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

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], Graham Casey[15], Xusheng Wang[16], Adetunji T. Toriola[17], Li Li[18], Bhuminder Singh[19], Ken S Lau[19], Li Zhou[20], 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*]

Affiliations:

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA

[2] Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Alberta, Canada

[3] Department of Biomedical Informatics, Vanderbilt University School of Medicine, 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

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

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 uncover five drugs (Haloperidol, Trazodone, Tranexamic Acid, Haloperidol, and Captopril) associated with increased cancer risk and two drugs (Caffeine and Acetazolamide) linked to reduced colorectal cancer risk. This study offers novel insights into therapeutic drugs targeting risk proteins for cancer prevention and intervention.

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 expression has 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.

Results

Overall analysis workflow

In Figure 1, we outlined several main steps of a comprehensive integrative analysis of GWAS, pQTLs, druggable proteins, and EHR data. First, we examined previously identified risk loci from six cancer types based on the most recent GWAS in breast (N = 247,173), ovary (N = 63,347), prostate (N = 140,306), colorectum (N = 254,791), lung (N = 85,716), and pancreas (N = 21,536). Through additional fine-mapping analysis using SuSiE[48], we characterized the most significantly associated variants (the lead variants) with independent association signals at each risk locus for each cancer (**Fig. 1a; Online Methods**). Second, we analyzed *cis*-pQTL results for the lead variants using proteomics data from individuals of European descent from ARIC[46], deCODE[47], and UKB-PPP[37]. We conducted fixed-effect meta-analyses of summary statistics *cis*-pQTLs from ARIC[46] and deCODE[47] through the same SOMAscan® platform (covering > 4,500 proteins) . We

combined them with the pQTL results from UKB-PPP through the Olink platform to identify potential risk proteins. For proteins that satisfied the significance threshold after multiple testing corrections, we further performed colocalization analyses to determine cancer risk proteins with high confidence through evaluating the likelihood of shared causal variants between pQTLs and GWAS (**Fig. 1a)**. Third, these risk proteins with evidence of colocalization were further annotated based on drug-protein information from four drugs/compounds databases (DrugBank[49], ChEMBL[50], the Therapeutic Target Database[51] (TTD) and OpenTargets[52] (**Fig. 1b**). We next identified druggable proteins that are therapeutic targets of approved drugs or undergoing clinical trials for cancer treatment or other indications. Finally, we focused on drugs approved for other indications. We built emulation of treated-control drug trials under the Inverse Probability of Treatment weighting (IPTW) framework[53] through the analysis of over 3.5 million EHRs at VUMC. In these emulations, we used the Cox proportional hazard model for each trial to evaluate the hazard ratio (HR) of the specific cancer risk between the treated focal drug and the control drug. The significance of each focal drug (HR and *P* value) was derived from a random-effects meta-analysis of results across its balanced trials (**Fig. 1c, Online Methods**).

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

To characterize lead variants at each locus for each cancer type, we collected the reported risk variants from previous fine-mapping or GWAS. Using breast cancer as an example, we included 196 lead variants with independent association signals at loci from a previous fine-mapping study based on conditional association analysis[54] and additional 32 genetic variants identified from a recent GWAS[6]. We then performed additional fine-mapping analysis using SuSiE[48] based on summary statistics of GWAS (N = 247,173) from the Breast Cancer Association Consortium (BCAC, **Supplementary Table 1**). After integrating the previous results with new fine-mapping efforts, we identified 227 lead variants with independently associated with cancer risk at each locus through several processing steps, which included lead variant selection ($P < 1 \times 10^{-6}$ in European populations) and evaluating a linkage disequilibrium (LD) ($r^2 < 0.1$) among the identified risk variants (**Extended Data Fig. 1; Online Methods**). Similarly, we characterized lead variants from previous GWAS and our fine-mapping studies for colorectal[55] and other cancers (**Extended Data Fig. 1; Supplementary Table 1; Online Methods**). In our analysis, 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 (**Fig. 2** and **Supplementary Tables 2 and 3; Online Methods**).

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 (**Fig. 2; Supplementary Table 4**). Notably, 312 of the identified proteins (85.4% of 365) among these cancer types have not been reported in previous proteomics-based MR studies[32-36,56,57] (**Supplementary Table 5**). 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. 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 (**Fig. 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 (**Online 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 (**Fig. 3a-b**, **Supplementary Table 6**). Of these, 74 proteins (73.2% of 101) have not been previously linked to cancer risk (**Supplementary Table 7**). 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 ($r = 0.66$, $P = 2 \times 10^{-4}$; **Fig. 3c**). In particular, seven proteins were highlighted as cancer-driver proteins[58,59] and Cancer Gene Census (CGC)[60], 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 next 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 (**Online 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 (**Supplementary Table 7**). 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 (**Online Methods**) Our results showed that 17 genes were likely regulated by the closest putative regulatory variants with either promoter and/or enhancer activities (**Supplementary Table 8**). We further investigated the potential distal regulatory effects of putative functional variants on these genes by analyzing chromatin-chromatin interaction data (**Online Methods**). We found that 39 genes were regulated distally by putative functional variants through long-term promoter-enhancer interactions (**Supplementary Table 9**). 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 partially or wholly supported by functional genomics analyses (**Supplementary Table 10)**.

Identifying druggable proteins

Using data from DrugBank[49], ChEMBL[50], the Therapeutic Target Database[51] (TTD) and OpenTargets[52], we comprehensively annotated our proteins as therapeutic targets of approved or clinical-stage drugs (**Online Methods**). 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 (**Fig. 4**, **Supplementary Table 11**). Specifically, we found 19 proteins targeted by 133 drugs either approved or under clinical trials to treat cancers (**Fig. 5, Supplementary Table 12**). Our results also provide evidence that the remaining draggable proteins are targeted by 197 drugs used for treating indications other than cancer (**Extended Data Fig. 3**).

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 indications based on real-world EHRs from the VUMC Synthetic Derivative (SD) database. Given a focal drug, we first emulated its trials by building control patient groups who were exposed to similar treated drugs under the same ATC-L2 category (**Online**

Methods; **Supplementary Table 13**). To mimic randomized controlled trials (RCT) to evaluate the focal drug's effect, we applied the Inverse Probability of Treatment Weighting (IPTW) framework[61] to create a pseudo-population wherein confounding variables are evenly distributed between the treated and control groups (**Online Methods**). After discarding the trials with less than 500 eligible patients in either of the groups, we analyzed 14 treated drugs with 335 balanced trials. Our analysis revealed that five drugs were linked to an increased risk of cancer: Haloperidol (HR = 1.76; $P = 1.6 \times 10^{-23}$; targeting HLA-A protein) and Trazodone (HR = 1.32; $P = 2.3 \times 10^{-12}$; targeting HLA-A protein) for breast cancer; Tranexamic Acid (HR = 1.53; $P = 1.1 \times 10^{-3}$; targeting PLG protein) and Sirolimus (HR = 1.71; $P = 1.1 \times 10^{-28}$; targeting TYRO3 protein) for prostate cancer; and Haloperidol (HR = 2.62, $P = 6.6 \times 10^{-20}$, targeting HLA-A protein) and Captopril (HR = 1.65; $P = 2.2 \times 10^{-9}$; targeting TF protein) for colorectal cancer (**Fig. 6**). In contrast, we also found that two drugs associated with a decreased risk of colorectal cancer: Caffeine (HR = 0.74, $P = 9.3 \times 10^{-5}$, targeting ALDH2 protein) and Acetazolamide (HR = 0.72; $P = 1.1 \times 10^{-20}$; targeting HLA-A protein) (**Fig. 6**).

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)[62].

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.

We uncovered five non-cancer drugs associated with increased cancer risk: Tranexamic Acid (PLG), Sirolimus (TYRO3), Haloperidol (HLA-A), Trazodone (HLA-A), and Captopril (TF). Our genetic evidence or previous pharmacological studies further support these findings. Specifically, Tranexamic Acid, an antifibrinolytic agent used to block the breakdown of blood clots and prevent bleeding, was associated with an increased risk of prostate cancer. Our findings suggest its potential inhibition of the protein expression of human plasminogen (PLG), based on data from the ChEMBL database. Our GWAS and pQTL results indicate that PLG may serve as a potential tumor suppressor, supported by evidence of the risk allele C of rs9347480 being associated with increased prostate cancer risk ($P = 1.41 \times 10^{-07}$) and decreased protein expression ($P = 4.57 \times 10^{-33}$). Additionally, this protein also shows notable evidence of decreased gene expression ($P = 0.038$) in prostate tumor samples compared to normal samples, as observed in data from TCGA. The drug Sirolimus, primarily used to treat immune system and eye diseases, can potentially affect receptor tyrosine kinases, TYRO3, a known protein important for prostate cancer development[63,64]. In breast and colorectal cancers, we identified two candidate drugs (Haloperidol and Trazodone) targeting the major histocompatibility complex, HLA-A, an essential protein for the immune system's defense against cancer development. Interestingly, our analysis suggests that Haloperidol, a type of antipsychotic treatment, is highly likely to increase the risk of both breast and colon cancer. Haloperidol, the first-generation antipsychotics, has been reported to be a carcinogenic compound[65] and its exposure of five years or more was associated with an increased risk of breast cancer in a Finland nationwide study[62]. Consistently, another prior study showed its notably increased risks of colorectal cancer in patients with schizophrenia who take antipsychotic medications[66]. We also found that Captopril, originally for cardiovascular diseases and promisingly repurposed for cancer treatment in clinical trials and several studies[67-70], has the potential to increase the risk of colorectal cancer, aligning with a previous study[71].

