Identifying proteomic risk factors for cancer using prospective and exome analyses: 1,463 circulating proteins and risk of 19 cancers in the UK Biobank

Authors: Keren Papier[1]*, Joshua R Atkins[1]*, Tammy YN Tong[1], Kezia Gaitskell [1], Trishna Desai [1], Chibuzor F Ogamba [1], Mahboubeh Parsaeian [1], Gillian K Reeves [1], Ian G Mills[2,3], Tim J Key[1], Karl Smith-Byrne[1+], Ruth C Travis[1+]

[1] Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

[2] Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

[3] Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK

*Keren Papier and Joshua R Atkins are shared first authors. [+]Karl Smith-Byrne and Ruth C Travis contributed equally.

Conflict of interests: None

Running title: Plasma proteins and cancer risk

Keywords: cancer; protein, proteomics, prospective, cohort, incidence, risk, exome, genetic risk score ….

Financial support:

This work was supported by Cancer Research UK (grant numbers C8221/A29017 and C8221/A29186) to fund the centralized pooling, checking and data analysis. Tammy Tong is supported by an NDPH fellowship. Keren Papier is supported by Wellcome, Our Planet Our Health (Livestock, Environment and People—LEAP) [grant number 205212/Z/16/Z]. Trishna Desai is supported by a Cancer Research UK studentship (grant number C8221/A30904). Chibuzor F. Ogamba is supported by an NDPH studentship. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Correspondence to: Keren Papier, Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Richard Doll Building, Roosevelt Drive, Oxford OX3 7LF, UK.

Tel: +44 1865 289641

Email: keren.papier@ndph.ox.ac.uk

Conflict of interest statement:

The authors disclose no potential conflicts of interest.

Abbreviations: CI, confidence interval; *cis*-*p*QTL, *cis* protein quantitative trait locus; exGS, exome-wide genetic score; HR, hazard ratio; pLI, probable loss-of-function intolerant.

Abstract (n=390)

Background

Proteins are essential for the development and progression of cancer and for the human body's defense against tumor onset. The availability of a large panel of protein measurements and whole exome sequence data in the UK Biobank has enabled the simultaneous examination of plasma protein associations with risk across multiple cancer sites and their potential role in cancer etiology.

Methods

We investigated the associations of plasma proteins with incidence of 19 cancers and 9 cancer subsites in up to 44,645 middle-aged adults in the UK Biobank, who had measurements of 1,463 plasma proteins generated using Olink Explore Proximity Extension Assay in baseline blood samples (2006-2010). Using multivariable-adjusted Cox regression, we estimated the risk of each protein with each cancer overall and by time-to-diagnosis after correction for multiple-testing. Identified protein-cancer associations were further assessed in an analysis of cancer risk using *cis*-pQTL and exome-wide protein genetic scores (exGS) in all UK Biobank participants (n=337,543).

Results

We identified 371 proteins associated with the risk of at least one incident cancer, represented by a total of 621 protein-cancer associations. These proteins were associated with cancers of the blood (201 proteins), liver (131), kidney (51), lung (28), esophagus (22), colorectum (15), stomach (8), breast (5), prostate (3), endometrium (3), ovary (2), bladder (1), head and neck (1), and brain (1). 100 of these 621 protein-cancer associations persisted for cases diagnosed more than seven years after blood draw. Of these 621 associations, there was further support from *cis*-pQTL analyses for the etiological role of TNFRSF14 in risk of non-Hodgkin lymphoma (NHL), and from whole exome protein score (exGS) analyses for 28 other protein-cancer associations, including SRP14 and risk of leukemia. Proteins with directionally concordant evidence from long time-to-diagnosis analyses and from both *cis*-pQTL and exGS analyses were SFTPA2 for lung cancer, TNFRSF1B and CD74 for NHL, and ADAM8 for leukemia.

