an increased colorectal cancer risk. Captopril may influence cancer risk by inhibiting Angiotensin-converting enzyme (ACE), thereby affecting downstream TF protein levels through the regulation of its receptors[83; 84]. Consistent with our findings, a previous study also highlighted that ACE inhibition may play a risk-promoting role in colorectal cancer[85].

Our result indicated that Caffeine, a drug prioritized by the risk protein ALDH2, may have a protective role in colorectal cancer when compared to several control drugs. These findings align with previous studies suggesting a potential protective effect of Caffeine in colorectal cancer[86; 87]. Specifically, Caffeine demonstrated a significant protective role in colorectal cancer risk when compared to Paroxetine (P = 5.66 x 10^{-9}) and Magnesium Sulfate (P = 9.10 x 10^{-12}). This is supported by previous studies highlighting associations between ALDH2 variants and Caffeine consumption[88; 89] and another study demonstrating that Caffeine inhibits ALDH2 activity, resulting in reduced nucleophilicity and partial alterations in the enzyme's secondary structure[90]. Additionally, ALDH2 has been shown to enhance colorectal cancer stemness by activating β-catenin signaling[91; 92]. Conversely, our results also revealed that Caffeine exhibits a risk-promoting role in colorectal cancer when compared to Finasteride (P = 4.72 x 10^{-7}). This suggested that Finasteride may have a stronger protective effect against colorectal cancer risk than Caffeine, as previous studies have shown that Finasteride may play a role in the causal pathways of colorectal neoplasms[93] and demonstrated effectiveness in the treatment of prostate neoplasms[94-96].

Our meta-analysis results showed that Haloperidol, which targets the known cancer risk immune protein, HLA-A, is associated with a reduced breast cancer risk. This contrasts with previous studies suggesting that Haloperidol increases breast cancer risk[78; 97; 98]. The meta-analysis result appears to be primarily driven by the most significant individual trial (Haloperidol vs. Prochlorperazine, HR = 0.69, $P = 4.72 \times 10^{-2}$), which may overshadow other signals showing the risk-promoting role played by Haloperidol. Interestingly, our meta-analysis results showed that Haloperidol plays a risk-promoting role in colorectal cancer (HR = 1.25, CI = [1.01, 1.54], nominal P = 4.0 x 10^{-2}). Although these findings are in line with the observation that HLA-A is associated with decreased risk in breast cancer and increased risk in colorectal cancer, replication of these findings in an independent dataset is necessary for future studies.

We also uncovered an additional drug, Vilazodone, which is associated with an increased colorectal cancer risk by targeting AGER, a protein known to promote chronic inflammation by activating the NF-κB pathway[99]. Our study also found Sirolimus, a mechanistic target of rapamycin (mTOR) inhibitor, exhibits a risk-promoting role in prostate cancer, consistent with previous research[100; 101]. While our findings highlight potential etiological pathways involving these drugs, it is essential to evaluate the effects of the reported candidate drugs through both *in vitro* and *in vivo* assays in future research.

Data availability

Table S1 provides the download information for the summary statistics of GWAS data for the six common cancers, including breast, ovary, prostate, colorectum, lung, and pancreas.

Metadata and pQTL summary statistics from UKB-PPP can be downloaded from Synapse: Project SynID: syn51364943; pQTL from ARIC[46] and deCODE genetics[47] can be accessed through previous publications (PMID: 34857953 and PMID: 35501419). Functional genomic data includes: TCGA and CPTAC differential expression results accessible through https://ualcan.path.uab.edu/index.html; 4DGenome: https://4dgenome.research.chop.edu/; Depmap : https://depmap.org/portal/; FANTOM5: http://fantom.gsc.riken.jp/5/. HaploReg: https://pubs.broadinstitute.org/mammals/haploreg/.
GTEx: https://gtexportal.org/home/. GENCODE (v26.GRCh38) was downloaded from https://www.gencodegenes.org/human/release_26.html. National Cancer Institute can be accessed through https://www.cancer.gov/about-cancer/treatment/drugs; CGC can be accessed accessed via COSMIC website: https://cancer.sanger.ac.uk/census. Drugs and compounds data can be downloaded from the following URLs: ChEMBL: https://www.ebi.ac.uk/chembl/; Therapeutic Target Database: https://db.idrblab.net/ttd/; Open Targets: https://www.opentargets.org/; DrugBank: https://go.drugbank.com/. The EHR data, containing de-identified clinical information, can be accessed through the VUMC SD database. Data is available through restricted access for approved studies and researchers who agree to specific conditions of use.

Code availability

The developed pipeline and main source R codes that are used in this work are available from the GitHub website of Xingyi Guo's lab: https://github.com/XingyiGuo/PQTL_EHR/

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

This work was supported by the US National Institutes of Health grant 1R37CA227130-01A1 and R01CA269589-01A1 to X.G. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. New Frontiers in Research Fund (NFRFE-2018-00748) and NSERC Discovery Grant (RGPIN-2024-04679) to Q.L. The computational infrastructure was partly supported by a Canada Foundation for Innovation JELF grant (36605) to Q.L.

References

1. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nature genetics 47, 856-860.
2. Diogo, D., Tian, C., Franklin, C.S., Alanne-Kinnunen, M., March, M., Spencer, C.C.A., Vangjeli, C., Weale, M.E., Mattsson, H., Kilpelainen, E., et al. (2018). Phenome-wide association studies across large population cohorts support drug target validation. Nature communications 9, 4285.
3. Finan, C., Gaulton, A., Kruger, F.A., Lumbers, R.T., Shah, T., Engmann, J., Galver, L., Kelley, R., Karlsson, A., Santos, R., et al. (2017). The druggable genome and support for target identification and validation in drug development. Sci Transl Med 9.
4. Peters, U., and Tomlinson, I. (2024). Utilizing Human Genetics to Develop Chemoprevention for Cancer-Too Good an Opportunity to be Missed. Cancer Prev Res (Phila) 17, 7-12.
5. Jia, G., Ping, J., Shu, X., Yang, Y., Cai, Q., Kweon, S.S., Choi, J.Y., Kubo, M., Park, S.K., Bolla, M.K., et al. (2022). Genome- and transcriptome-wide association studies of 386,000 Asian and European-ancestry women provide new insights into breast cancer genetics. Am J Hum Genet 109, 2185-2195.
6. Zhang, H., Ahearn, T.U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T.A., Zhao, N., Bolla, M.K., et al. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature genetics 52, 572-581.
7. Fernandez-Rozadilla, C., Timofeeva, M., Chen, Z., Law, P., Thomas, M., Schmit, S., Díez-Obrero, V., Hsu, L., Fernandez-Tajes, J., Palles, C., et al. (2023). Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat Genet 55, 89-99.
8. Conti, D.V., Darst, B.F., Moss, L.C., Saunders, E.J., Sheng, X., Chou, A., Schumacher, F.R., Olama, A.A.A., Benlloch, S., Dadaev, T., et al. (2021). Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nat Genet 53, 65-75.
9. Byun, J., Han, Y., Li, Y., Xia, J., Long, E., Choi, J., Xiao, X., Zhu, M., Zhou, W., Sun, R., et al. (2022). Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. Nat Genet 54, 1167-1177.
10. Klein, A.P., Wolpin, B.M., Risch, H.A., Stolzenberg-Solomon, R.Z., Mocci, E., Zhang, M., Canzian, F., Childs, E.J., Hoskins, J.W., Jermusyk, A., et al. (2018). Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun 9, 556.
11. Phelan, C.M., Kuchenbaecker, K.B., Tyrer, J.P., Kar, S.P., Lawrenson, K., Winham, S.J., Dennis, J., Pirie, A., Riggan, M.J., Chornokur, G., et al. (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet 49, 680-691.
12. Lawrenson, K., Song, F., Hazelett, D.J., Kar, S.P., Tyrer, J., Phelan, C.M., Corona, R.I., Rodriguez-Malave, N.I., Seo, J.H., Adler, E., et al. (2019). Genome-wide association studies identify susceptibility loci for epithelial ovarian cancer in east Asian women. Gynecol Oncol 153, 343-355.
13. McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C., Caporaso, N.E., Johansson, M., Xiao, X., Li, Y., et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 49, 1126-1132.
14. Wen, W., Chen, Z., Bao, J., Long, Q., Shu, X.O., Zheng, W., and Guo, X. (2021). Genetic variations of DNA bindings of FOXA1 and co-factors in breast cancer susceptibility. Nat Commun 12, 5318.