Conversely, our study also identified two non-cancer drugs (Caffeine and Acetazolamide) associated with a reduced risk of colorectal cancer. Acetazolamide, prioritized by the risk protein named TF, exhibited a notable effect in preventing colorectal cancer development (HR = 0.72, $P = 1.1 \times 10^{-20}$). In line with our findings, prior studies demonstrated its role in inhibiting cell viability, migration, and colony formation ability of

colorectal cancer cells[72], as well as its ability to suppress the development of intestinal polyps in Min Mice[73]. In addition, Caffeine, a drug prioritized by the risk protein ALDH2 in colorectal cancer, has been shown to exert a protective effect on colorectal cancer by prior studies[74,75]. However, such low-risk association may vary by colon subsites[76] and specific populations[77]. Of note, a clinical trial (NCT05692024) is undergoing the recruitment phase to evaluate the effects of instant coffee on the gut microbiome, metabolome, liver fat, and fibrosis in colorectal cancer patients.

Although the larger sample size of the European-ancestry study available for both GWAS and proteomics enabled us to identify a larger number of association signals for risk protein discovery based on colocalization analysis, our study was primarily limited to individuals of European ancestry and further investigations are needed to assess the relevance of these proteins in non-European populations. Millions of EHRs provide an unprecedented opportunity to systematically evaluate non-cancer drugs' effect on risk of cancer development. Especially these therapeutic drugs have been used for a long time to treat diseases other than cancers, which can provide appropriate statistical power for the analysis. While this approach is limited in only examining the cancer risk of common approved drugs, it serves as an efficient complementary method to the pre-clinical data analysis for cancer prevention and treatment. Despite the supportive evidence of Acetazolamide *in vitro* and *in vivo*, it remains necessary to evaluate the effects of our reported candidate drugs through both *in vitro* and *in vivo* assays in future investigations.

Online Methods

Data resources

The GWAS summary statistics data of European descendants for breast, prostate, ovarian, and lung cancers were downloaded and compiled from their corresponding consortia, including the Breast Cancer Association Consortium (BCAC)[6] (N = 247,173, 133,384 cases and 113,789 controls), the Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3)[13] (N = 85,716, 29,266 cases and 56,450 controls), the Ovary Cancer Association Consortium (OCAC)[11] (N = 63,347, 22,406 cases and 40,941 controls), and the Pancreatic Cancer Case-Control Consortium (PanC4)[10] (N = 21,536, 9,040 cases and 12,496 controls), and the Prostate Cancer Association Group Investigate Cancer Associated Alterations in the Genome (PRACTICAL)[78] (N = 140,306, 79,194 cases and 61,112 controls). For colorectal cancer, we included GWAS data of 125,487 subjects from the European population.[19,79,80] In addition, the GWAS data (N = 254,791) consisting of 100,204

colorectal cancer cases and 154,587 controls from European and Asian populations[7] were also used in our analysis.

The large-scale *cis*-protein quantitative trait loci (*cis*-pQTLs) among European-ancestry populations were analyzed based on three proteomics datasets: UKB-PPP[37] (N = 34,557, 2,922 plasma proteins), ARIC[44] (N = 7,213, 4,657 plasma proteins) and deCODE genetics[45] (N = 35,559, 4,907 plasma proteins). Detailed descriptions of sample collection and processes of the *cis*-pQTL analyses from the above proteomics datasets have been described in previous studies[37,46,47].

We utilized the synthetic derivative (SD) database at Vanderbilt University Medical Center (VUMC)[81]. This VUMC SD database contains de-identified clinical information derived from Vanderbilt's electronic medical record The SD has longitudinal clinical data for over 3.5 million individuals, including patient demographics, medical history, laboratory results, and medication history.

Characterization of lead variants in six types of cancer

In breast cancer, we included 196 strong independent association signals at $P < 1 \times 10^{-6}$ from a fine-mapping study[54] and 32 risk variants from a GWAS[6]. We first combined the reported lead variants from these two studies after removing those variants in LD ($r^2 < 0.1$ in European populations). We further included additional lead variants from SuSiE fine-mapping analysis on GWAS (N = 247,173)[48], 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)[82]. 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.

For colorectal cancer, we analyzed 238 lead variants from our recent fine-mapping study[55] based on the GWAS data from 254,791 participants in both European and Asian populations. We characterized 233 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. 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 next 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. 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 26 risk variants at $P < 1 \times 10^{-6}$ in European populations from the trans-ancestry GWAS[9].

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]. 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[83]. *Cis*-pQTLs from the UKB-PPP (assayed through Olink platform) were independently analyzed. In 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* [84] and summary data-based Mendelian Randomization (SMR)[85]. 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 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[86]. 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[87], 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[88], FANTOM5[89], EnhancerAtlas[90], and super-enhancer[91]. 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[92,93]. Similarly, we analyzed their differential gene expression between tumor and normal tissue using data from TCGA, also through the UALCAN website.

Inclusion of patients for a focal drug and its control drugs

To evaluate the impact of a focal drug on cancer development, we conducted comparisons between its effects and those of its 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. We formulated emulation trials, each containing one treated patient group (taking the focal drug) and one control patient group (taking the control drug). One focal drug may have multiple trials, depending on the number of potential control drugs belonging to the same ATC-L2 category. Next, we enrolled patients for the treated group and the control group from the VUMC SD based on the following criteria: 1) patients aged ≥ 40 at the time of the latest EHRs or the initial diagnosis of cancer; 2) availability of at least one year of EHRs before the first prescription of the treated/control drug (index date); 3) for cancer patients, a minimum of two exposures to the treated or control drug during the follow-up period (from the index date to the three months before cancer diagnosis); 4) for non-cancer individuals, a minimum of two exposures to control drugs during the follow-up period (from the index date to the date of the latest EHRs). Finally, we precluded patients who were prescribed both treated and control drugs and discarded the trials with less than 500 eligible patients in either patient group[40].

Emulation of treated-control drugs balanced trials

In the IPTW framework[61], individuals are assigned weights based on the inverse of their propensity scores (PS), which represent their probability of being exposed to risk factors or a specific intervention, such as a treated drug, based on their baseline characteristics. In this study, we followed Zang's work[40] and trained a logistic regression propensity score (LR-PS) model with L1 or L2 regularization on patients' treatment assignments Z and covariates, including age, gender, comorbidities, etc. (**Supplementary Table 14**). We trained and selected the logistic model (Eq.1) with the highest area under curve (AUC) using a 10-folder cross-validation. We used the selected model to calculate all patient's stabilized weights (Eq. 2). These weights are used to calculate the standardized mean difference (SMD, Eq.3) of the covariate's prevalence in treated and control groups. A covariate d is defined as unbalanced if $SMD(d) > 0.1$ in IPTW framework (Eqs. 3, 4). A trial is balanced if it contains ≤10% unbalanced covariates (Eq. 5).

The logistic regression is defined as follows:

$$\log\left(\frac{P(\boldsymbol{Z}=1)}{1-P(\boldsymbol{Z}=1)}\right) = \beta_0 + \beta_1\boldsymbol{X}_1 + \beta_2\boldsymbol{X}_2 + \cdots + \beta_n\boldsymbol{X}_n \quad (1)$$

$$P(\boldsymbol{Z}=1) = \frac{1}{1+\ e^{-(\sum \beta_j \boldsymbol{X}_j + \beta_0)}}$$

where $\boldsymbol{Z}$ refers to treatment assignment (1 for treated patient group and 0 for control patient group) and $\boldsymbol{X}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ for baseline covariates. The propensity score is defined as $P(\boldsymbol{Z} = 1|\, \boldsymbol{X})$ and the stabilized IPTW of each individual is calculated as follows:

$$\mathbf{w} = \frac{\boldsymbol{Z} \times P(\boldsymbol{Z} = 1)}{P(\boldsymbol{Z} = 1|\, \boldsymbol{X})} + \frac{(1 - \boldsymbol{Z}) \times (1 - P(\boldsymbol{Z} = 1))}{1 - P(\boldsymbol{Z} = 1|\, \boldsymbol{X})} \quad (2)$$

Standardized mean difference is calculated as following:

$$SMD(\boldsymbol{x}_{treat} - \boldsymbol{x}_{control}) = \frac{|\boldsymbol{\mu}_{treat} - \boldsymbol{\mu}_{control}|}{\sqrt{(\boldsymbol{S}^2_{treat} + \boldsymbol{S}^2_{control})/2}} \quad (3)$$

$\boldsymbol{x}_{treat}, \boldsymbol{x}_{control} \in R^D$, representing vectors of D number of covariates of treated group and control group respectively; $\boldsymbol{\mu}_{treat}, \boldsymbol{\mu}_{control}$ are their sample means, and $\boldsymbol{S}^2_{treat}, \boldsymbol{S}^2_{control}$ are their sample variances. In IPTW framework, the weighted sample mean $\boldsymbol{\mu}_w$ and sample variance $\boldsymbol{S}^2_w$ are calculated as following:

$$\boldsymbol{\mu}_w = \frac{\sum \boldsymbol{w}_i \boldsymbol{x}_i}{\sum \boldsymbol{w}_i} \quad (4)$$

$$\boldsymbol{S}^2_w = \frac{\sum \boldsymbol{w}_i}{(\sum \boldsymbol{w}_i)^2 - \sum \boldsymbol{w}_i^2} \sum \boldsymbol{w}_i (\boldsymbol{x}_i - \boldsymbol{\mu}_w)^2$$

Number of unbalanced covariates are calculated as following:

$$n = \sum_{d=1}^{D} \mathbb{1}[SMD(d) > 0.1] \quad (5)$$

$$where\ D\ is\ the\ total\ number\ of\ covariates\ in\ the\ model, d\ is\ one\ covariate$$

Logistic regression propensity score (LR-PS) hyperparameter selection and model training

To select the optimal regulation penalty weight (λ), we applied 10-fold cross-validation on a list of lambda elements $\lambda \in$ [0.005, 0.01, 0.05, 0.1, 0.5]. Specifically, the logistic model was trained on 9 training folders, and learnable parameters (β) were estimated through minimizing the binary cross-entropy loss with L1 (Eq. 6) or L2 penalty (Eq. 7). On the left-one out validation folder (k), we calculated SMD_k (Eq. 3) values for D covariates based on individuals' weights (Eq. 2) as well as the number unbalanced covariates n_k (Eq. 5). In addition, we evaluated trained model's prediction performance using area under curve (AUC_k) on the validation dataset. The same processes were repeated 10 times. We defined the optimal hyperparameter value is the value generates the smallest averaged n_k. For two hyperparameter values generate approximate n_k, the one with larger averaged AUC_k is the optimal. Finally, we trained LR-PS model's learnable

parameters (β) on all subjects with the optimal hyperparameter and leveraged this trained model to compute weights and the proportion of imbalanced covariates to pinpoint balanced trials.