Conclusions

For the first time using an integrated multi-omics and cross-cancer approach, we have comprehensively assessed the plasma proteome in relation to cancer risk and identified multiple novel etiological candidates. Differences in the levels of many circulating proteins were detectable more than seven years before cancer diagnosis; while some of these are likely to be markers of early cancer processes that may inform risk stratification, and/or risk factors, concordant evidence from genetic analyses suggests that some may have a role in cancer development.

Introduction

Proteins are integral to most biological processes including many that lead to carcinogenesis, such as tissue growth and proliferation. Previous prospective studies of individual or small panels of blood proteins have identified etiological cancer proteins, such as insulin-like growth factor-I, which is a causal risk factor for breast, colorectal, and prostate cancers, and microseminoprotein-beta, which is associated with lower prostate cancer risk. [1–3] Other cancer biomarkers identified include protein markers for early detection, progression, recurrence and prognosis, for example CA-125, CEACAM5, CA19-9 and prostate-specific antigen. [4–7] However, new multiplex proteomics methods allow for the simultaneous measurement of thousands of proteins, many of which have not previously been assessed for their associations with risk across multiple cancer sites.

Identifying etiological markers of cancer risk using prospective data alone can be challenging due to the potential for confounding and other epidemiological biases. However, the abundance of many proteins in the circulation can be at least partially explained by inherited genetic variation; these genetic predictors of protein levels can be used to generate complementary evidence, with orthogonal biases, on protein-cancer associations.[8–10] Many of these genetic variants lie in a protein's cognate gene (known as *cis* protein quantitative trait loci [*cis*-pQTL]) and likely influence biological processes directly, such as by transcription or translation, and can be highly robust and specific predictors of protein concentrations.[11–13] Such genetic analyses complement traditional prospective epidemiology, and the combination of observational and genetic approaches can improve our ability to identify proteins most likely to have a causal role in cancer development and progression.[14]

Here, we use an integrated multi-omics approach combining prospective cohort and exome-variant study designs to identify proteins with a role in cancer etiology: we describe the association of 1,463 protein biomarkers quantified using the Olink platform with risk of 19 common cancers and 9 cancer subsites in 44,645 UK Biobank participants, overall and by time to diagnosis. We further assess the identified protein-cancer associations as etiological risk factors using exome *cis*-pQTL variant and exome-wide genetic score analyses (exGS).

Methods

Observational data

Study population

This study is based on data from the UK Biobank participants, a prospective cohort of 503,317 adults aged between 39 and 73, recruited between 2006 and 2010 from across the UK. The study design and rationale have been described elsewhere.[15,16] Briefly, eligible participants were those registered with the National Health Service in England, Scotland or Wales who lived within travelling distance of one of the 22 assessment centers in these regions. In total, ~5% of invited participants joined the study by attending a baseline visit, where they completed a touchscreen questionnaire, had anthropometric data and biological samples taken by trained staff, and gave informed written consent to be followed up through national record linkage. The study was approved by the National Information Governance Board for Health and Social Care and the National Health Service Northwest Multicenter Research Ethics Committee (06/MRE08/65).

Exposure and outcome assessment

Non-fasting blood samples were collected from all participants at recruitment and plasma was prepared and stored at −80°C. Protein measurements were generated using the Olink Proximity Extension Assay in 54,306 participants selected as part of the UK Biobank Pharma Proteomics Project (UKB-PPP). Samples were selected for inclusion in the UKB-PPP based on a number of factors described in detail elsewhere.[17] In brief, an initial 5,500 were pre-selected by UKB-PPP members. A further 44,502 representative participant samples were selected from the UK Biobank, stratified by age, sex, and recruitment center. The remaining samples were chosen as part of a second picking process based on a variety of criteria including membership of a COVID-19 case-control imaging study. Plasma samples were transferred to the Olink Analyses Service, Uppsala, Sweden for measurements.