607 15. Moreno, V., Alonso, M.H., Closa, A., Vallés, X., Diez-Villanueva, A., Valle, L., Castellví-Bel, S., Sanz-
608 Pamplona, R., Lopez-Doriga, A., Cordero, D., et al. (2018). Colon-specific eQTL analysis to
609 inform on functional SNPs. Br J Cancer 119, 971-977.
610 16. Chen, Z., Wen, W., Beeghly-Fadiel, A., Shu, X.O., Díez-Obrero, V., Long, J., Bao, J., Wang, J., Liu,
611 Q., Cai, Q., et al. (2019). Identifying Putative Susceptibility Genes and Evaluating Their
612 Associations with Somatic Mutations in Human Cancers. Am J Hum Genet 105, 477-492.
613 17. Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., Yuan, Y., Shi, J., Sun, Y., Han, M.R., et al. (2018). A
614 Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci
615 Identified in Genome-wide Association Studies. Am J Hum Genet 102, 890-903.
616 18. He, J., Wen, W., Beeghly, A., Chen, Z., Cao, C., Shu, X.O., Zheng, W., Long, Q., and Guo, X. (2022).
617 Integrating transcription factor occupancy with transcriptome-wide association analysis
618 identifies susceptibility genes in human cancers. Nat Commun 13, 7118.
619 19. Guo, X., Lin, W., Wen, W., Huyghe, J., Bien, S., Cai, Q., Harrison, T., Chen, Z., Qu, C., Bao, J., et al.
620 (2021). Identifying Novel Susceptibility Genes for Colorectal Cancer Risk From a
621 Transcriptome-Wide Association Study of 125,478 Subjects. Gastroenterology 160, 1164-
622 1178.e1166.
623 20. Yuan, Y., Bao, J., Chen, Z., Villanueva, A.D., Wen, W., Wang, F., Zhao, D., Fu, X., Cai, Q., Long, J., et
624 al. (2021). Multi-omics analysis to identify susceptibility genes for colorectal cancer. Hum
625 Mol Genet 30, 321-330.
626 21. Bien, S.A., Su, Y.R., Conti, D.V., Harrison, T.A., Qu, C., Guo, X., Lu, Y., Albanes, D., Auer, P.L.,
627 Banbury, B.L., et al. (2019). Genetic variant predictors of gene expression provide new
628 insight into risk of colorectal cancer. Hum Genet 138, 307-326.
629 22. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q.,
630 et al. (2018). A transcriptome-wide association study of 229,000 women identifies new
631 candidate susceptibility genes for breast cancer. Nat Genet 50, 968-978.
632 23. Gao, G., Fiorica, P.N., McClellan, J., Barbeira, A.N., Li, J.L., Olopade, O.I., Im, H.K., and Huo, D.
633 (2023). A joint transcriptome-wide association study across multiple tissues identifies
634 candidate breast cancer susceptibility genes. Am J Hum Genet 110, 950-962.
635 24. Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K.L., Kote-Jarai, Z., Eeles, R., Freedman,
636 M., Haiman, C., Pasaniuc, B., et al. (2018). Large-scale transcriptome-wide association study
637 identifies new prostate cancer risk regions. Nat Commun 9, 4079.
638 25. Liu, D., Zhu, J., Zhou, D., Nikas, E.G., Mitanis, N.T., Sun, Y., Wu, C., Mancuso, N., Cox, N.J., Wang,
639 L., et al. (2022). A transcriptome-wide association study identifies novel candidate
640 susceptibility genes for prostate cancer risk. Int J Cancer 150, 80-90.
641 26. Bosse, Y., Li, Z., Xia, J., Manem, V., Carreras-Torres, R., Gabriel, A., Gaudreault, N., Albanes, D.,
642 Aldrich, M.C., Andrew, A., et al. (2020). Transcriptome-wide association study reveals
643 candidate causal genes for lung cancer. Int J Cancer 146, 1862-1878.
644 27. Lu, Y., Beeghly-Fadiel, A., Wu, L., Guo, X., Li, B., Schildkraut, J.M., Im, H.K., Chen, Y.A., Permuth,
645 J.B., Reid, B.M., et al. (2018). A Transcriptome-Wide Association Study Among 97,898
646 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. Cancer
647 Res 78, 5419-5430.
648 28. Gusev, A., Lawrenson, K., Lin, X., Lyra, P.C., Jr., Kar, S., Vavra, K.C., Segato, F., Fonseca, M.A.S.,
649 Lee, J.M., Pejovic, T., et al. (2019). A transcriptome-wide association study of high-grade
650 serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat
651 Genet 51, 815-823.
652 29. Zhong, J., Jermusyk, A., Wu, L., Hoskins, J.W., Collins, I., Mocci, E., Zhang, M., Song, L., Chung,
653 C.C., Zhang, T., et al. (2020). A Transcriptome-Wide Association Study Identifies Novel
654 Candidate Susceptibility Genes for Pancreatic Cancer. J Natl Cancer Inst 112, 1003-1012.
655 30. Zheng, C.J., Han, L.Y., Yap, C.W., Ji, Z.L., Cao, Z.W., and Chen, Y.Z. (2006). Therapeutic targets:
656 Progress of their exploration and investigation of their characteristics (vol 58, pg 259, 2006).
657 Pharmacol Rev 58, 682-682.

31. Zhu, J., Shu, X., Guo, X., Liu, D., Bao, J., Milne, R.L., Giles, G.G., Wu, C., Du, M., White, E., et al. (2020). Associations between Genetically Predicted Blood Protein Biomarkers and Pancreatic Cancer Risk. Cancer Epidemiol Biomarkers Prev 29, 1501-1508.
32. Shu, X., Bao, J., Wu, L., Long, J., Shu, X.O., Guo, X., Yang, Y., Michailidou, K., Bolla, M.K., Wang, Q., et al. (2020). Evaluation of associations between genetically predicted circulating protein biomarkers and breast cancer risk. Int J Cancer 146, 2130-2138.
33. Wu, L., Shu, X., Bao, J., Guo, X., Kote-Jarai, Z., Haiman, C.A., Eeles, R.A., Zheng, W., and Practical, C.B.P.C.C.P.C. (2019). Analysis of Over 140,000 European Descendants Identifies Genetically Predicted Blood Protein Biomarkers Associated with Prostate Cancer Risk. Cancer Res 79, 4592-4598.
34. Gregga, I., Pharoah, P.D., Gayther, S.A., Manichaikul, A., Im, H.K., Kar, S.P., Schildkraut, J.M., and Wheeler, H.E. (2023). Predicted proteome association studies of breast, prostate, ovarian, and endometrial cancers implicate plasma protein regulation in cancer susceptibility. Cancer Epidemiol Biomarkers Prev.
35. Jia, G., Yang, Y., Ping, J., Xu, S., Liu, L., Guo, X., Tao, R., Long, J., and Zheng, W. (2023). Identification of target proteins for breast cancer genetic risk loci and blood risk biomarkers in a large study by integrating genomic and proteomic data. Int J Cancer 152, 2314-2320.
36. Considine, D.P.C., Jia, G., Shu, X., Schildkraut, J.M., Pharoah, P.D.P., Zheng, W., Kar, S.P., and Ovarian Cancer Association, C. (2021). Genetically predicted circulating protein biomarkers and ovarian cancer risk. Gynecol Oncol 160, 506-513.
37. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature.
38. Tautermann, C.S. (2020). Current and Future Challenges in Modern Drug Discovery. Methods Mol Biol 2114, 1-17.
39. Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. (2019). Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18, 41-58.
40. Zang, C., Zhang, H., Xu, J., Zhang, H., Fouladvand, S., Havaldar, S., Cheng, F., Chen, K., Chen, Y., Glicksberg, B.S., et al. (2023). High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. Nat Commun 14, 8180.
41. Wu, Y., Warner, J.L., Wang, L., Jiang, M., Xu, J., Chen, Q., Nian, H., Dai, Q., Du, X., Yang, P., et al. (2019). Discovery of Noncancer Drug Effects on Survival in Electronic Health Records of Patients With Cancer: A New Paradigm for Drug Repurposing. JCO Clin Cancer Inform 3, 1-9.
42. Xu, H., Aldrich, M.C., Chen, Q., Liu, H., Peterson, N.B., Dai, Q., Levy, M., Shah, A., Han, X., Ruan, X., et al. (2015). Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc 22, 179-191.
43. Bejan, C.A., Cahill, K.N., Staso, P.J., Choi, L., Peterson, J.F., and Phillips, E.J. (2021). DrugWAS: Drug-wide Association Studies for COVID-19 Drug Repurposing. Clin Pharmacol Ther 110, 1537-1546.
44. Reznikov, L.R., Norris, M.H., Vashisht, R., Bluhm, A.P., Li, D., Liao, Y.J., Brown, A., Butte, A.J., and Ostrov, D.A. (2021). Identification of antiviral antihistamines for COVID-19 repurposing. Biochem Biophys Res Commun 538, 173-179.
45. Liu, R., Wei, L., and Zhang, P. (2021). A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nat Mach Intell 3, 68-75.
46. Zhang, J., Dutta, D., Kottgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., Consortium, C.K., Yu, B., Boerwinkle, E., et al. (2022). Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat Genet 54, 593-602.

47. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat Genet 53, 1712-1721.
48. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S., Pooley, K.A., Dennis, J., Michailidou, K., Turman, C., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet 52, 56-73.
49. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Series B Stat Methodol 82, 1273-1300.
50. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., van de Geijn, B., Reshef, Y., Marquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat Genet 52, 1355-1363.
51. Chen, Z., Guo, X., Tao, R., Huyghe, J.R., Law, P.J., Fernandez-Rozadilla, C., Ping, J., Jia, G., Long, J., Li, C., et al. (2024). Fine-mapping analysis including over 254,000 East Asian and European descendants identifies 136 putative colorectal cancer susceptibility genes. Nat Commun 15, 3557.
52. Schwarzer, G., Carpenter, J.R., and Rücker, G. (2015). Meta-analysis with R.(Springer).
53. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 10, e1004383.
54. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48, 481-487.
55. Li, L., Huang, K.L., Gao, Y., Cui, Y., Wang, G., Elrod, N.D., Li, Y., Chen, Y.E., Ji, P., Peng, F., et al. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat Genet 53, 994-1005.
56. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40, D930-934.
57. He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A 111, E2191-2199.
58. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16, 22.
59. Gao, T., and Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res 48, D58-D64.
60. Wang, Y., Song, C., Zhao, J., Zhang, Y., Zhao, X., Feng, C., Zhang, G., Zhu, J., Wang, F., Qian, F., et al. (2023). SEdb 2.0: a comprehensive super-enhancer database of human and mouse. Nucleic Acids Res 51, D280-D290.
61. Chandrashekar, D.S., Bashel, B., Balasubramanya, S.A.H., Creighton, C.J., Ponce-Rodriguez, I., Chakravarthi, B., and Varambally, S. (2017). UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia 19, 649-658.
62. Chandrashekar, D.S., Karthikeyan, S.K., Korla, P.K., Patel, H., Shovon, A.R., Athar, M., Netto, G.J., Qin, Z.S., Kumar, S., Manne, U., et al. (2022). UALCAN: An update to the integrated cancer data analysis platform. Neoplasia 25, 18-27.
63. Knox, C., Wilson, M., Klinger, C.M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E.L., Strawbridge, S.A., et al. (2024). DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res 52, D1265-D1275.
64. Zdrazil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D.M., Mosquera, J.F., et al. (2024). The ChEMBL Database in 2023: a drug discovery

758 platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52,
759 D1180-D1192.
760 65. Zhou, Y., Zhang, Y., Zhao, D., Yu, X., Shen, X., Zhou, Y., Wang, S., Qiu, Y., Chen, Y., and Zhu, F.
761 (2024). TTD: Therapeutic Target Database describing target druggability information. Nucleic
762 Acids Res 52, D1465-D1477.
763 66. Buniello, A., Suveges, D., Cruz-Castillo, C., Llinares, M.B., Cornu, H., Lopez, I., Tsukanov, K.,
764 Roldan-Romero, J.M., Mehta, C., Fumis, L., et al. (2025). Open Targets Platform: facilitating
765 therapeutic hypotheses building in drug discovery. Nucleic Acids Res 53, D1467-D1475.
766 67. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys,
767 D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized
768 medicine. Clin Pharmacol Ther 84, 362-369.
769 68. Austin, P.C., and Stuart, E.A. (2015). Moving towards best practice when using inverse probability
770 of treatment weighting (IPTW) using the propensity score to estimate causal treatment
771 effects in observational studies. Stat Med 34, 3661-3679.
772 69. Shu, X., Zhou, Q., Sun, X., Flesaker, M., Guo, X., Long, J., Robson, M.E., Shu, X.O., Zheng, W., and
773 Bernstein, J.L. (2022). Associations between circulating proteins and risk of breast cancer by
774 intrinsic subtypes: a Mendelian randomisation analysis. Br J Cancer 127, 1507-1514.
775 70. Sun, J., Zhao, J.H., Jiang, F.Y., Wang, L.J., Xiao, Q., Han, F.Y., Chen, J., Yuan, S., Wei, J.S., Larsson,
776 S.C., et al. (2023). Identification of novel protein biomarkers and drug targets for colorectal
777 cancer by integrating human plasma proteome with genome. Genome Med 15.
778 71. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico,
779 A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of
780 Cancer Driver Genes and Mutations. Cell 174, 1034-1035.
781 72. Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., Lander, E.S., Van
782 Allen, E.M., and Sunyaev, S.R. (2020). Identification of cancer driver genes based on
783 nucleotide context. Nature genetics 52, 208-218.
784 73. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC
785 Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev
786 Cancer 18, 696-705.
787 74. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson,
788 M., Neveu, V., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic
789 Acids Res 42, D1091-1097.
790 75. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S.,
791 Michalovich, D., Al-Lazikani, B., et al. (2012). ChEMBL: a large-scale bioactivity database for
792 drug discovery. Nucleic Acids Res 40, D1100-1107.
793 76. Zhou, Y., Zhang, Y.T., Zhao, D.H., Yu, X.Y., Shen, X.Y., Zhou, Y., Wang, S.S., Qiu, Y.Q., Chen, Y.Z.,
794 and Zhu, F. (2023). TTD: Therapeutic Target Database describing target druggability
795 information. Nucleic Acids Res.
796 77. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda,
797 A., Cruz-Castillo, C., Fumis, L., et al. (2023). The next-generation Open Targets Platform:
798 reimagined, redesigned, rebuilt. Nucleic Acids Res 51, D1353-D1359.
799 78. Taipale, H., Solmi, M., Lahteenvuo, M., Tanskanen, A., Correll, C.U., and Tiihonen, J. (2021).
800 Antipsychotic use and risk of breast cancer in women with schizophrenia: a nationwide
801 nested case-control study in Finland. Lancet Psychiatry 8, 883-891.
802 79. Durdagi, S., Vullo, D., Pan, P., Kahkonen, N., Maatta, J.A., Hytonen, V.P., Scozzafava, A., Parkkila,
803 S., and Supuran, C.T. (2012). Protein-protein interactions: inhibition of mammalian carbonic
804 anhydrases I-XV by the murine inhibitor of carbonic anhydrase and other members of the
805 transferrin family. J Med Chem 55, 5529-5535.
806 80. Eckenroth, B.E., Mason, A.B., McDevitt, M.E., Lambert, L.A., and Everse, S.J. (2010). The structure
807 and evolution of the murine inhibitor of carbonic anhydrase: a member of the transferrin
808 superfamily. Protein Sci 19, 1616-1626.

81. Karakus, F., Eyol, E., Yilmaz, K., and Unuvar, S. (2018). Inhibition of cell proliferation, migration and colony formation of LS174T Cells by carbonic anhydrase inhibitor. Afr Health Sci 18, 1303-1310.
82. Noma, N., Fujii, G., Miyamoto, S., Komiya, M., Nakanishi, R., Shimura, M., Tanuma, S.I., and Mutoh, M. (2017). Impact of Acetazolamide, a Carbonic Anhydrase Inhibitor, on the Development of Intestinal Polyps in Min Mice. Int J Mol Sci 18.
83. Shen, Y., Li, X., Dong, D., Zhang, B., Xue, Y., and Shang, P. (2018). Transferrin receptor 1 in cancer: a new sight for cancer therapy. Am J Cancer Res 8, 916-931.
84. Chen, Y.J., Qian, Z.M., Sheng, Y., Zheng, J., and Liu, Y. (2020). Angiotensin II down-regulates transferrin receptor 1 and ferroportin 1 expression in Neuro-2a cells via activation of type-1 receptor. Neurosci Lett 716, 134684.
85. Yarmolinsky, J., Diez-Obrero, V., Richardson, T.G., Pigeyre, M., Sjaarda, J., Pare, G., Walker, V.M., Vincent, E.E., Tan, V.Y., Obon-Santacana, M., et al. (2022). Genetically proxied therapeutic inhibition of antihypertensive drug targets and risk of common cancers: A mendelian randomization analysis. PLoS Med 19, e1003897.
86. Schmit, S.L., Rennert, H.S., Rennert, G., Gruber, S.B.J.C.E., Biomarkers, and Prevention. (2016). Coffee consumption and the risk of colorectal cancer. 25, 634-639.
87. Sartini, M., Bragazzi, N.L., Spagnolo, A.M., Schinca, E., Ottria, G., Dupont, C., and Cristina, M.L. (2019). Coffee Consumption and Risk of Colorectal Cancer: A Systematic Review and Meta-Analysis of Prospective Studies. Nutrients 11.
88. Low, J.J., Tan, B.J., Yi, L.X., Zhou, Z.D., and Tan, E.K. (2024). Genetic susceptibility to caffeine intake and metabolism: a systematic review. J Transl Med 22, 961.
89. Chen, C.H., Kraemer, B.R., and Mochly-Rosen, D. (2022). ALDH2 variance in disease and populations. Dis Model Mech 15.
90. Laskar, A.A., Alam, M.F., Ahmad, M., and Younus, H. (2018). Kinetic and biophysical investigation of the inhibitory effect of caffeine on human salivary aldehyde dehydrogenase: Implications in oral health and chemotherapy. J Journal of Molecular Structure 1157, 61-68.
91. Wei, P.L., Prince, G., Batzorig, U., Huang, C.Y., and Chang, Y.J. (2023). ALDH2 promotes cancer stemness and metastasis in colorectal cancer through activating beta-catenin signaling. J Cell Biochem 124, 907-920.
92. Li, R., Zhao, Z., Sun, M., Luo, J., and Xiao, Y. (2016). ALDH2 gene polymorphism in different types of cancers and its clinical significance. Life Sci 147, 59-66.
93. Keum, N., Cao, Y., Lee, D.H., Park, S.M., Rosner, B., Fuchs, C.S., Wu, K., and Giovannucci, E.L. (2016). Male pattern baldness and risk of colorectal neoplasia. Br J Cancer 114, 110-117.
94. Thompson, I.M., Jr., Goodman, P.J., Tangen, C.M., Parnes, H.L., Minasian, L.M., Godley, P.A., Lucia, M.S., and Ford, L.G. (2013). Long-term survival of participants in the prostate cancer prevention trial. N Engl J Med 369, 603-610.
95. Wang, L., Lei, Y., Gao, Y., Cui, D., Tang, Q., Li, R., Wang, D., Chen, Y., Zhang, B., and Wang, H. (2020). Association of finasteride with prostate cancer: A systematic review and meta-analysis. Medicine (Baltimore) 99, e19486.
96. Thompson, I.M., Lucia, M.S., Redman, M.W., Darke, A., La Rosa, F.G., Parnes, H.L., Lippman, S.M., and Coltman, C.A. (2007). Finasteride decreases the risk of prostatic intraepithelial neoplasia. J Urol 178, 107-109; discussion 110.
97. Rahman, T., Sahrmann, J.M., Olsen, M.A., Nickel, K.B., Miller, J.P., Ma, C., and Grucza, R.A. (2022). Risk of Breast Cancer With Prolactin Elevating Antipsychotic Drugs: An Observational Study of US Women (Ages 18-64 Years). J Clin Psychopharmacol 42, 7-16.
98. Hicks, B.M., Busby, J., Mills, K., O'Neil, F.A., McIntosh, S.A., Zhang, S.D., Liberante, F.G., and Cardwell, C.R. (2020). Post-diagnostic antipsychotic use and cancer mortality: a population based cohort study. Bmc Cancer 20, 804.