Binary cross-entropy loss function with LASSO (L1) penalization:

$$argmin_{\boldsymbol{\beta}}(\,Loss) = -\frac{1}{n}\left(\sum_{i=1}^{n} z_i \log(P(z_i = 1)) + (1 - z_i)\log(1 - P(z_i = 1))\right) + \lambda_1 \sum_{j=1}^{b} |\beta_j| \quad (6)$$

Binary cross-entropy loss function with ridge (L2) penalization:

$$argmin_{\boldsymbol{\beta}}(\,Loss) = -\frac{1}{n}\left(\sum_{i=1}^{n} z_i \log\big(P(z_i = 1)\big) + (1 - z_i)\log\big(1 - P(z_i = 1)\big)\right) + \lambda_2 \sum_{j=1}^{b} \beta_j^2 \quad (7)$$

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

We evaluate subjects' hazard of developing/preventing cancer in balanced treated-control trials through survival analyses. We applied weighted Cox proportional hazard model[94] using the lifelines 0.28.0 Python package to systematically evaluate the hazard ratio of developing cancer for patients taking treated drug vs patients taking control drugs (time-to-event). The time windows utilized in this study start from the earliest date in EHRs for prescription of the treated/control drug to patients and end at the date of the diagnosis of cancer (event) or the end of EHRs records (censored). We included unbalanced covariates (if exist) into Cox models. For a treated drug, its overall hazard ratio and p-value were obtained by applying a random effect meta-analysis on the hazard ratios from its eligible trials using the meta 7.0 R package (**Supplementary Table 15**). We reported that a treated drug has a significantly increasing or decreasing risk of cancer development, contrasting to its control drugs, if the overall hazard ratio has a $P < 0.05$ after Bonferroni correction (nominal $P = 3.5 \times 10^{-3}$ corresponding to 14 tests).

Data availability

Supplementary Table 1 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., *et al.* The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860 (2015).
2. Diogo, D., *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nature communications* **9**, 4285 (2018).
3. Finan, C., *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**(2017).
4. Peters, U. & Tomlinson, I. Utilizing Human Genetics to Develop Chemoprevention for Cancer-Too Good an Opportunity to be Missed. *Cancer Prev Res (Phila)* **17**, 7-12 (2024).
5. Jia, G., *et al.* 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 (2022).
6. Zhang, H., *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature genetics* **52**, 572-581 (2020).
7. Fernandez-Rozadilla, C., *et al.* 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 (2023).
8. Conti, D.V., *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature genetics* **53**, 65-75 (2021).
9. Byun, J., *et al.* Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nature genetics* **54**, 1167-1177 (2022).
10. Klein, A.P., *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat Commun* **9**, 556 (2018).
11. Phelan, C.M., *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature genetics* **49**, 680-691 (2017).
12. Lawrenson, K., *et al.* Genome-wide association studies identify susceptibility loci for epithelial ovarian cancer in east Asian women. *Gynecol Oncol* **153**, 343-355 (2019).
13. McKay, J.D., *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature genetics* **49**, 1126-1132 (2017).
14. Wen, W., *et al.* Genetic variations of DNA bindings of FOXA1 and co-factors in breast cancer susceptibility. *Nature communications* **12**, 5318 (2021).
15. Moreno, V., *et al.* Colon-specific eQTL analysis to inform on functional SNPs. *Br J Cancer* **119**, 971-977 (2018).
16. Chen, Z., *et al.* Identifying Putative Susceptibility Genes and Evaluating Their Associations with Somatic Mutations in Human Cancers. *American journal of human genetics* **105**, 477-492 (2019).
17. Guo, X., *et al.* A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *American journal of human genetics* **102**, 890-903 (2018).
18. He, J., *et al.* Integrating transcription factor occupancy with transcriptome-wide association analysis identifies susceptibility genes in human cancers. *Nature communications* **13**, 7118 (2022).
19. Guo, X., *et al.* Identifying Novel Susceptibility Genes for Colorectal Cancer Risk From a Transcriptome-Wide Association Study of 125,478 Subjects. *Gastroenterology* **160**, 1164-1178 e1166 (2021).
20. Yuan, Y., *et al.* Multi-omics analysis to identify susceptibility genes for colorectal cancer. *Human molecular genetics* **30**, 321-330 (2021).

21. Bien, S.A., *et al.* Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Hum Genet* **138**, 307-326 (2019).
22. Wu, L., *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics* **50**, 968-978 (2018).
23. Gao, G., *et al.* A joint transcriptome-wide association study across multiple tissues identifies candidate breast cancer susceptibility genes. *Am J Hum Genet* **110**, 950-962 (2023).
24. Mancuso, N., *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature communications* **9**, 4079 (2018).
25. Liu, D., *et al.* A transcriptome-wide association study identifies novel candidate susceptibility genes for prostate cancer risk. *Int J Cancer* **150**, 80-90 (2022).
26. Bosse, Y., *et al.* Transcriptome-wide association study reveals candidate causal genes for lung cancer. *International journal of cancer* **146**, 1862-1878 (2020).
27. Lu, Y., *et al.* A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. *Cancer research* **78**, 5419-5430 (2018).
28. Gusev, A., *et al.* A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature genetics* **51**, 815-823 (2019).
29. Zhong, J., *et al.* A Transcriptome-Wide Association Study Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer. *J Natl Cancer Inst* **112**, 1003-1012 (2020).
30. Zheng, C.J., Han, L.Y., Yap, C.W., Ji, Z.L., Cao, Z.W. & Chen, Y.Z. Therapeutic targets: Progress of their exploration and investigation of their characteristics (vol 58, pg 259, 2006). *Pharmacol Rev* **58**, 682-682 (2006).
31. Zhu, J., *et al.* Associations between Genetically Predicted Blood Protein Biomarkers and Pancreatic Cancer Risk. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **29**, 1501-1508 (2020).
32. Shu, X., *et al.* Evaluation of associations between genetically predicted circulating protein biomarkers and breast cancer risk. *International journal of cancer* **146**, 2130-2138 (2020).
33. Wu, L., *et al.* Analysis of Over 140,000 European Descendants Identifies Genetically Predicted Blood Protein Biomarkers Associated with Prostate Cancer Risk. *Cancer research* **79**, 4592-4598 (2019).
34. Gregga, I., *et al.* Predicted proteome association studies of breast, prostate, ovarian, and endometrial cancers implicate plasma protein regulation in cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* (2023).
35. Jia, G., *et al.* 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 (2023).
36. Considine, D.P.C., *et al.* Genetically predicted circulating protein biomarkers and ovarian cancer risk. *Gynecol Oncol* **160**, 506-513 (2021).
37. Sun, B.B., *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* (2023).
38. Tautermann, C.S. Current and Future Challenges in Modern Drug Discovery. *Methods Mol Biol* **2114**, 1-17 (2020).
39. Pushpakom, S., *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41-58 (2019).
40. Zang, C., *et al.* High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. *Nature communications* **14**, 8180 (2023).
41. Wu, Y., *et al.* 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 (2019).

42. Xu, H., *et al.* 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 (2015).
43. Bejan, C.A., Cahill, K.N., Staso, P.J., Choi, L., Peterson, J.F. & Phillips, E.J. DrugWAS: Drug-wide Association Studies for COVID-19 Drug Repurposing. *Clin Pharmacol Ther* **110**, 1537-1546 (2021).
44. Reznikov, L.R., *et al.* Identification of antiviral antihistamines for COVID-19 repurposing. *Biochem Biophys Res Commun* **538**, 173-179 (2021).
45. Liu, R., Wei, L. & Zhang, P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat Mach Intell* **3**, 68-75 (2021).
46. Zhang, J., *et al.* 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 (2022).
47. Ferkingstad, E., *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* **53**, 1712-1721 (2021).
48. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. 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 (2020).
49. Law, V., *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**, D1091-1097 (2014).
50. Gaulton, A., *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100-1107 (2012).
51. Zhou, Y., *et al.* TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* (2023).
52. Ochoa, D., *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res* **51**, D1353-D1359 (2023).
53. Chesnaye, N.C., *et al.* An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* **15**, 14-20 (2022).
54. Fachal, L., *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature genetics* **52**, 56-73 (2020).
55. Chen, Z., *et al.* Fine-mapping analysis including over 254,000 East Asian and European descendants identifies 136 putative colorectal cancer susceptibility genes. *Nat Commun* **15**, 3557 (2024).
56. Shu, X., *et al.* Associations between circulating proteins and risk of breast cancer by intrinsic subtypes: a Mendelian randomisation analysis. *Br J Cancer* **127**, 1507-1514 (2022).
57. Sun, J., *et al.* Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. *Genome Med* **15**(2023).
58. Bailey, M.H., *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034-1035 (2018).
59. Dietlein, F., *et al.* Identification of cancer driver genes based on nucleotide context. *Nature genetics* **52**, 208-218 (2020).
60. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I. & Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).
61. Austin, P.C. & Stuart, E.A. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* **34**, 3661-3679 (2015).
62. Taipale, H., Solmi, M., Lahteenvuo, M., Tanskanen, A., Correll, C.U. & Tiihonen, J. Antipsychotic use and risk of breast cancer in women with schizophrenia: a nationwide nested case-control study in Finland. *Lancet Psychiatry* **8**, 883-891 (2021).