Olink assay technology and analyses are described in detail elsewhere.[18] In brief, the relative abundance of 1,463 proteins was quantified using antibodies distributed across four 384-plex panels: inflammation, oncology, cardiometabolic, and neurology. Blood samples were assayed in four 384-well plates consisting of four abundance blocks for each of the four panels per 96 samples using the Olink Explore platform, which is based on proximity extension assays (PEA) that are highly sensitive and reproducible with low cross-reactivity. Relative concentrations of the 1,463 unique proteins were readout by next-generation sequencing. Measurements are expressed as normalized protein expression (NPX) values that are log-base-2 transformed. Protein values below the limit of detection (LOD) were replaced with the LOD divided by the square root of 2 and each protein was rescaled to have a mean of 0 and a standard deviation (SD) of 1.[19] Protein values were subsequently inverse rank normal transformed.

Cancer registration and death data were obtained through record linkage to national registries (NHS Digital for England and Wales using participants' NHS numbers, and NHS Central Register for Scotland using the Community Health Index). Data were available until the censoring date (December 31, 2020, in England and Wales and November 30, 2021, in Scotland) or until participants died, withdrew consent for future linkage or were reported to have left the United Kingdom. Further information on data linkage is available from https://biobank.ndph.ox.ac.uk/crystal/crystal/docs/CancerLinkage.pdf). For the observational analyses, the endpoints were defined as the first incident cancer diagnosis, or cancer first recorded in death certificate if there was no previous record of a cancer diagnosis [all coded using the 10th revision of the World Health Organization's International Statistical Classification of Diseases (ICD-10)]: head and neck (C00–14, C32), esophagus (C15), stomach (C16), colorectum (C18–20), liver (C22), pancreas (C25), lung (C34), malignant melanoma (C43), breast in women (C50), uterine (C54), ovary (C56), prostate (C61), kidney (C64–65), bladder (C67), brain (C71), thyroid (C73), and the blood cancer subgroups non-Hodgkin lymphoma (NHL; C82–85), multiple myeloma (C90), and leukemia (C91–95). The following subclassifications of these cancer groupings were also considered: oral (C00–14) and lip and oral cavity (C00–06) within head and neck cancers (C00–14, C32); adenocarcinoma of esophagus (C15, morphology codes ICD-O-3 8140–8573) within esophageal cancer (C15); colon (C18) and rectum (including rectosigmoid junction, C19–20) within colorectal cancer (C18–20); adenocarcinoma of lung (C34, morphology codes ICD-O-3 8140, 8211, 8250–8260, 8310, 8323, 8480–8490 and 8550), squamous cell carcinoma (C34, morphology codes ICD-O-3 8070-8072), small cell carcinoma (C34, morphology codes ICD-O-3 8041-8042) within lung cancer (C34); and DLBCL (C83) within NHL (C82–85). The person-years of follow-up were calculated from the date of recruitment until the date of first registration of malignant cancer, death due to cancer, death, loss or end of follow-up, or censoring date, whichever came first.

Exome-sequencing in the UK Biobank and exonic pQTL discovery

Exome-sequencing data preparation and quality control procedures in the UK Biobank have been previously described.[20] In brief, exome capture was done using the IDT xGen Exome Research Panel v1.0 that underwent 75bp paired-end Illumina sequencing on the NovaSeq 6000 platform using the S2 and S4 flow cells. BWA-MEM was used to map reads to GRCh38 with variant calling performed by DeepVariant using a 100bp buffer at each site of the custom target regions. We extracted 27,335 exome variants associated with circulating protein concentrations on the Olink Explore panel at $p < 5x10^{-8}$ reported by Dhindsa *et al.* for 50,829 UK Biobank participants.[21] The exome variants reported by Dhindsa *et al.* underwent a different pipeline using AstraZeneca's Genomics Research (CGR) bioinformatics pipeline.[21] Single Nucleotide variants (SNV) and small insertions and deletions (INDEL) were additionally annotated to SnpEFF v4.3 against Ensembl Build 38.92.[22]

Exclusion and inclusion criteria

Of the 54,306 participants who were selected for proteomic profiling as part of the UKB-PPP, 1,601 had samples that did not pass quality control. From the remaining 52,705, we further excluded 2,996 participants due to cancer diagnosis at or prior to baseline (except non-melanoma skin cancer C44), 242 who had missing information on height or weight, 2,113 who were currently using hormone replacement therapy or oral contraceptives, and 2,709

who reported having diabetes at baseline. Following these exclusions, the maximal analysis cohort included 44,645 participants (see Extended Figure 1 for participant flowchart).