99. Zhou, M., Zhang, Y., Shi, L., Li, L., Zhang, D., Gong, Z., and Wu, Q. (2024). Activation and modulation of the AGEs-RAGE axis: Implications for inflammatory pathologies and therapeutic interventions - A review. Pharmacol Res 206, 107282.
100. Chamie, K., Ghosh, P.M., Koppie, T.M., Romero, V., Troppmann, C., and deVere White, R.W. (2008). The effect of sirolimus on prostate-specific antigen (PSA) levels in male renal transplant recipients without prostate cancer. Am J Transplant 8, 2668-2673.
101. Yanik, E.L., Siddiqui, K., and Engels, E.A. (2015). Sirolimus effects on cancer incidence after kidney transplantation: a meta-analysis. Cancer Med 4, 1448-1459.

Figures

Figure 1: Overview of the Analytical Framework.

A, An illustration depicting the identification of proteins associated with the risk of the six major cancers: breast, lung, colorectal, ovarian, pancreatic, and prostate. Population-based proteomics data (for pQTLs) and GWAS data resources (for identifying lead variants) utilized in this study are shown in the left panels. Meta-analyses of *cis*-pQTLs from ARIC and deCODE, conducted through the SOMAscan® platform, were combined with pQTL results from the UKB-PPP to identify potential risk proteins, as depicted in the middle panels. Colocalization analyses between GWAS summary statistics and *cis*-pQTLs were performed to identify cancer risk proteins with high confidence, as illustrated in the right panel. **B**, The proteins with evidence of colocalization annotated based on drug-protein information from four databases: DrugBank, ChEMBL, TTD, and OpenTargets. **C**, The framework for evaluating the effects of drugs approved for indications on cancer risk. The Inverse Probability of Treatment weighting (IPTW) framework was utilized to construct emulations of treated-control drug trials based on millions of patients' Electronic Health Records stored at VUMC SD (left pane). In these emulations, the Cox proportional hazard model was conducted for each trial to assess the hazard ratio (HR) of cancer risk between the treated focal drug and the control drug (right panels).

Figure 2: Genome-wide distribution of lead variants and putative risk proteins among six types of cancer. Proteins identified for each cancer are represented by different colors. Each circle represents a single lead variant-protein pair. Proteins marked with an asterisk (*) denote multiple proteins associated with the lead variants. A dashed box highlights several well-known cancer-related proteins, such as HLA-A and HLA-E, which are linked to lead variants located in the major histocompatibility complex (MHC).

Figure 3: Identification of 101 cancer risk proteins through pQTL and colocalization analysis

A, Number of proteins showing evidence of colocalizations between pQTLs and GWAS association signals for six cancer types**. B**, Percentage of proteins showing evidence of colocalizations between pQTLs and GWAS summary statistics for six cancer types. **C,** A plot illustrating the high consistency of pQTL p-values for 22 cancer risk proteins between the ARIC+deCODE and the UKB-PPP (proteins commonly assays from SOMAscan® and Olink platforms).

Figure 4: A circular plot showing 36 druggable proteins potentially targeted by 404 approved drugs or undergoing clinical trials for cancer treatment or other indications

Presented from inner to outer layers are cancer types, proteins, and drugs. Each drug-protein interaction is annotated by DrugBank, ChEMBL, TTD, and OpenTargets, with lines in different colors representing each database. Interactions where proteins are annotated by two databases are linked to drugs with thick lines.

Figure 5: A circular plot showing 19 druggable proteins potentially targeted by 133 approved drugs or undergoing clinical trials for cancer treatment. Presented from inner to outer layers are cancer types, proteins, drugs and cancers. Drugs approved and undergoing clinical trials for cancer treatment are highlighted on green and gray, respectively. Drug approved indications are formatted in bold, while indications under clinical trial are in regular font.

Figure 6: Drugs approved for treated indications showing significant effects on cancer risk

A, A table showing cancer risk alleles of lead variants, risk proteins, and drug name approved for indications. Positive associations are indicated by upward arrows, while negative associations are indicated by downward arrows. Odds ratio (OR) refers to the exponential transformation of logistic regression coefficient and beta refers to linear regression coefficient. **B**, An illustration of six drugs, showing significant associations with cancer risk in meta-analysis, and their potential targeted risk proteins. **C**, Survival plots depict the statistically significant difference in the probability of being cancer-free for patients in the treated group (taking a focal drug, shown in green) and those in the control group (primarily selected by the smallest *P* value across eligible treated-control trials for each significant drug in the meta-analyses, shown in purple; all detailed results see Table S13). The shaded area represents the 95% confidence interval. The hazard ratio and *P* for the focal drug compared to the control drug, determined through weighted Cox proportional hazards models, are presented in the top right corner of each panel.

Supplementary Figures

Figure S1: A flowchart for characterizing lead variants with independent risk signals in six types of cancer. The analysis for each of the six major cancers: breast, lung, colorectal, ovarian, pancreatic, and prostate is separated by dashed lines. The detailed protocols of new efforts from our additional fine-mapping analysis using SuSiE and a collection of previously identified risk variants from GWAS or fine-mapping studies are indicated in Box A, Box B and Box C, respectively.

Figure S2: Common cancer risk proteins identified across six cancer types. A, A Venn plot showing common proteins in breast, colorectal, lung and prostate cancer. **B**, A heatmap showing common proteins observed from at least two of six types of cancer.

Figure S3: A circular plot showing 28 druggable proteins potentially targeted by 197 approved drugs or undergoing clinical trials for treated indications rather than cancers. Presented from inner to outer layers are cancer types, proteins, drugs and cancers. Drugs approved and undergoing clinical trials for cancer treatment are highlighted on blue and gray, respectively. Drug approved indications are formatted in bold, while indications under clinical trial are in regular font.

Supplementary Tables

Supplementary Table 1: GWAS summary statistics data of European descendants for breast, ovarian, colorectal, pancreatic, prostate and lung cancers used in this study.

Supplementary Table 2: List 710 lead variants with independent risk signals across six types of cancers

Supplementary Table 3: List of identified 365 putative cancer risk proteins across six types of cancers

Supplementary Table 4: List of novel and reported putative cancer risk proteins in breast, colorectal, ovarian, and prostate cancers

Supplementary Table 5: List of 101 cancer risk proteins identified by GWAS and pQTLs colocalization for six types of cancers

Supplementary Table 6: The identified cancer risk proteins supported by the evidence of eQTLs, sQTLs and apaQTLs

Supplementary Table 7: List of the identified cancer risk proteins potentially regulated by putative proximal regulatory variants.

Supplementary Table 8: List of the identified cancer risk proteins potentially regulated by putative distal regulatory variants.

Supplementary Table 9: Cancer risk proteins supported by the evidence of functional genomic analysis

Supplementary Table 10: List of 36 cancer druggable risk proteins

Supplementary Table 11: List of cancer drugs and non-cancer drugs (for other indications) that potentially target cancer risk proteins

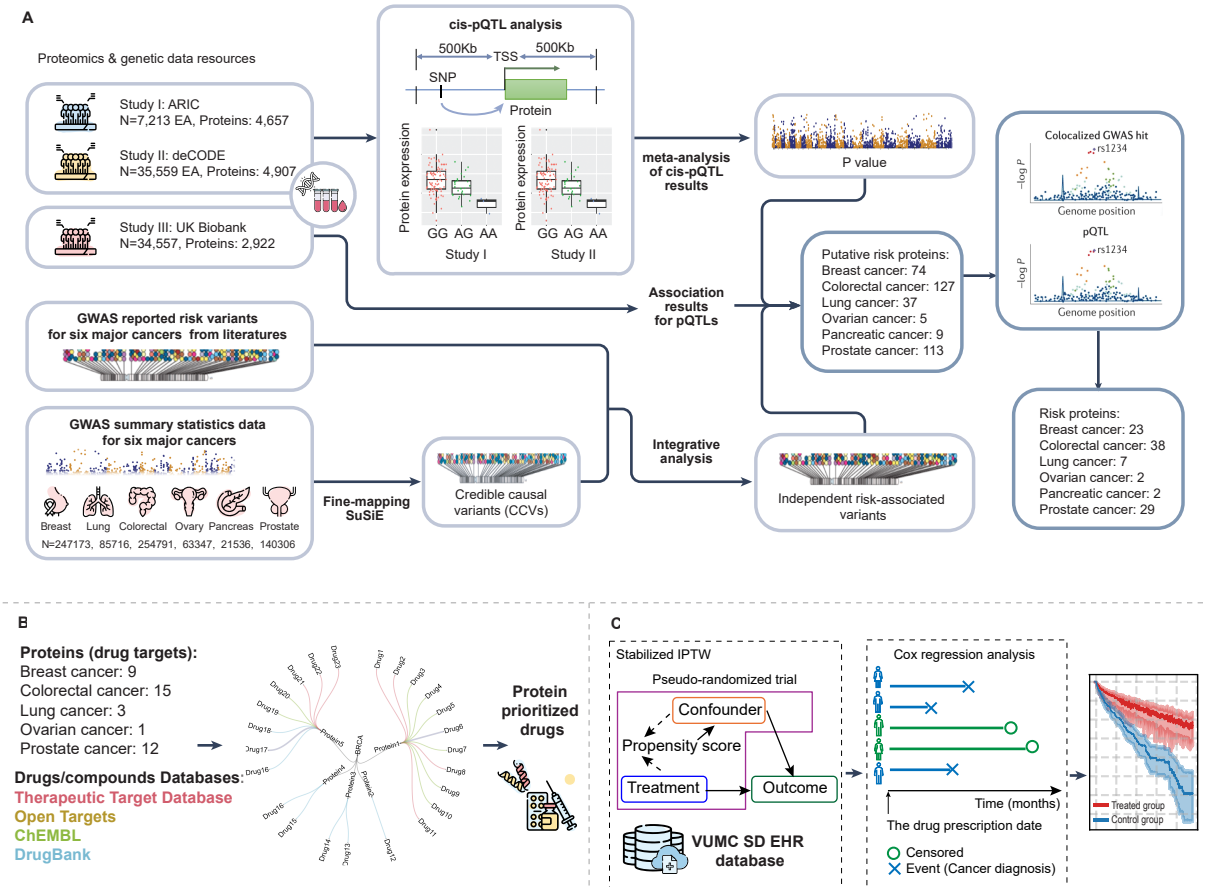
Supplementary Table 12: List of covariates utilized in LR-PS models