63. Wu, G.*, et al.* Targeting Gas6/TAM in cancer cells and tumor microenvironment. *Mol Cancer* **17**, 20 (2018).
64. Jansen, F.H.*, et al.* Profiling of antibody production against xenograft-released proteins by protein microarrays discovers prostate cancer markers. *J Proteome Res* **11**, 728-735 (2012).
65. Chiang, J.Y.*, et al.* Haloperidol Instigates Endometrial Carcinogenesis and Cancer Progression by the NF-kappaB/CSF-1 Signaling Cascade. *Cancers (Basel)* **14**(2022).
66. Hippisley-Cox, J., Vinogradova, Y., Coupland, C. & Parker, C. Risk of malignancy in patients with schizophrenia or bipolar disorder: nested case-control study. *Arch Gen Psychiatry* **64**, 1368-1376 (2007).
67. Koh, S.L., Ager, E.I., Costa, P.L., Malcontenti-Wilson, C., Muralidharan, V. & Christophi, C. Blockade of the renin-angiotensin system inhibits growth of colorectal cancer liver metastases in the regenerating liver. *Clin Exp Metastasis* **31**, 395-405 (2014).
68. Childers, W.K. Interactions of the renin-angiotensin system in colorectal cancer and metastasis. *Int J Colorectal Dis* **30**, 749-752 (2015).
69. Riddiough, G.E.*, et al.* Captopril, a Renin-Angiotensin System Inhibitor, Attenuates Features of Tumor Invasion and Down-Regulates C-Myc Expression in a Mouse Model of Colorectal Cancer Liver Metastasis. *Cancers (Basel)* **13**(2021).
70. Kristensen, K.B., Hicks, B., Azoulay, L. & Pottegard, A. Use of ACE (Angiotensin-Converting Enzyme) Inhibitors and Risk of Lung Cancer: A Nationwide Nested Case-Control Study. *Circ Cardiovasc Qual Outcomes* **14**, e006687 (2021).
71. Yarmolinsky, J.*, et al.* Genetically proxied therapeutic inhibition of antihypertensive drug targets and risk of common cancers: A mendelian randomization analysis. *PLoS Med* **19**, e1003897 (2022).
72. Karakus, F., Eyol, E., Yilmaz, K. & Unuvar, S. Inhibition of cell proliferation, migration and colony formation of LS174T Cells by carbonic anhydrase inhibitor. *Afr Health Sci* **18**, 1303-1310 (2018).
73. Noma, N.*, et al.* Impact of Acetazolamide, a Carbonic Anhydrase Inhibitor, on the Development of Intestinal Polyps in Min Mice. *Int J Mol Sci* **18**(2017).
74. Schmit, S.L., Rennert, H.S., Rennert, G., Gruber, S.B.J.C.E., Biomarkers & Prevention. Coffee consumption and the risk of colorectal cancer. **25**, 634-639 (2016).
75. Sartini, M.*, et al.* Coffee Consumption and Risk of Colorectal Cancer: A Systematic Review and Meta-Analysis of Prospective Studies. *Nutrients* **11**(2019).
76. Um, C.Y., McCullough, M.L., Guinter, M.A., Campbell, P.T., Jacobs, E.J. & Gapstur, S.M. Coffee consumption and risk of colorectal cancer in the Cancer Prevention Study-II Nutrition Cohort. *Cancer Epidemiol* **67**, 101730 (2020).
77. Micek, A., Gniadek, A., Kawalec, P. & Brzostek, T. Coffee consumption and colorectal cancer risk: a dose-response meta-analysis on prospective cohort studies. *Int J Food Sci Nutr* **70**, 986-1006 (2019).
78. Schumacher, F.R.*, et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics* **50**, 928-936 (2018).
79. Huyghe, J.R.*, et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nature genetics* **51**, 76-87 (2019).
80. Chen, Z.*, et al.* Novel insights into genetic susceptibility for colorectal cancer from transcriptome-wide association and functional investigation. *J Natl Cancer Inst* **116**, 127-137 (2024).
81. Roden, D.M.*, et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84**, 362-369 (2008).
82. Weissbrod, O.*, et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat Genet* **52**, 1355-1363 (2020).
83. Schwarzer, G., Carpenter, J.R. & Rücker, G. *Meta-analysis with R*, (Springer, 2015).

84. Giambartolomei, C.*, et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).
85. Zhu, Z.*, et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-487 (2016).
86. Li, L.*, et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature genetics* **53**, 994-1005 (2021).
87. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934 (2012).
88. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2191-2199 (2014).
89. Lizio, M.*, et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22 (2015).
90. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic acids research* **48**, D58-D64 (2020).
91. Wang, Y.*, et al.* SEdb 2.0: a comprehensive super-enhancer database of human and mouse. *Nucleic Acids Res* **51**, D280-D290 (2023).
92. Chandrashekar, D.S.*, et al.* UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* **19**, 649-658 (2017).
93. Chandrashekar, D.S.*, et al.* UALCAN: An update to the integrated cancer data analysis platform. *Neoplasia* **25**, 18-27 (2022).
94. Lin, D.Y. & Wei, L.-J. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* **84**, 1074-1078 (1989).

Figures

Fig. 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).

Fig.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).

Fig. 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).

Fig. 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.

Fig. 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.

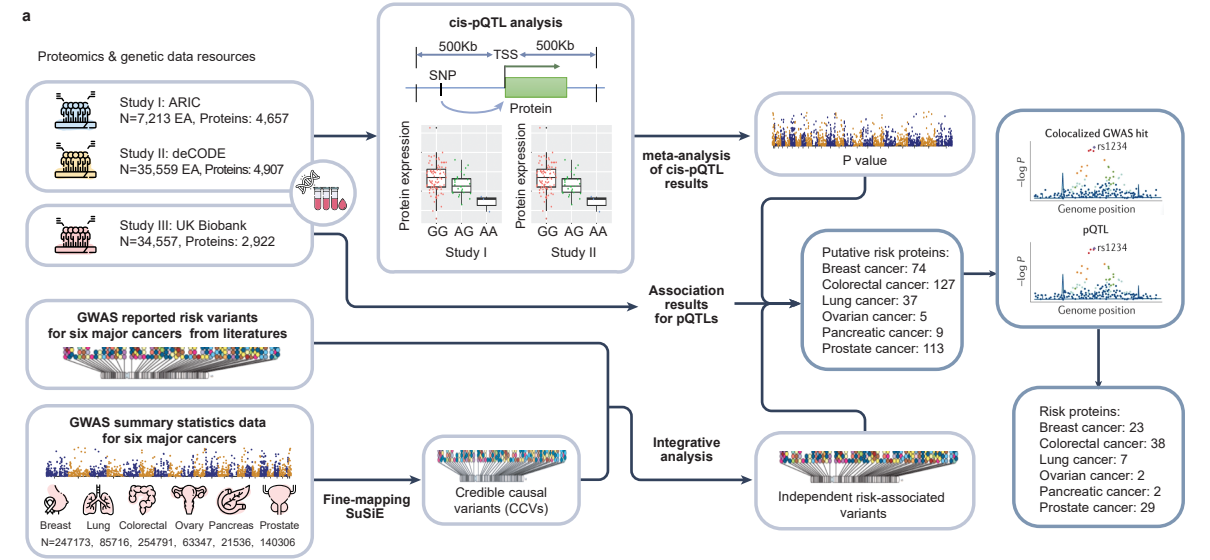
Fig. 6: Drugs approved for treated indications showing significant effects on cancer risk

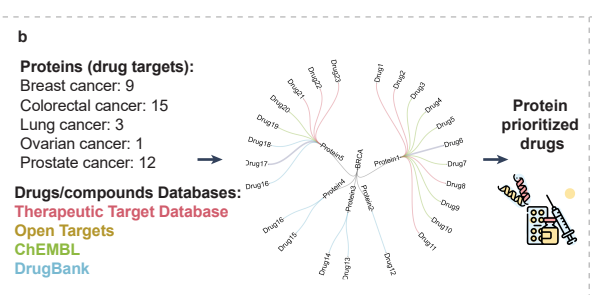
a, A table showing cancer risk alleles of lead variants, risk proteins, and drug name approved for indications. Positive associations are indicated by upward arrows, while negative associations are indicated by downward arrows. **b**, Boxplots showing differentially expressed proteins between normal and tumor colon tissues using data from CPTAC. **c**, An illustration of drugs linked to specific cancers based on the risk proteins targeted by the drugs. **d**, 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) compared to control groups (shown in purple). The shaded area represents the 95% confidence interval. The overall hazard ratio and P-value for the focal drug, determined through Cox proportional hazard models, are presented in the top right corner of each panel.