Statistical analysis

Observational analyses

All analyses were conducted using Stata release 17.1 and R version 4.1.2. We estimated hazard ratios (HRs) and 95% confidence intervals (CI) for each cancer site separately using Cox proportional hazards regression models with age as the underlying time variable. Missing data in covariates were handled by assigning participants to an "unknown" category for each respective variable. The minimally adjusted models were stratified by age group at recruitment (<45, 45–49, 50–54, 55–59, 60–64, and ≥65 years) and sex where applicable and adjusted for geographical region (London, North-West, North-East, Yorkshire and Humber, West Midlands, East Midlands, South-East, South-West, Wales, and Scotland), and Townsend deprivation index (fifths, unknown). Multivariable-adjusted models were additionally adjusted for cancer-specific risk factors (see Extended Methods). Cancer specific risk factors were chosen upon review of the literature and restricted to variables available in the UK Biobank. We used an effective number of tests (ENT) correction for multiple testing, applied in a family-wise manner by cancer type. The ENT method accounts for multiple testing by applying a Bonferroni correction that determines the number of independent tests as the number of principal components needed to explain 95% of the variance in protein abundance. In this case, this was 639 independent tests.[19]

We examined protein and cancer-risk associations by time to diagnosis (diagnosed in <3 years, 3-7 years, and >7 years of follow-up) to investigate potential effects of reverse causality. We also conducted a sensitivity analysis by self-reported sex (women and men) to investigate potential sex differences for protein-cancer associations that passed multiple testing correction. We tested the heterogeneity of risk coefficients between the subgroups in each stratified analysis using inverse variance weighting, testing for statistical significance with a $\chi 2$ test with k-1 degrees of freedom, where k is the number of subgroups.

Integrating existing publicly available datasets on gene expression

To provide greater biological context for identified protein-cancer associations, we extracted single cell RNA expression from the Human Protein Atlas to describe mRNA expression in cancer-free individuals for genes that code for the identified protein markers in our main observational analyses.[23] Normalized expression levels were extracted for genes in 30 different human tissues and 82 cell-types. Gene expression specificity at the cell or tissue type level was calculated as the ratio of each gene cell-type or tissue expression to the total expression of each gene across all cell or tissue types. We subsequently grouped genes into majority expression (more than 50% of total expression in each cell or tissue type) and enriched expression (between 10% and 50% of total expression in each cell-type or tissue). For proteins with either mRNA enriched or majority expression in at least one cell or tissue type, we also mapped these to their likely candidate cell and tissue of origin where possible.

Integrating existing publicly available datasets on drug targets

We gathered information on the potential *druggability* of proteins with evidence of a cancer risk association in our main analyses by extracting information on whether a protein was the target of a known drug from the Open Targets Platform.[24] Subsequently, we filtered information from Open Targets to identify drugs that were approved and on the market by additionally cross-referencing against the ChEMBL database and other drug databases including DailyMed and the Electronic Medicines Compendium.[25–27] Proteins identified as the target of an available drug were additionally annotated with information on whether the effect of the drug would act to reduce or increase the proposed protein association with cancer risk.