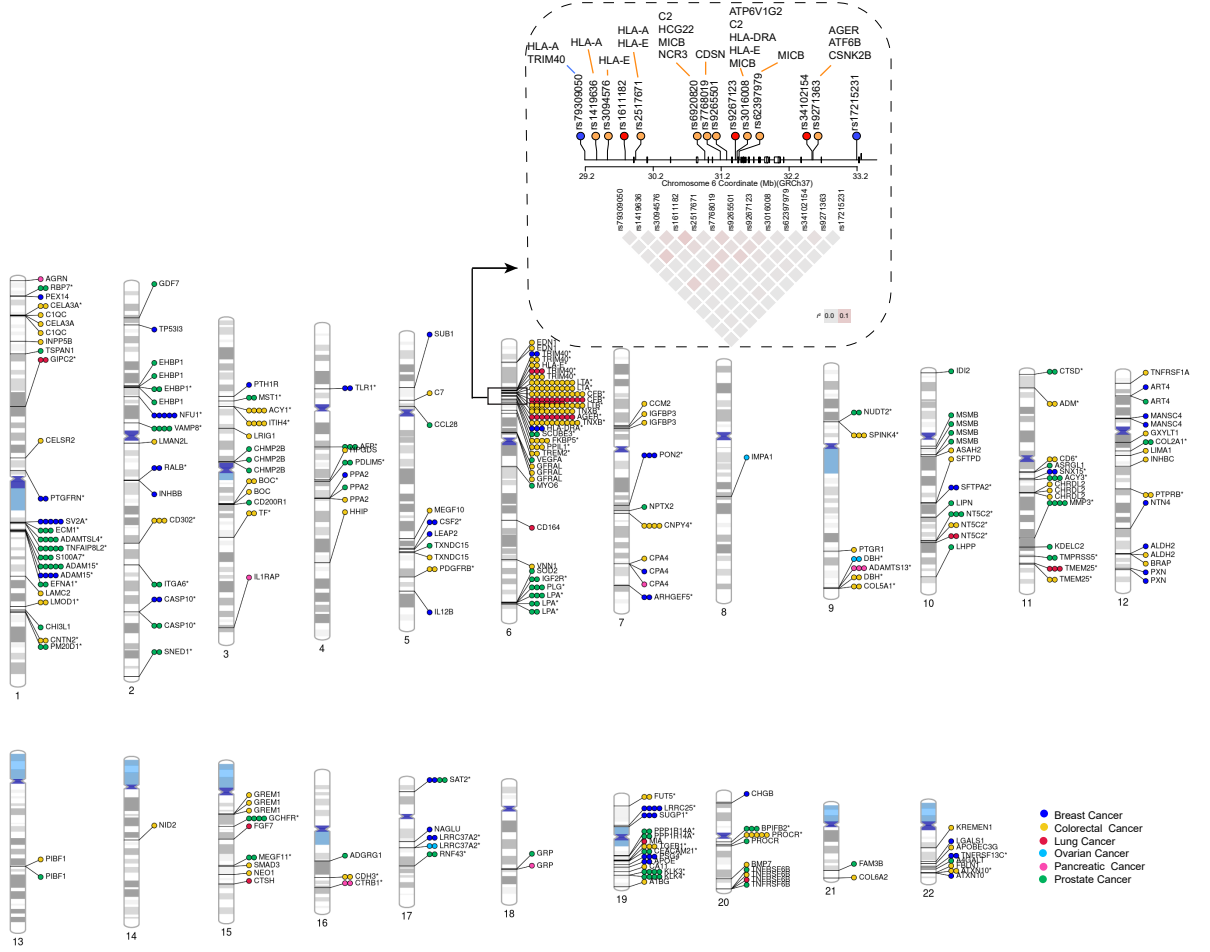
Supplementary Table 13: Individual Cox proportional hazards analysis results for treated and control drugs (10 years follow-up period)

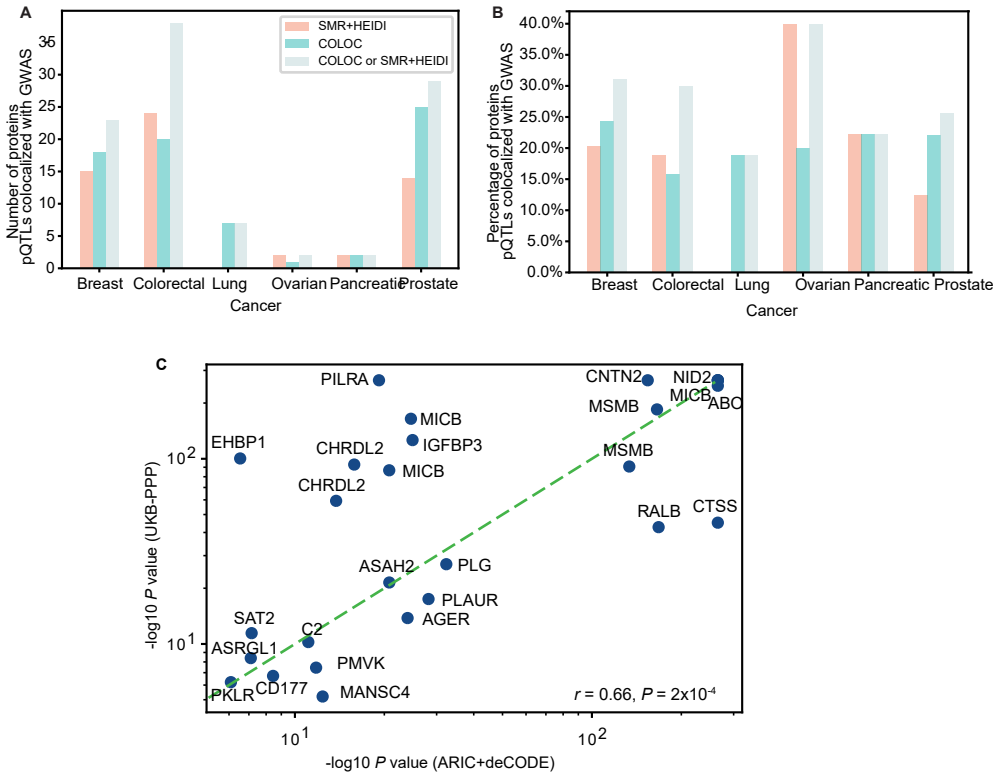
Supplementary Table 14: Meta-analysis of Cox proportional hazards for treated vs. control drugs by level 2 ATC codes