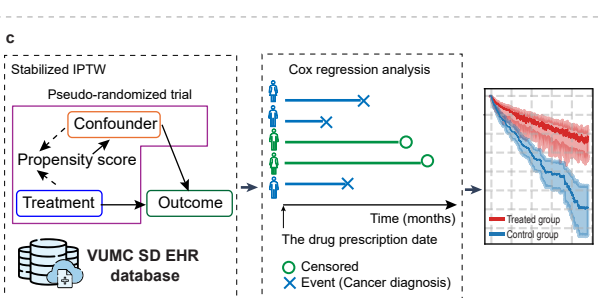
Extended Data Fig. 1: A flowchart for characterizing lead variants with independent risk signals in six cancer types. The analysis for each of the six major cancers: breast, lung, colorectal, ovarian, pancreatic, and prostate is separated by dashed lines. The detailed protocols of new efforts from our 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.

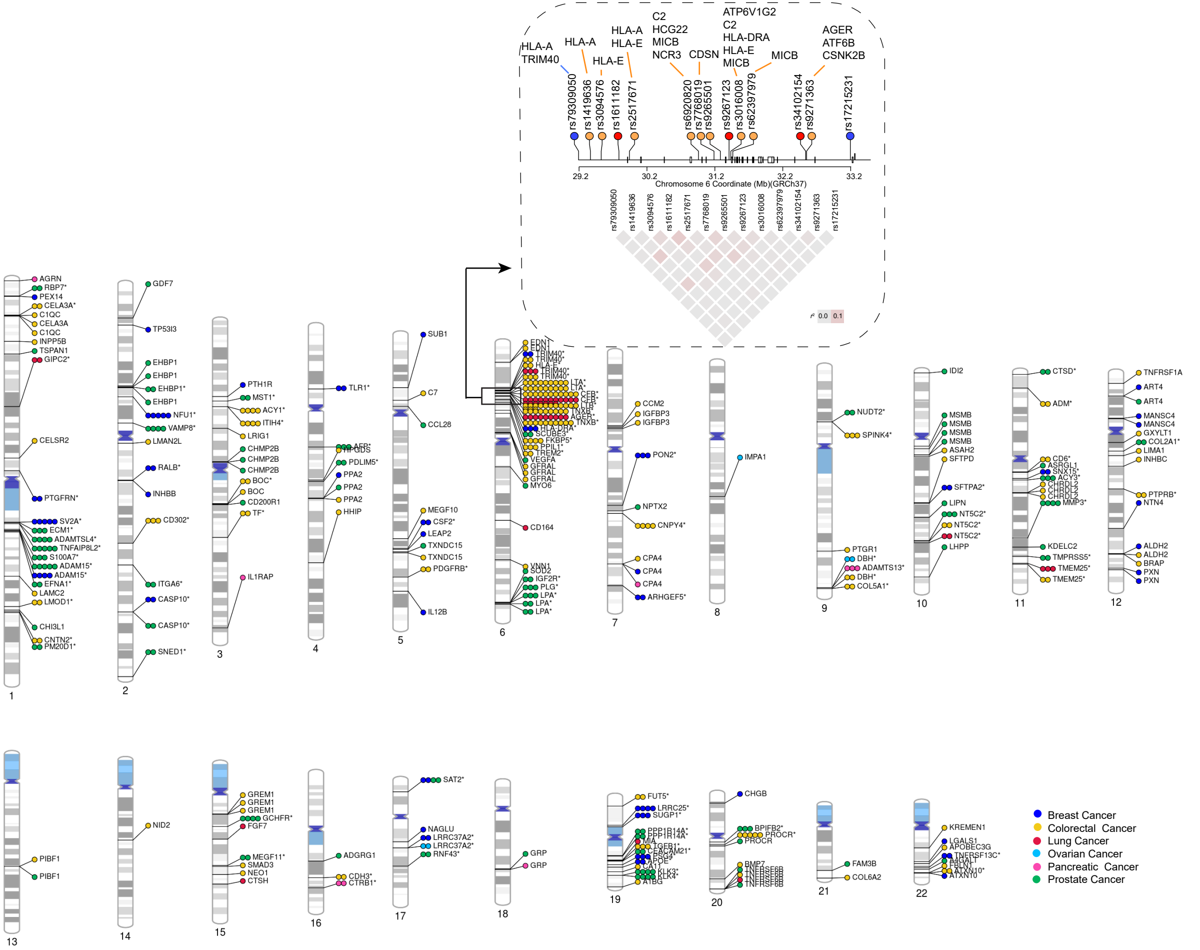
Extended Data Fig. 2: Common cancer risk proteins identified across six cancer types **a**, A Venn plot showing common proteins in breast, colorectal, lung and prostate cancer. **b**, A heatmap showing common proteins observed from at least two of six types of cancer.

Extended Data Fig. 3: 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.

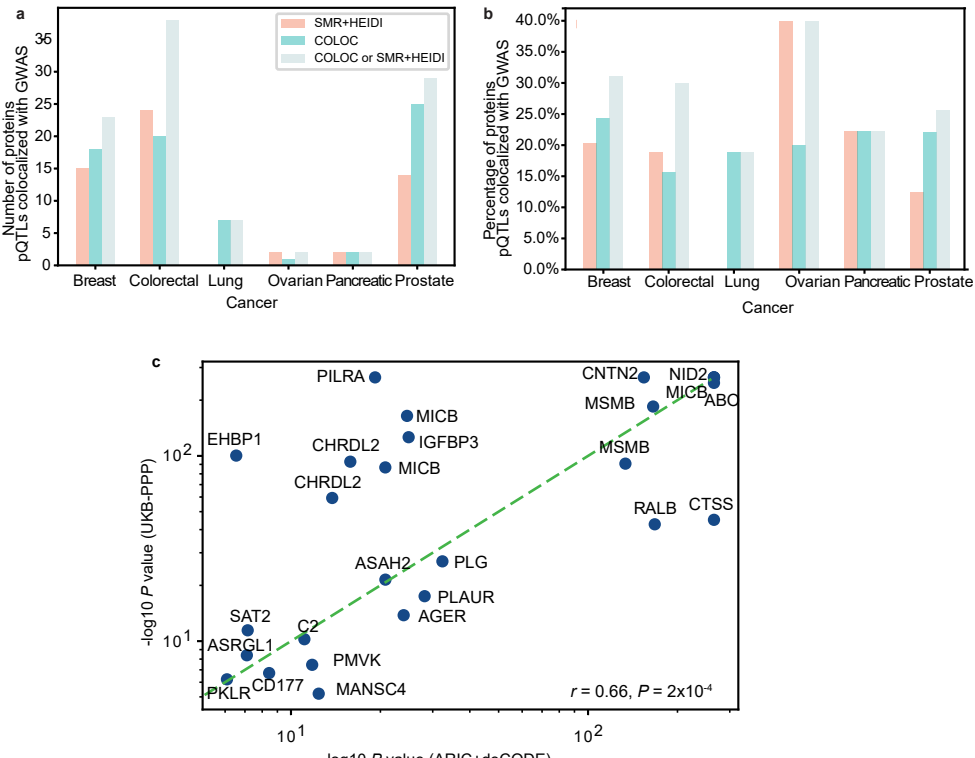

a
Proteomics & genetic data resources
Study I: ARIC
N=7,213 EA, Proteins: 4,657
Study II: deCODE
N=35,559 EA, Proteins: 4,907
Study III: UK Biobank
N=34,557, Proteins: 2,922
cis-pQTL analysis
500Kb
TSS
500Kb
SNP
Protein
Protein expression
GG AG AA
Study I
Study II
meta-analysis of cis-pQTL results
P value
Colocalized GWAS hit
rs1234
−log P
Genome position
pQTL
Association results for pQTLs
Putative risk proteins:
Breast cancer: 74
Colorectal cancer: 127
Lung cancer: 37
Ovarian cancer: 5
Pancreatic cancer: 9
Prostate cancer: 113
Risk proteins:
Breast cancer: 23
Colorectal cancer: 38
Lung cancer: 7
Ovarian cancer: 2
Pancreatic cancer: 2
Prostate cancer: 29
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