Cis-pQTL and exome-wide genetic score on cancer outcomes

We further investigated protein-cancer associations identified after correction for multiple testing in observational analyses using two genetic approaches: single *cis*-pQTL risk analyses, where *cis*-pQTL were available for the protein of interest, and using an exome-wide genetic score approach. No exonic variants were identified by Dhindsa *et al.* for PREB, ING1, NPM1, PQBP1, SEPTIN9, KRT14 and ARTN and so were not considered in these analyses. In all exome-wide analyses, variants were oriented to the protein-increasing allele and exGS were calculated by summing the number of independent (clumping $r^2 < 0.01$, 10,000KB) protein-increasing alleles, weighted by betas reported in Dhindsa *et al.*, and projected in up to 337,543 European UK Biobank participants with exome-sequencing (Extended Table 2) using PLINK2.[28] We subsequently used logistic regression models to estimate the association of each genetically predicted protein with cancer risk, using both *cis*-pQTL and exGS models, for each protein-cancer association identified in observational analyses. Models were adjusted for by age, sex, and the first 10 genetic principal components of ancestry. For sex-specific cancers (breast, prostate, ovary and uterine), sex was excluded from the model. *Trans*-pQTL single variant analyses were conducted to contextualize which genes may drive protein associations with cancer risk from exGS analyses. Additionally, we annotated exGS and single variant analyses with probable loss-of-function intolerance scores (pLI) from Gnomad and used IntOGen to annotate driver genes.[29,30] In the exome analysis, conventional significance was defined as $p< 0.05$, while Bonferroni correction was used as the threshold for multiple test correction across the number of *cis*-pQTL or exGS analyzed for *cis*-pQTL or exome-wide genetic scores, respectively.

Combined evidence from prospective and genetic analyses

To enhance our understanding of a protein's likelihood of having a role in cancer etiology, we combined evidence from observational long time-to-diagnosis analyses (> 7 years between blood drawn and diagnosis), *cis*-pQTL analyses, and exGS analyses, and categorized protein-cancer associations by degree of directionally concordant support from each of these three analyses. Acknowledging that not all proteins may have *cis*-pQTL, we ranked proteins as most likely to be etiological risk factors if all three types of analyses supported an association at conventional significance, followed by long time-to-diagnosis and *cis*-pQTL analyses, then long time-to-diagnosis and exGS, exGS and cis-pQTL, and finally any one of long time-to-diagnosis, *cis*-pQTL, or exGS analyses.

RESULTS

Observational analyses

Our prospective analyses included a total of 4,921 incident malignant cancer cases with a mean follow-up of 12 years (SD 2.7). The median age at any cancer diagnosis was 66.9 years (Interquartile range (IQR) 9.9) [youngest median diagnosis was for breast cancer in women (median 64.5, IQR 12.5) and oldest for squamous cell carcinoma of the lung in women (median 71.8, IQR 9.9)]. Extended Table 1 shows the median ages at diagnosis for all cancer subsites.

Baseline characteristics of the analysis sample overall, by sex and in those who developed a malignant cancer over follow-up are shown in Table 1. Compared with the total analysis sample, participants who developed a cancer were on average older and a higher proportion of them were former or current smokers, moderate to high alcohol consumers, and had a family history of various cancers; among the women, they reported having fewer children, were younger at menarche, and a higher proportion of them were postmenopausal, had used hormone replacement therapy, and had never used the oral contraceptive pill.

From the 1,463 proteins included in our analyses, we identified an association for 371 proteins with risk of at least one cancer after correction for multiple testing, which amounted to 621 protein-cancer associations (Figure 1 & Extended Table 3). Almost half of these associations (304) were for proteins enriched (greater than 10% of total body expression) for mRNA expression in either the tissue or candidate cell of origin for the cancer indicated in our

analyses (Figure 2). For 83 of the protein-cancer associations, the proteins were majority expressed (i.e. > 50%) in either the tissue or candidate cell of origin. Many of these associations were for proteins that were associated with risk of hematological cancers with high mRNA expression in either B-cells or T-cells. However, we also identified proteins that both associated with risk for cancer and either had enriched or majority mRNA expression in the liver, lung, colorectum, kidneys, brain, stomach, esophagus, and endometrium (Figure 2).