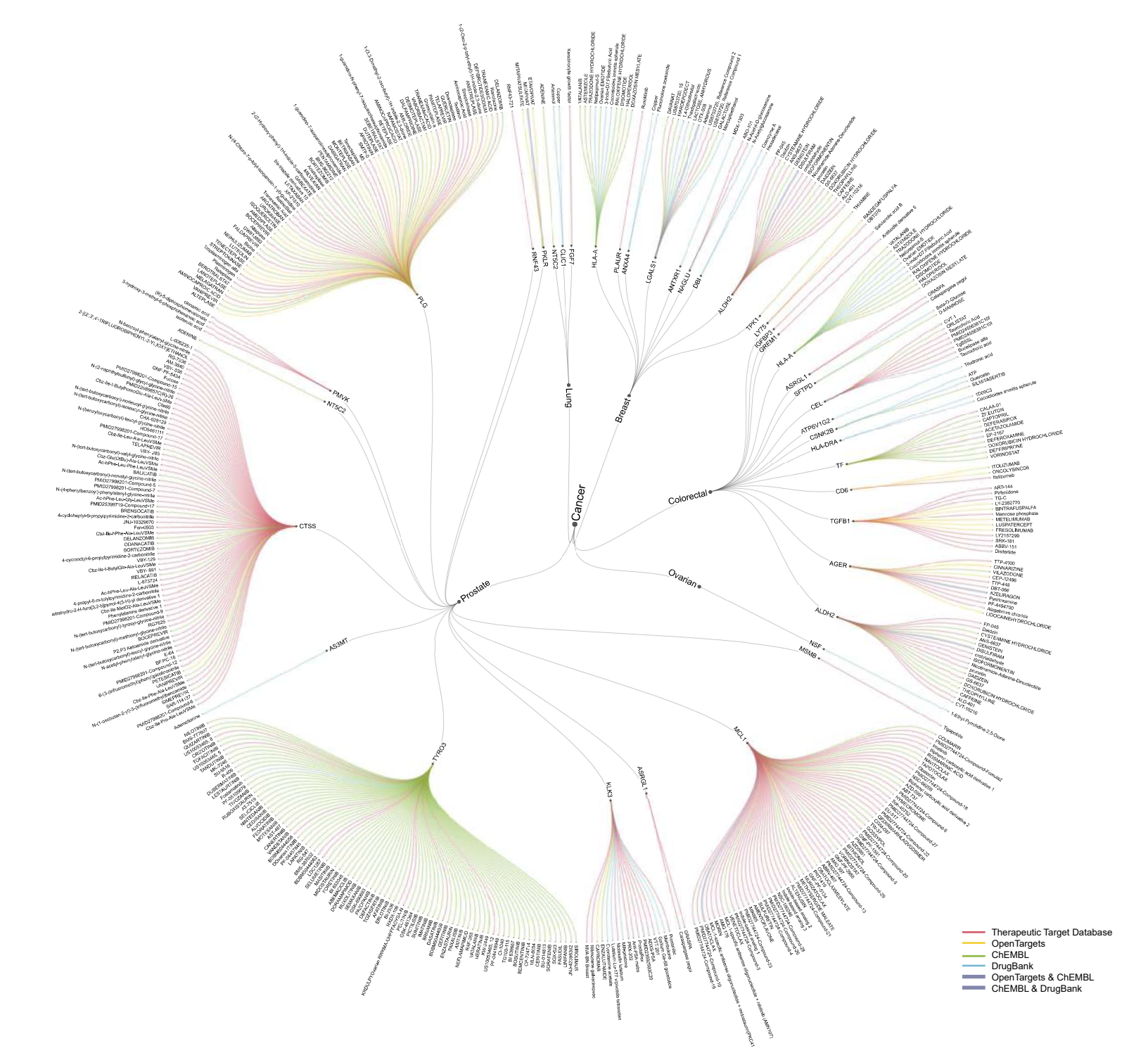
Supplementary Table 15: Individual Cox proportional hazards analysis results for treated and control drugs (5 years follow-up period)

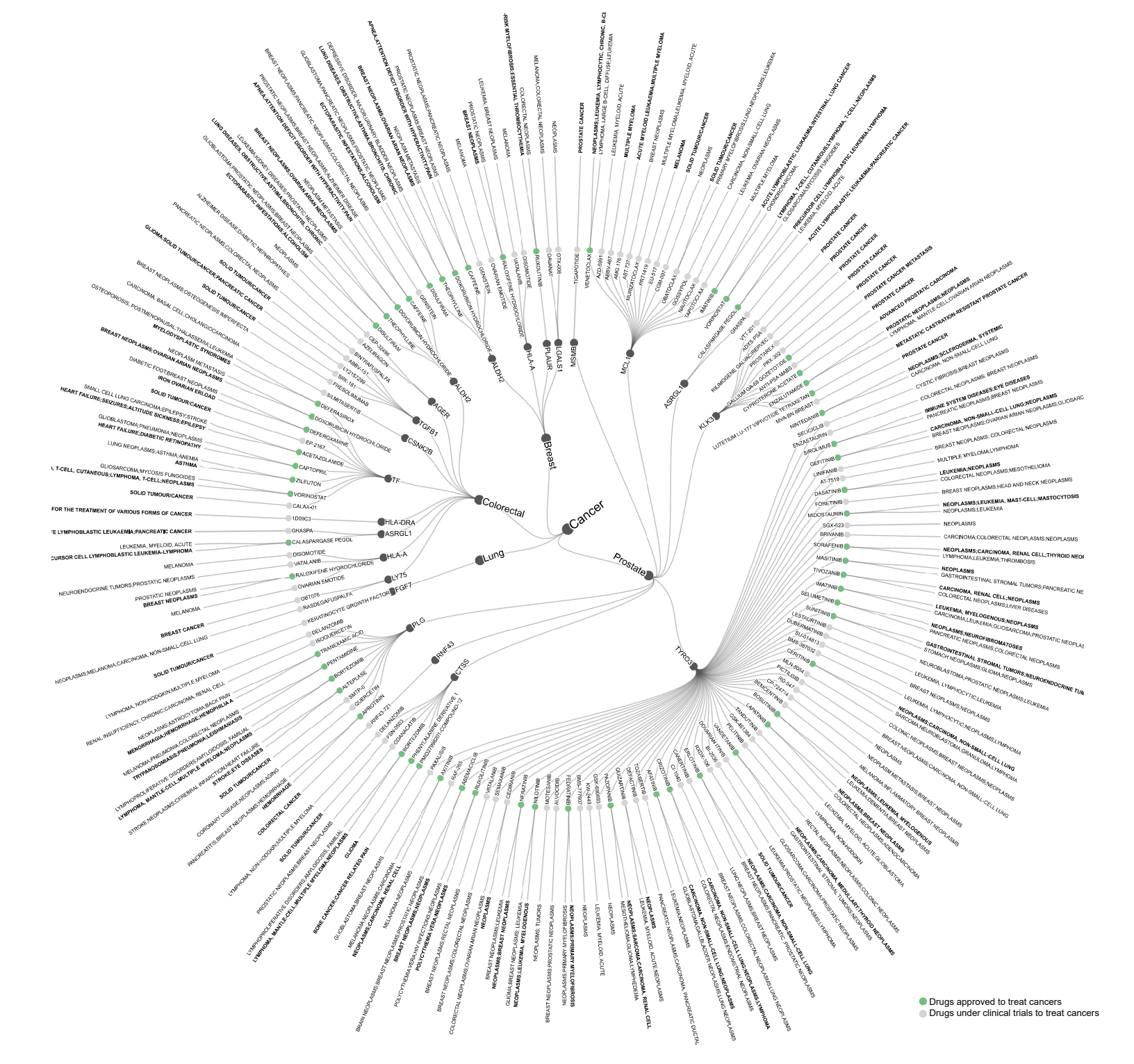
Supplementary Table 16: Meta-analysis of Cox proportional hazards for treated vs. control drugs by level 3 ATC codes


A
Proteomics & genetic data resources
Study I: ARIC
N=7,213 EA, Proteins: 4,657
Study II: deCODE
N=35,559 EA, Proteins: 4,907
Study III: UK Biobank
N=34,557, Proteins: 2,922
cis-pQTL analysis
500Kb
TSS
500Kb
SNP
Protein
Protein expression
GG AG AA
Study I
Study II
meta-analysis of cis-pQTL results
P value
Colocalized GWAS hit
rs1234
-log P
Genome position
pQTL
Putative risk proteins:
Breast cancer: 74
Colorectal cancer: 127
Lung cancer: 37
Ovarian cancer: 5
Pancreatic cancer: 9
Prostate cancer: 113
Association results for pQTLs
GWAS reported risk variants for six major cancers from literatures
GWAS summary statistics data for six major cancers
Breast Lung Colorectal Ovary Pancreas Prostate
N=247173, 85716, 254791, 63347, 21536, 140306
Fine-mapping SuSiE
Credible causal variants (CCVs)
Integrative analysis
Independent risk-associated variants
Risk proteins:
Breast cancer: 23
Colorectal cancer: 38
Lung cancer: 7
Ovarian cancer: 2
Pancreatic cancer: 2
Prostate cancer: 29
B
Proteins (drug targets):
Breast cancer: 9
Colorectal cancer: 15
Lung cancer: 3
Ovarian cancer: 1
Prostate cancer: 12
Drugs/compounds Databases:
Therapeutic Target Database
Open Targets
ChEMBL
DrugBank
Protein prioritized drugs
C
Stabilized IPTW
Pseudo-randomized trial
Confounder
Propensity score
Treatment
Outcome
VUMC SD EHR database
Cox regression analysis
Time (months)
The drug prescription date
Censored
Event (Cancer diagnosis)
Treated group
Control group


HLA-A
TRIM40
HLA-A
HLA-A
HLA-E
HLA-E
C2
HCG22
MICB
NCR3
CDSN
ATP6V1G2
C2
HLA-DRA
HLA-E
MICB
MICB
AGER
ATF6B
CSNK2B
rs7930050
rs1419636
rs3094576
rs1611182
rs2517671
rs6920820
rs7768019
rs9265501
rs9267123
rs3016008
rs6023979
rs34102154
rs9271363
rs17215231
29.2
30.2
31.2
32.2
33.2
Chromosome 6 Coordinate (Mb)(GRCh37)
r² 0.0 0.1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Breast Cancer
Colorectal Cancer
Lung Cancer
Ovarian Cancer
Pancreatic Cancer
Prostate Cancer


A
SMR+HEIDI
COLOC
COLOC or SMR+HEIDI
Number of proteins pQTLs colocalized with GWAS
0
5
10
15
20
25
30
35
Breast
Colorectal
Lung
Ovarian
Pancreatic
Prostate
Cancer
B
Percentage of proteins pQTLs colocalized with GWAS
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
30.0%
35.0%
40.0%
Breast
Colorectal
Lung
Ovarian
Pancreatic
Prostate
Cancer
C
PILRA
CNTN2
NID2
MICB
ABC
MSMB
MICB
IGFBP3
EHBP1
CHRDL2
MICB
MSMB
CHRDL2
RALB
CTSS
ASAH2
PLG
PLAUR
AGER
SAT2
C2
ASRGL1
PMVK
PKLR
CD177
MANSC4
r = 0.66, P = 2x10-4
-log10 P value (UKB-PPP)
10 2
10 1
10 1
10 2
-log10 P value (ARIC+deCODE)