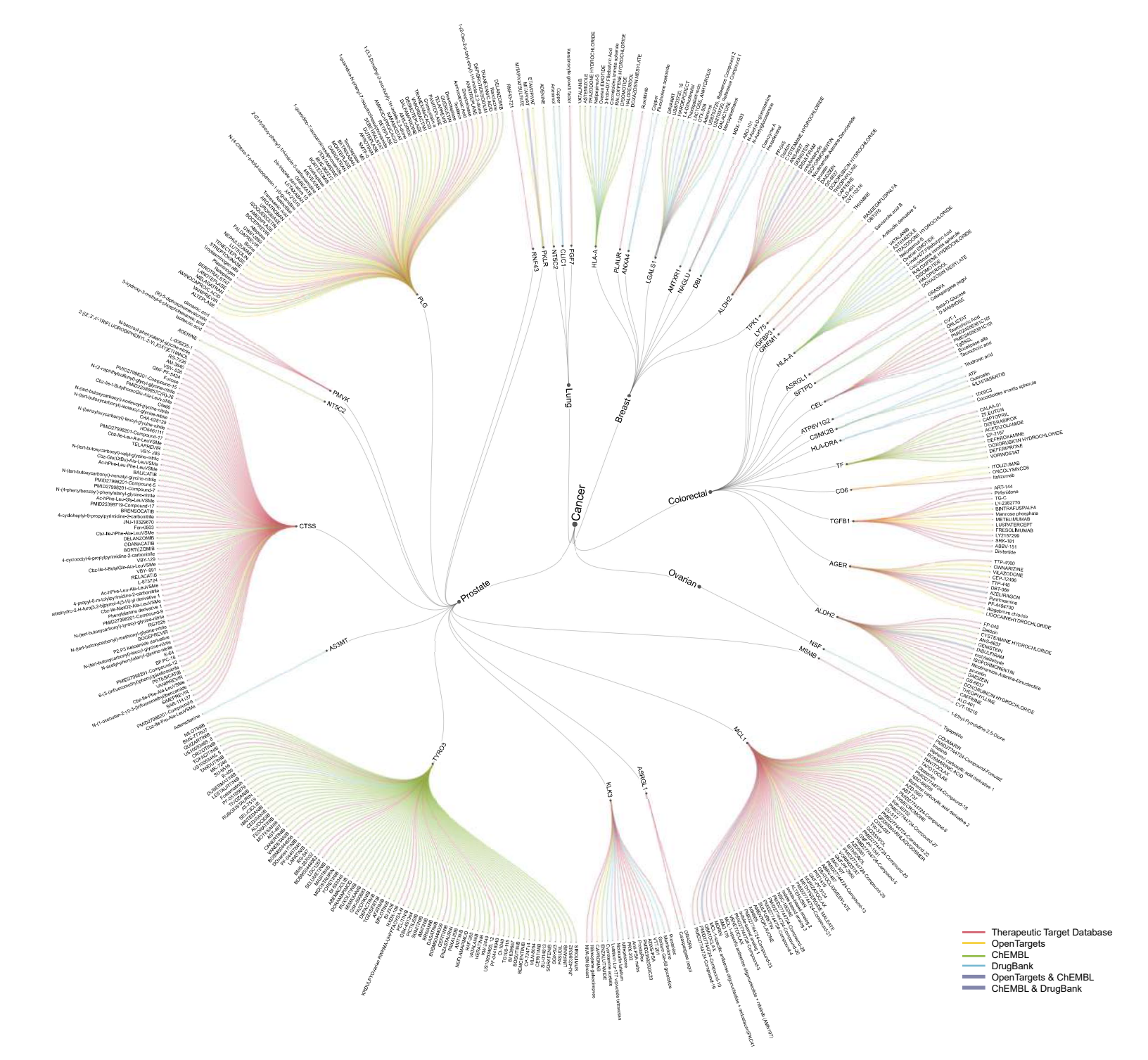

b
Proteins (drug targets):
Breast cancer: 9
Colorectal cancer: 15
Lung cancer: 3
Ovarian cancer: 1
Prostate cancer: 12
Drugs/compounds Databases:
Therapeutic Target Database
Open Targets
ChEMBL
DrugBank
Protein prioritized drugs


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

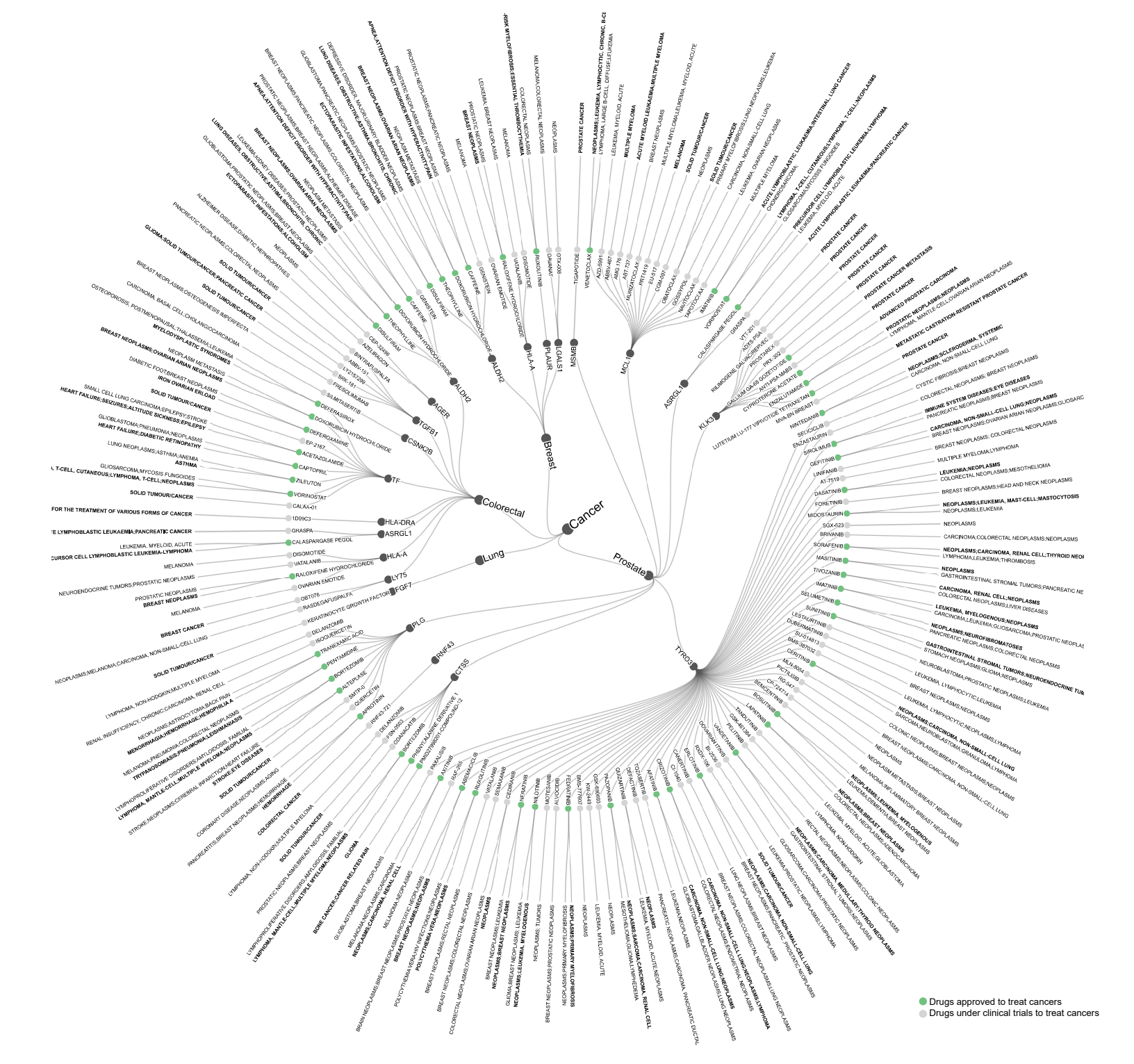


HLA-A
TRIM40
HLA-A
HLA-E
HLA-A
HLA-E
C2
HCG22
MICB
NCR3
CDSN
ATP6V1G2
C2
HLA-DRA
HLA-E
MICB
MICB
AGER
ATF6B
CSNK2B
rs79309050
rs1419636
rs3094576
rs1611182
rs2517671
rs6920820
rs7768019
rs9265501
rs9267123
rs3016008
rs62397979
rs34102154
rs9271363
rs17215231
29.2
30.2
31.2
32.2
33.2
Chromosome 6 Coordinate (Mb)(GRCh37)
r² 0.0 0.1
AGRN
RBP7*
PEX14
CELA3A*
C1QC
CELA3A
C1QC
INPP5B
TSPAN1
GIPC2*
CELSR2
PTGFRN*
SV2A*
ECM1*
ADAMTSL4*
TNFAIP8L2*
S100A7*
ADAM15*
ADAM15*
EFNA1*
LAMC2
LMOD1*
CHI3L1
CNTN2*
PM20D1*
1
GDF7
TP53I3
EHBP1
EHBP1
EHBP1*
EHBP1
NFU1*
VAMP8*
LMAN2L
RALB*
INHBB
CD302*
ITGA6*
CASP10*
CASP10*
SNED1*
2
PTH1R
MST1*
ACY1*
ITIH4*
LRIG1
CHMP2B
CHMP2B
CHMP2B
BOC*
BOC
CD200R1
TF*
IL1RAP
3
TLR1*
AFP*
PDLIM5*
PPA2
PPA2
PPA2
HHIP
4
SUB1
C7
CCL28
MEGF10
CSF2*
LEAP2
TXNDC15
TXNDC15
PDGFRB*
IL12B
5
EDN1
EDN1
TRIM40*
TRIM40*
HLA-E
TRIM40*
TRIM40*
LTA*
LTA*
CFB*
CFB*
LTB*
TNXB*
AGER*
TNXB*
HLA-DRA*
SCUBE3*
FKBP5*
PPIL1*
TREM2*
VEGFA
GFRAL
GFRAL
GFRAL
MYO6
CD164
VNN1
SOD2
IGF2R*
PLG*
LPA*
LPA*
LPA*
6
CCM2
IGFBP3
IGFBP3
PON2*
NPTX2
CNPY4*
CPA4
CPA4
CPA4
ARHGEF5*
7
IMPA1
8
NUDT2*
SPINK4*
PTGR1
DBH*
ADAMTS13*
DBH*
COL5A1*
9
IDI2
MSMB
MSMB
MSMB
MSMB
ASAH2
SFTPD
SFTPA2*
LIPN
NT5C2*
NT5C2*
NT5C2*
LHPP
10
CTSD*
ADM*
CD6*
ASRGL1
SNX15*
ACY3*
CHRDL2
CHRDL2
CHRDL2
MMP3*
KDELC2
TMPRSS5*
TMEM25*
TMEM25*
11
TNFRSF1A
ART4
ART4
MANSC4
MANSC4
GXYLT1
COL2A1*
LIMA1
INHBC
PTPRB*
NTN4
ALDH2
ALDH2
BRAP
PXN
PXN
12
PIBF1
PIBF1
13
NID2
14
GREM1
GREM1
GREM1
GCHFR*
FGF7
MEGF11*
SMAD3
NEO1
CTSH
15
ADGRG1
CDH3*
CTRB1*
16
SAT2*
NAGLU
LRRC37A2*
LRRC37A2*
RNF43*
17
GRP
GRP
18
FUT5*
LRRC25*
SUGP1*
PPP1R14A*
PPP1R14A*
MIA
TGFB1*
CEACAM21*
PSG4*
APOE
CA11
KLK3*
KLK4*
A1BG
19
CHGB
BPIFB2*
PROCR*
PROCR
BMP7
TNFRSF6B
TNFRSF6B
TNFRSF6B
TNFRSF6B
20
FAM3B
COL6A2
21
KREMEN1
LGALS1
APOBEC3G
TNFRSF13C*
ATXN10*
ATXN10
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
EHBP1
CHRDL2
IGFBP3
MSMB
CHRDL2
MICB
RALB
CTSS
ASAH2
PLG
PLAUR
SAT2
C2
AGER
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)