More than half of our ENT-significant protein-cancer associations (320) were for hematological malignancies (non-Hodgkin overall (NHL) [124], diffuse large B-cell non-Hodgkin (DLBCL) [50], leukemia [87], and multiple myeloma [59]). These included the associations of TNFRSF13B and SLAMF7 with risk of multiple myeloma [HR (95%CI): 2.09 (1.96-2.24) and 3.07 (2.73-3.46), respectively], PDCD1 and TNFRSF9 with risk of NHL [1.99 (1.87-2.11) and 1.98 (1.85-2.11), respectively], and FCER2 and FCRL2 with risk of leukemia [2.12 (1.98-2.29) and 2.10 (1.95-2.26), respectively].

We also observed associations between 131 proteins and risk of liver cancer that included IGFBP7 and IGFBP3 [1.65 (1.48-1.84) and 0.46 (0.39-0.54), respectively], and 51 proteins and risk of kidney cancer, such as HAVCR1 and ESM1 [2.88 (2.55-3.24) and 1.84 (1.55-2.19)]. We identified 28 proteins associated with risk of lung cancer overall and/or at least one histological subtype that included WFDC2 and CEACAM5 [1.52 (1.39-1.67) and 1.44 (1.33-1.56)]. Although most protein-cancer associations did not differ greatly between minimally and fully adjusted models, some proteins associated with risk of lung cancer after ENT correction were attenuated by more than 50% compared with minimally adjusted models, which may imply a potential risk for residual confounding stemming from measurement error in smoking behaviors (Extended Figure 2).

Twenty-two proteins were associated with risk of esophageal cancer and/or esophageal adenocarcinoma, including REG4 and ST6GAL1 [2.02 (1.66-2.45) and 1.83 (1.53-2.19)]. We identified 15 proteins associated with colorectal, colon, and/or rectal cancer, such as AREG and GDF15 [1.30 (1.19-1.42) and 1.32 (1.20-1.45)]. Eight proteins were associated with risk of stomach cancer including ANXA10 and TFF1 [1.76 (1.53-2.03) and 1.95 (1.63-2.33)]. We found five proteins associated with risk of breast cancer, such as STC2 and CRLF1 [1.33 (1.23-1.44) and 1.31 (1.22-1.41)]. Three proteins were associated with risk of prostate cancer: GP2, TSPAN1, and FLT3LG [1.29 (1.21-1.36), 1.14 (1.09-1.18), and 0.87 (0.82-0.92)] and three were associated with endometrial cancer: CHRDL2, KLK4, and WFIKKN1 [1.42 (1.21-1.65), 1.41 (1.20-1.65), and 1.42 (1.20-1.68)]. Two proteins were associated with risk of ovarian cancer, DKK4 and WFDC2 [1.46 (1.28-1.70), 1.57 (1.26-1.96)]. We identified one protein for each of bladder [WAS, 0.54 (0.39-0.73)], brain [GFAP, 1.55 (1.31-1.86)], and head and neck cancers [TPP1, 1.33 (1.16-1.52)]. Little evidence for protein associations was observed in these data for cancers of the pancreas, thyroid, lip and oral cavity, or melanoma after correcting for multiple testing. Limited heterogeneity was observed after stratifying the protein-cancer associations by sex, however none survived multiple testing correction (Extended Table 4).

Analysis stratified by time between blood draw and diagnosis

In stratified analyses, we identified 100 of the 621 ENT significant protein-cancer associations as ENT significant in the analysis of cases diagnosed more than seven years after blood draw, representing 67 unique proteins [hematological cancers: 37, liver: 12, lung: 9, stomach: 4, breast: 3, esophagus: 2, kidney: 2, colorectum: 1] (Figure 3). Among the proteins associated with risk of hematological cancers, we identified associations with risk of multiple blood cancers for members of the fc-receptor protein [FCRL1, FCRL2, FCRL3, FCRL5, FCRLB] and TNF receptor families [TNFRSF4, TNFRSF9, TNFRSF13B, TNFRSF13C, TNFSF13B, TNFSF13]. Among the 621 ENT significant protein-cancer associations, 290 were also ENT significant in the analysis of cases diagnosed within three years of blood draw, representing 182 unique proteins [hematological cancers: 142, liver: 14, lung: 12, colorectum: 10, kidney: 5, prostate: 2, stomach: 2, bladder: 1, esophagus: 1, breast: 1, brain: 1, ovary: 1], which may indicate effects of reverse causation.