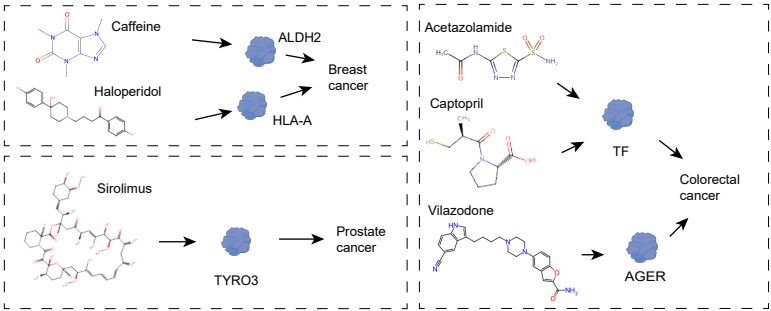

Cancer
Prostate
Lung
Breast
Colorectal
Ovarian
PLG
PMVK
NT5C2
CTSS
AS3MT
TYRO3
KLK3
ASRGL1
MCL1
NSF
MSMS
ALDH2
AGER
TGFB1
CD6
TF
HLA-DRA
CSNK2B
ATP6V1G2
CEL
SFTPD
ASRGL1
HLA-A
TPK1
LY75
IGFBP3
GREM1
ALDH2
LGALS1
ANTXR1
NAGLU
DBI
PLAUR
ANXA4
HLA-A
FGF7
CLIC1
NT5C2
PKLR
RNF43
Therapeutic Target Database
OpenTargets
ChEMBL
DrugBank
OpenTargets & ChEMBL
ChEMBL & DrugBank


Cancer
Colorectal
Lung
Breast
Prostate
Drugs approved to treat cancers
Drugs under clinical trials to treat cancers

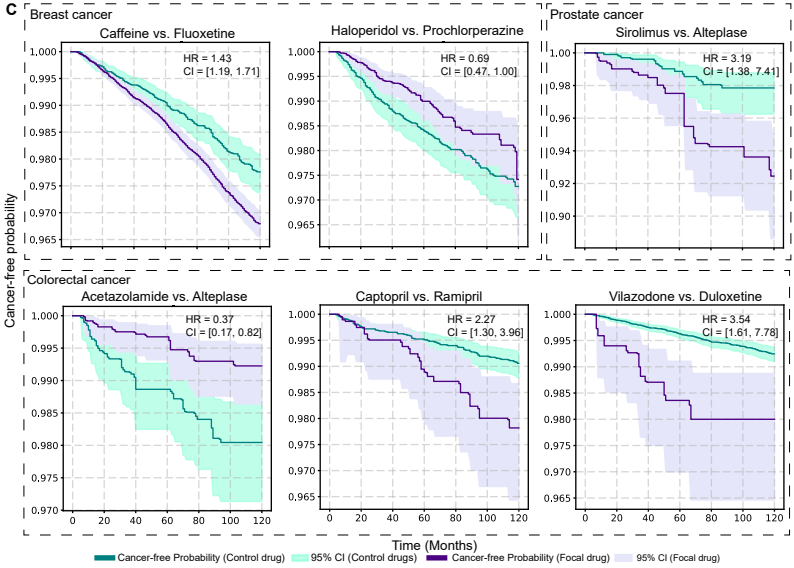
A

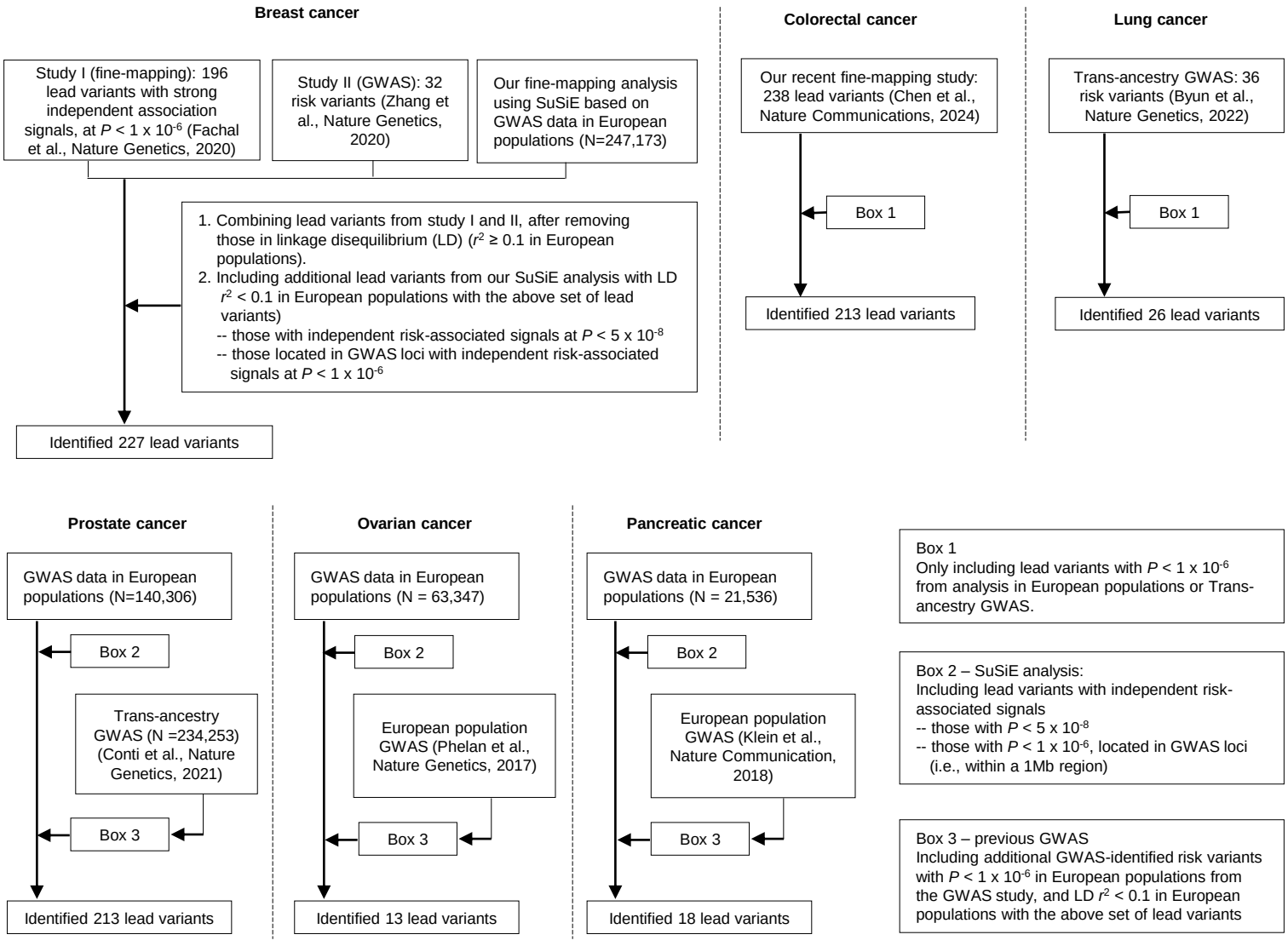
Cancer	Proteins	Lead variant	Risk allele	GWAS: OR (95% CI)		pQTL: beta (95% CI)		Drugs
Breast	ALDH2	rs11065822	G	1.04 (1.02, 1.05)	↑	-0.05 (-0.07, -0.04)	↓	Caffeine
Breast	HLA-A	rs79309050	T	1.13 (1.08, 1.18)	↑	-0.47 (-0.53, -0.41)	↓	Haloperidol, Trazodone Hydrochloride
Colorectal	AGER	rs9271363	C	1.09 (1.07, 1.24)	↑	-0.08 (-0.09, -0.06)	↓	Vilazodone
Colorectal	ALDH2	rs3858704	G	1.04 (1.02, 1.06)	↑	-0.04 (-0.05, -0.02)	↓	Caffeine, Theophylline
Colorectal	HLA-A	rs1419636	G	1.05 (1.03, 1.07)	↑	0.42 (0.41, 0.44)	↑	Theophylline, Trazodone Hydrochloride
Colorectal	HLA-A	rs2517671	G	1.04 (1.02, 1.05)	↑	0.83 (0.81, 0.84)	↑	Theophylline, Trazodone Hydrochloride
Colorectal	TF	rs4854776	C	1.04 (1.02, 1.06)	↑	-0.04 (-0.06, -0.03)	↓	Acetazolamide, Captopril
Prostate	TYRO3	rs11561564	G	1.06 (1.04, 1.08)	↑	0.14 (0.12, 0.16)	↑	Sirolimus