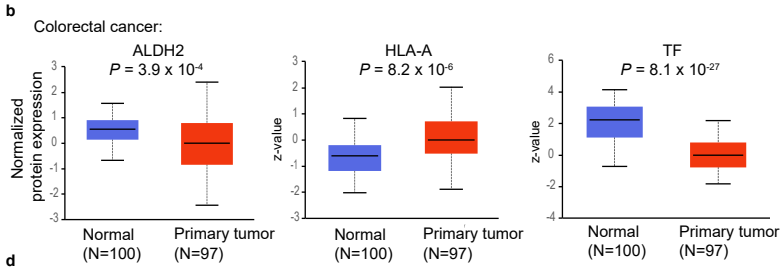


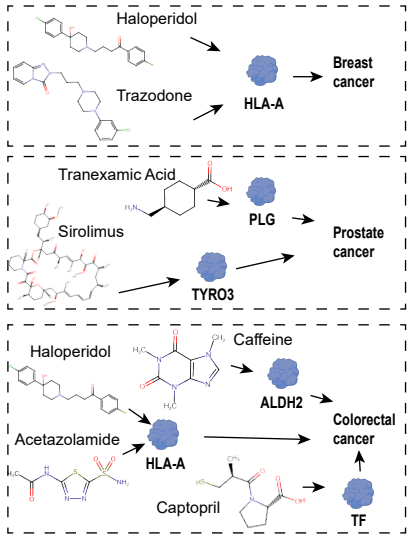
Therapeutic Target Database
OpenTargets
ChEMBL
DrugBank
OpenTargets & ChEMBL
ChEMBL & DrugBank

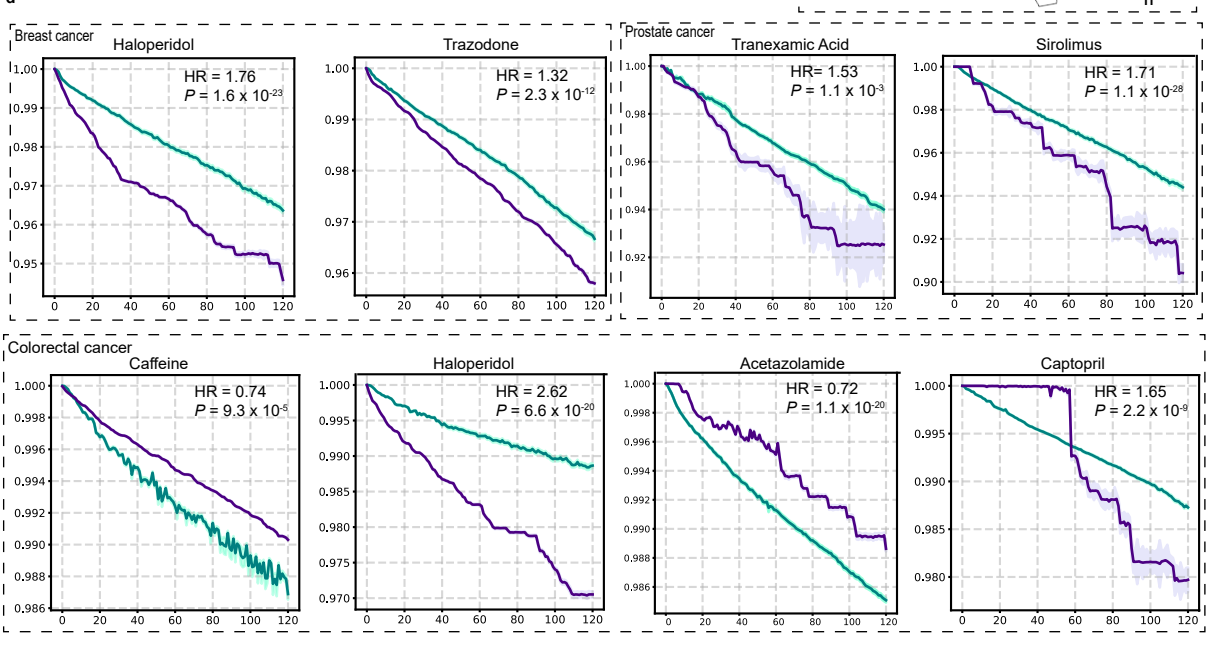

Cancer
Breast
Colorectal
Lung
Prostate
HLA-DRA
ASRGL1
HLA-A
LY75
FGF7
PLG
RNF43
CTSS
TF
TGFB1
CSNK2B
AGER
ALDH2
HLA-A
PLAUR
LGALS1
MSMB
MCL1
ASRGL1
KLK3
TYRO3
Drugs approved to treat cancers
Drugs under clinical trials to treat cancers