Integrating existing publicly available datasets on drug targets

We identified 38 proteins associated with the risk of at least one cancer that were also the target of a drug currently approved and available [hematological malignancies (20), liver (17), kidney cancer (7), esophageal adenocarcinoma (1), and lung cancer (1)]. Most of these proteins were the target of monoclonal antibodies (21) and small molecule inhibitors (13). The proposed action for most of these drugs would be to reduce the cancer risk as indicated in our observational analyses, i.e. the drug would inhibit a protein positively associated with cancer risk. Nine of these proteins are also the target of drugs currently indicated for the treatment of the cancers identified in our risk analyses. These include Dasatinib (EPHA2), Moxetumomab pasudotox (CD22) and Inotuzumab ozogamicin (CD22) indicated in the treatment of leukemia subtypes, Brentuximab vedotin (TNFRSF8), Polatuzumab vedotin (CD79B) and Pembrolizumab (PDCD1) indicated in the treatment of NHL subtypes including DLBCL, Elotuzumab (SLAMF7) indicated in the treatment of multiple myeloma, and Regorafenib (EPHA2, PDGFRA, FGFR2) indicated in the treatment of liver cancers (Extended Table 5).

Circulating proteins with both prospective and single *cis*-variant associations

Using 939 *cis*-pQTL, which represented 294 unique proteins, we investigated 498 of the 621 protein-cancer associations that were identified after multiple testing in the main analyses. Three *cis*-pQTL coding for higher TNFRSF14 were associated with a lower risk of NHL after correction for multiple testing ($p < 0.05/939$ tests based on *cis*-pQTL variants), 1:2559766:C:T [0.85 (0.79-0.91)]; 1:2559503:C:A, [0.85 (0.79-0.91)] and 1:2556714:A:G [0.86 (0.80-0.92)] (Figure 4). We found evidence to support the potential role of an additional 81 proteins in cancer risk as indicated by 106 protein-cancer associations at $p<0.05$ which did not meet correction for multiple testing (Extended Table 6).

Circulating proteins with both prospective and exome-score associations

We derived exGS that combined known *cis* and *trans*-pQTLs to predict circulating protein concentrations and assessed their associations with cancer risk. We were able to investigate 533 of the 621 protein-cancer associations across 324 unique proteins. After correcting for multiple testing (0.05/533 exGS tests), we identified 28 associations, including 24 for NHL, 2 for leukemia (SRP14, TREML2), 1 for both liver (KRT18) and lung (TNR) (Figure 4). The strongest association was for SRP14 with leukemia [1.22 (1.16-1.28)] followed by KRT18 for liver [1.29 (1.18-1.42)], CD1C for NHL [1.11 (1.06-1.16)] and TNR for lung [0.92 (0.89-0.95). Additionally, we found 115 conventionally significant protein-cancer associations, representing 96 unique proteins (Extended Table 7) of which 74 were directionally concordant with the results from the prospective analyses.

Integrated evidence of protein-cancer associations

We identified four proteins that were both associated with risk of cancer in main analyses and had directionally concordant, conventionally significant support from all three additional analyses, i.e. long (>7 year) time-to-diagnosis, *cis*-pQTL, and exGS analyses: SFTPA2 for lung [1.24 (1.14-1.35)], TNFRSF1B [1.28 (1.19-1.37)] and CD74 [1.68 (1.49-1.90)] for NHL and ADAM8 for leukemia [1.87 (1.69-2.06)] (Figure 5). Additionally, we found genetic and observational evidence supporting the role of 45 unique proteins in the risk of cancer which were associated with cancers of blood (22 proteins), liver (11), lung (6), kidney (5), colorectum (3), prostate (1) (Table 2).