B

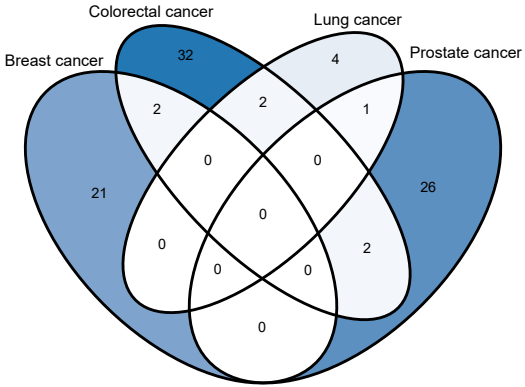


C

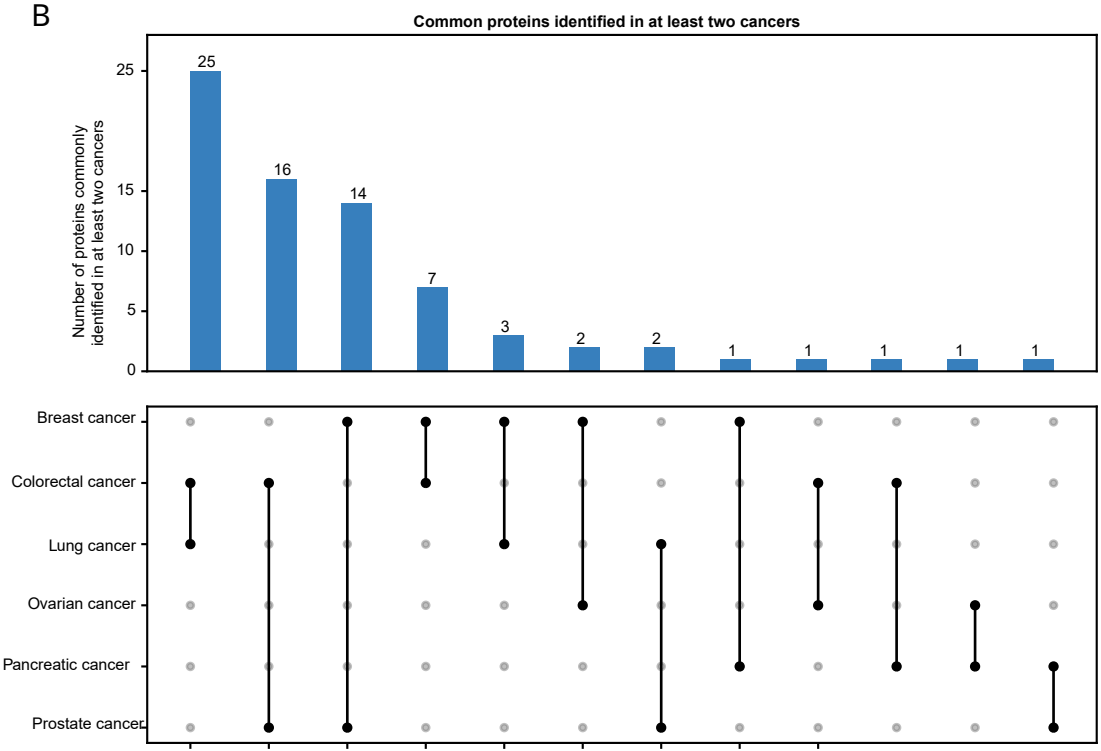


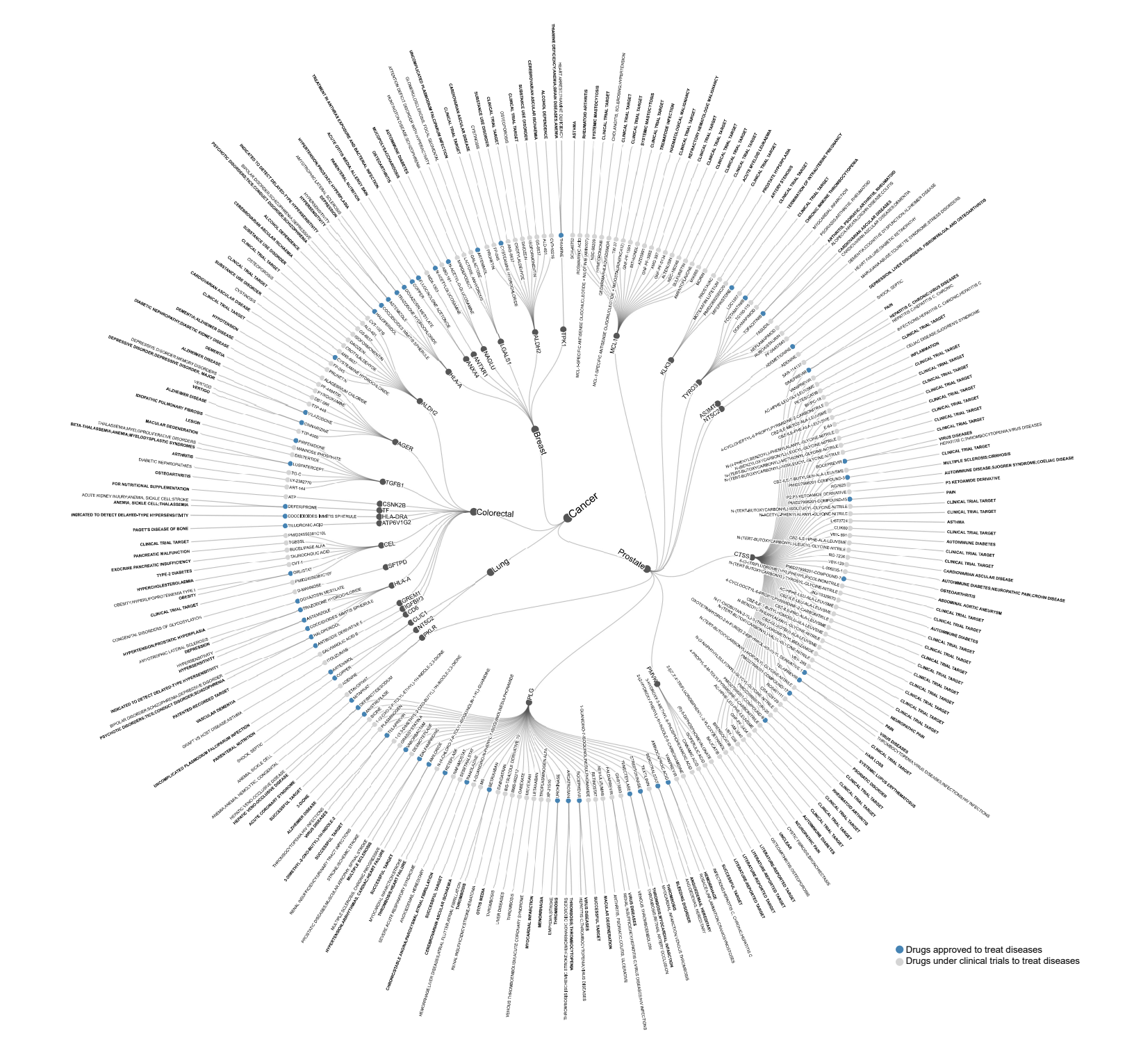

Breast cancer
Study I (fine-mapping): 196 lead variants with strong independent association signals, at $P < 1 \times 10^{-6}$ (Fachal et al., Nature Genetics, 2020)
Study II (GWAS): 32 risk variants (Zhang et al., Nature Genetics, 2020)
Our fine-mapping analysis using SuSiE based on GWAS data in European populations (N=247,173)
1. Combining lead variants from study I and II, after removing those in linkage disequilibrium (LD) ($r^2 \geq 0.1$ in European populations).
2. Including additional lead variants from our SuSiE analysis with LD $r^2 < 0.1$ in European populations with the above set of lead variants)
-- those with independent risk-associated signals at $P < 5 \times 10^{-8}$
-- those located in GWAS loci with independent risk-associated signals at $P < 1 \times 10^{-6}$
Identified 227 lead variants
Colorectal cancer
Our recent fine-mapping study: 238 lead variants (Chen et al., Nature Communications, 2024)
Box 1
Identified 213 lead variants
Lung cancer
Trans-ancestry GWAS: 36 risk variants (Byun et al., Nature Genetics, 2022)
Box 1
Identified 26 lead variants
Prostate cancer
GWAS data in European populations (N=140,306)
Box 2
Trans-ancestry GWAS (N =234,253) (Conti et al., Nature Genetics, 2021)
Box 3
Identified 213 lead variants
Ovarian cancer
GWAS data in European populations (N = 63,347)
Box 2
European population GWAS (Phelan et al., Nature Genetics, 2017)
Box 3
Identified 13 lead variants
Pancreatic cancer
GWAS data in European populations (N = 21,536)
Box 2
European population GWAS (Klein et al., Nature Communication, 2018)
Box 3
Identified 18 lead variants
Box 1
Only including lead variants with $P < 1 \times 10^{-6}$ from analysis in European populations or Trans-ancestry GWAS.
Box 2 – SuSiE analysis:
Including lead variants with independent risk-associated signals
-- those with $P < 5 \times 10^{-8}$
-- those with $P < 1 \times 10^{-6}$, located in GWAS loci (i.e., within a 1Mb region)
Box 3 – previous GWAS
Including additional GWAS-identified risk variants with $P < 1 \times 10^{-6}$ in European populations from the GWAS study, and LD $r^2 < 0.1$ in European populations with the above set of lead variants

A


Colorectal cancer
Lung cancer
Breast cancer
Prostate cancer
32
4
2
2
1
0
0
21
26
0
0
2
0
0
0

B


Common proteins identified in at least two cancers
Number of proteins commonly identified in at least two cancers
25
16
14
7
3
2
2
1
1
1
1
1
0
5
10
15
25
Breast cancer
Colorectal cancer
Lung cancer
Ovarian cancer
Pancreatic cancer
Prostate cancer


Cancer
Colorectal
Lung
Breast
Prostate
Drugs approved to treat diseases
Drugs under clinical trials to treat diseases