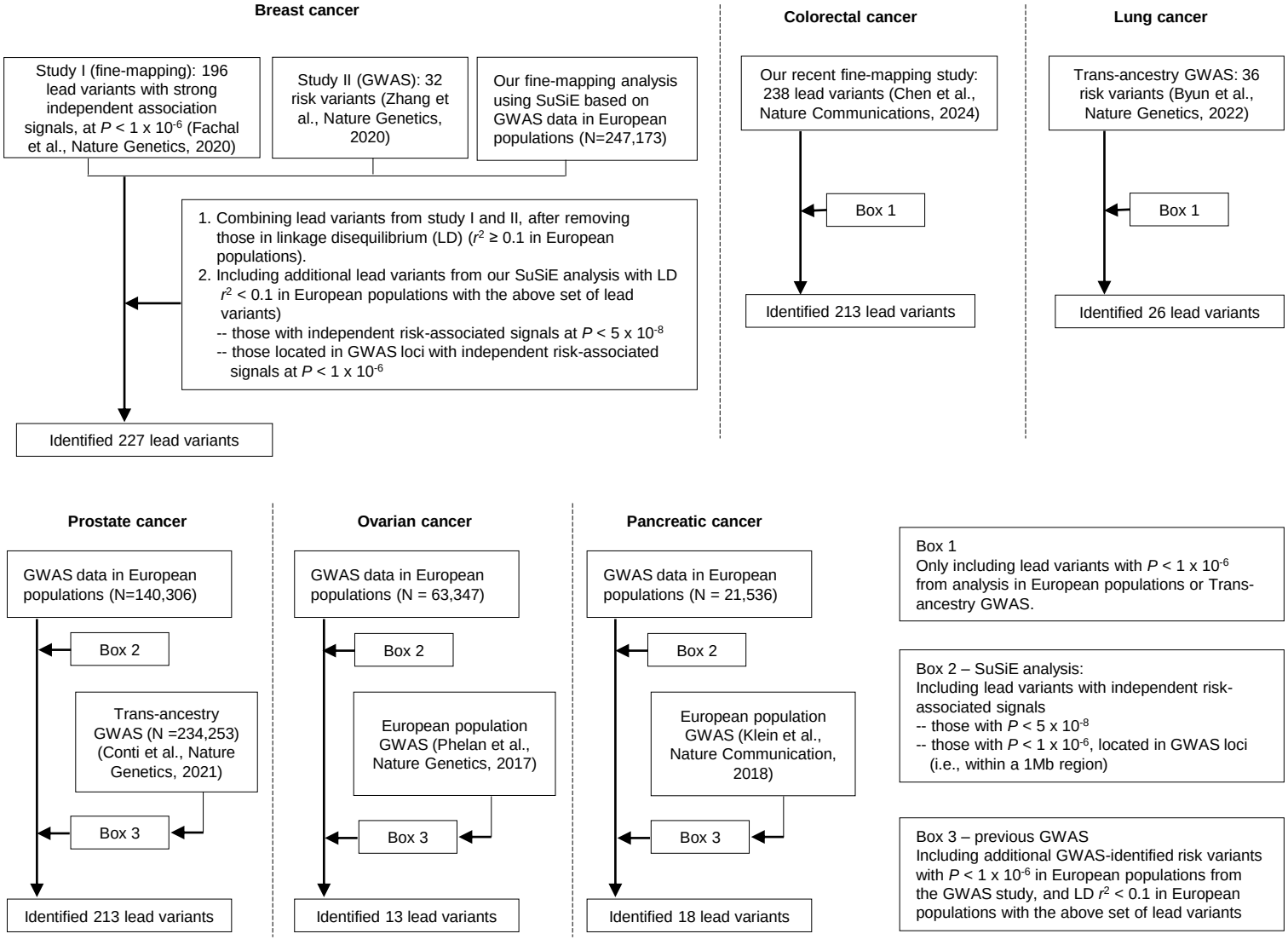
a

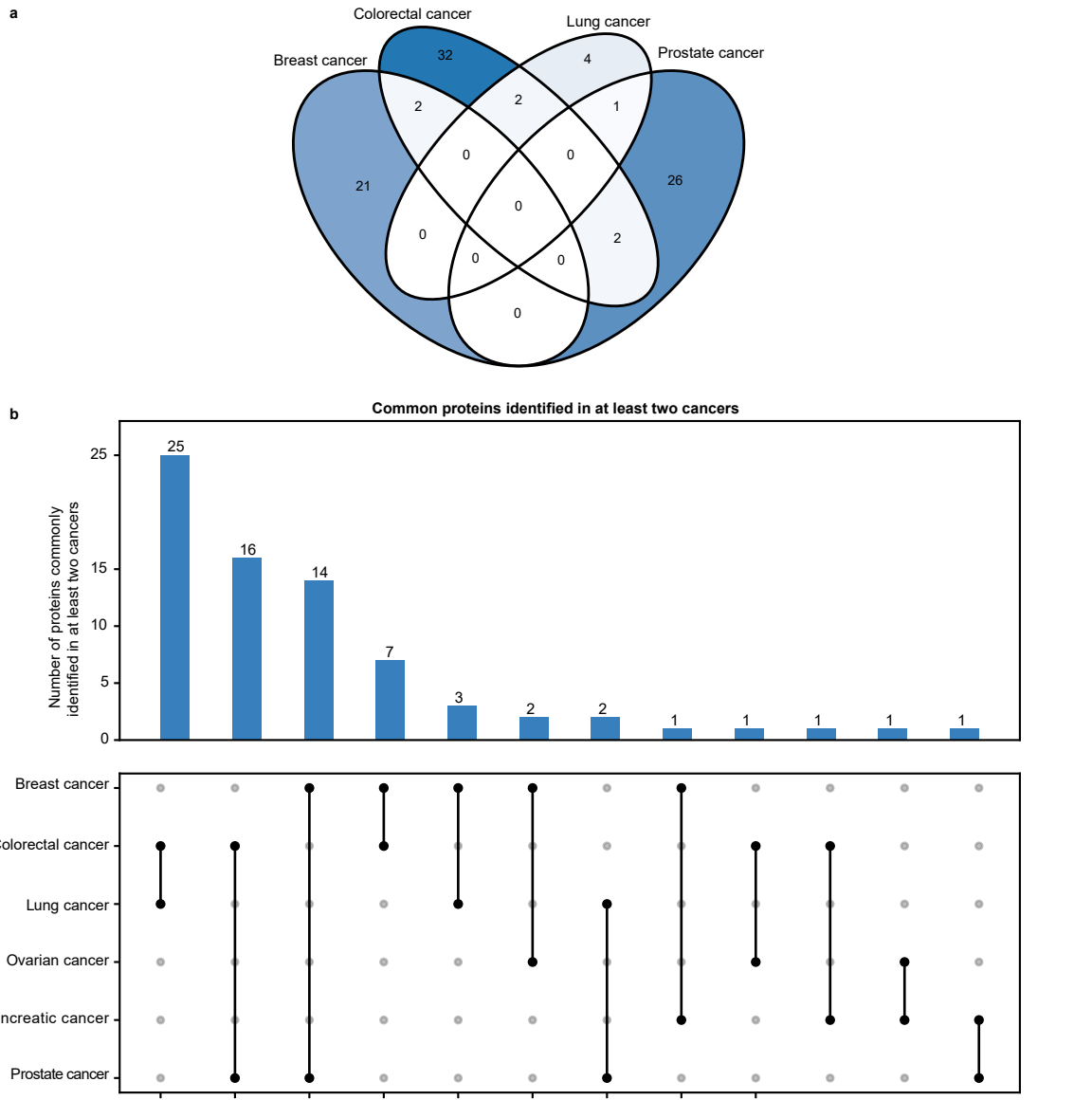
Cancer	Protein	Lead variant	Risk allele	OR (GWAS)		Beta (pQTL)		Drug
Breast	HLA-A	rs79309050	T	1.13	↑	-0.47	↓	Haloperidol
	HLA-A	rs79309050	T	1.13	↑	-0.47	↓	Trazodone
Prostate	PLG	rs9347480	C	1.06	↑	-0.12	↓	Tranexamic Acid
	TYRO3	rs11561564	G	1.06	↑	0.14	↑	Sirolimus
Colorectal	ALDH2	rs3858704	G	1.04	↑	-0.04	↓	Caffeine
	HLA-A	rs2517671	G	1.04	↑	0.83	↑	Haloperidol
	HLA-A	rs2517671	G	1.04	↑	0.83	↑	Acetazolamide
	TF	rs4854776	C	1.04	↑	-0.04	↓	Captopril

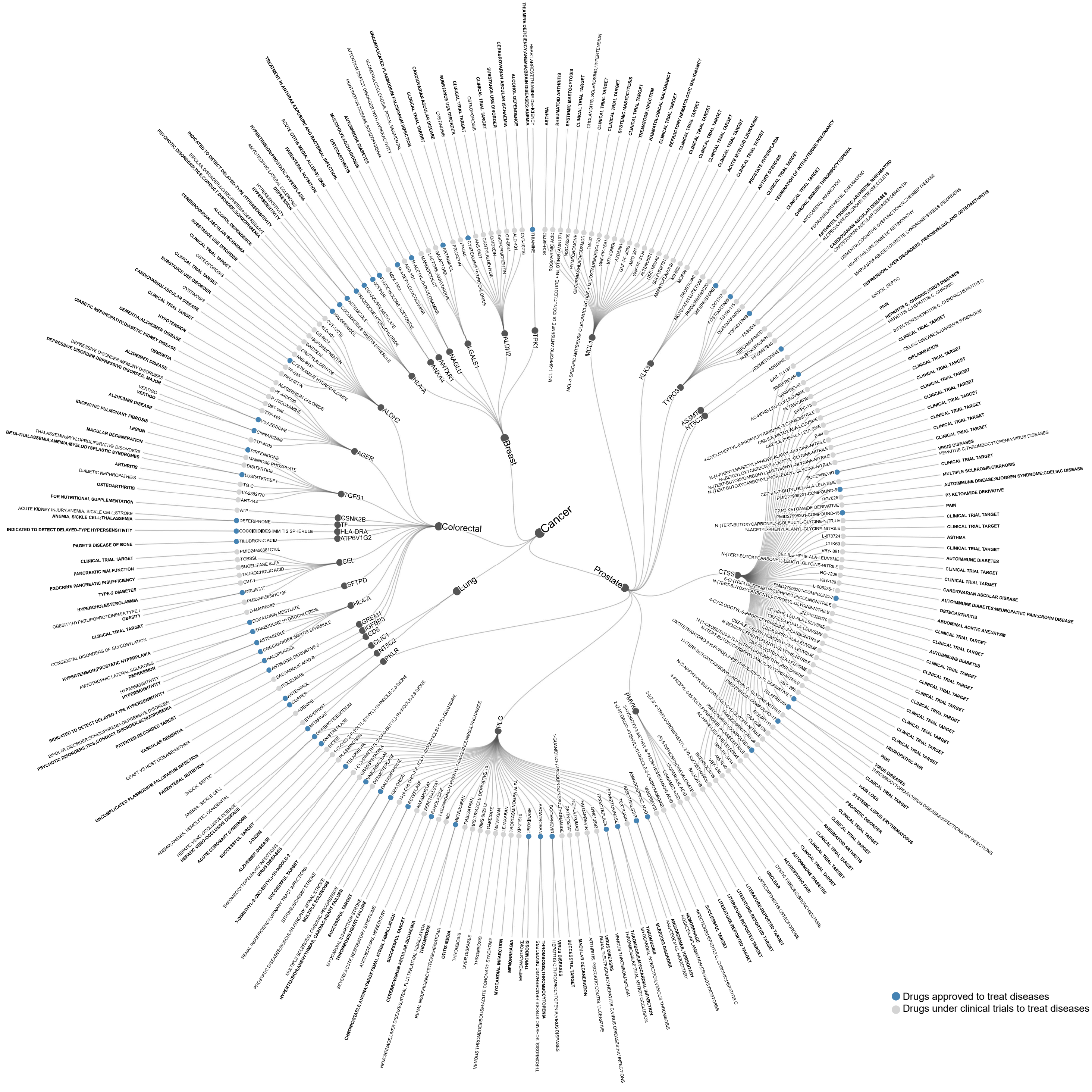







Breast cancer
Study I (fine-mapping): 196 lead variants with strong independent association signals, at P < 1 x 10-6 (Fachal et al., Nature Genetics, 2020)
Study II (GWAS): 32 risk variants (Zhang et al., Nature Genetics, 2020)
Our fine-mapping analysis using SuSiE based on GWAS data in European populations (N=247,173)
1. Combining lead variants from study I and II, after removing those in linkage disequilibrium (LD) (r2 ≥ 0.1 in European populations).
2. Including additional lead variants from our SuSiE analysis with LD r2 < 0.1 in European populations with the above set of lead variants)
-- those with independent risk-associated signals at P < 5 x 10-8
-- those located in GWAS loci with independent risk-associated signals at P < 1 x 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 x 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 x 10-8
-- those with P < 1 x 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 x 10-6 in European populations from the GWAS study, and LD r2 < 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
Breast cancer
Colorectal cancer
Lung cancer
Ovarian cancer
Pancreatic cancer
Prostate cancer


Cancer
Breast
Colorectal
Lung
Prostate
TPK1
ALDH2
LGALS1
NAGLU
ANTXR1
ANXA4
HLA-A
ALDH2
AGER
TGFB1
CSNK2B
TF
HLA-DRA
ATP6V1G2
CEL
SFTPD
HLA-A
GREM1
IGFBP3
CD6
CLIC1
NT5C2
PKLR
PLG
PMVK
CTSS
AS3MT
NT5C2
TYRO3
KLK3
MCL1
Drugs approved to treat diseases
Drugs under clinical trials to treat diseases