Discussion

In this large prospective study of 1,463 proteins with the risk of up to 19 cancers, we identified 371 plasma protein markers of cancer risk, including 100 that were associated with cancer diagnosed more than seven years after blood draw and many that also had support from complementary genetic analyses, which may suggest a role in etiology. Furthermore, 182 proteins were strongly associated with diagnosis within three years, suggesting potential relevance as biomarkers for early detection.

We identified both proteins that mark common processes across cancer sites and those with associations specific to a particular cancer. The proteins associated with risk of multiple cancers included GDF15, a stress-regulated hormone that we found to be associated with an increased risk of eight cancers (liver, aerodigestive and gastrointestinal tract, and hematological malignancies), and MMP12, an enzyme expressed on macrophages that was associated with an increased risk of cancers of the stomach, colon, lung, and NHL.[31] However, the majority of protein-cancer associations were cancer-site specific (225 of the 371 proteins), and many also had majority mRNA expression on the cell or tissue of cancer origin. We note, however, that further evidence for proteins and risk of less common cancers and cancer subtypes may emerge with further follow-up in the UK Biobank or other cohorts.

We found that protein-cancer associations were most prevalent for cancers related to the blood or in tissues with a role in the maintenance of blood composition or with a high throughput of blood, such as the liver, kidneys, and lungs. Further, the smaller number of protein associations for cancers with higher incidence in this study but whose organs are not directly involved in blood composition (such as breast and prostate) may indicate a more localized effect and highlight the limitation of only measuring blood protein levels when investigating diseases in other tissues. When, in the future, stage and histological grading information becomes available for cancers within the UK Biobank or other cohorts, it may be possible to identify proteins associated with disease that has progressed beyond the primary organ that may lead to more easily measurable effects in the circulation.

Integrating prospective observational and genetic evidence for candidate etiological proteins

We found four proteins that associated with cancer that in observational long time-to-diagnosis analyses, and *cis*-pQTL and exGS analyses; CD74 and TNFRSF1B were associated with NHL, and ADAM8 and SFTPA2, were associated with risk of leukemia and lung cancer, respectively. While each of these three complementary analyses have their own specific biases, the combination of concordant support from all methods may lead to greater confidence for a role in cancer development.[10] Each of these four also appear to have notable biological plausibility. CD74, TNFRSF1B, and ADAM8 all have important roles in the immune system and have enrichment for mRNA expression on candidate cells of origin for NHL and leukemia. Similarly, SFTPA2 has a well-described role in maintaining healthy lung function and is also majority expressed in alveolar cells, which are a candidate cell of origin for multiple common subtypes of lung cancer.[32]

SRP14 was associated with the risk of leukemia in both observational and exGS analyses and was more strongly associated with risk of leukemia in people diagnosed within the first three years. SRP14 has a well-described role in protein targeting in the endoplasmic reticulum, has a high pLi, and is essential for leukemia and lymphoid malignancy cell survival, as shown using CRISPR knockout models.[33] Notably, the SRP14 exGS association was explained by a single *trans* missense variant (9:5073770:G:T) in JAK2, that leads to constitutively active JAK2, which is known to predispose to various forms of leukemia.[34–36] Given *cis*-pQTL did not support a role for SRP14 with leukemia risk, it is therefore possible that SRP14, as a biomarker of imminent leukemia diagnosis, may indicate constitutively active JAK2.

Similarly, higher FLT3LG was associated with a lower risk of prostate cancer in both observational and exGS analyses. We found that the FLT3LG exGS was largely explained by *trans*-pQTL that lie in established cancer risk genes involved in the regulation of cell division and DNA repair (CHEK2 [22:28695868:AG:A], ATM [11:108267276:T:C], and TERT [5:1293971:C:T]). For example, carriers of the CHEK2 allele previously reported to increase risk of prostate cancer had lower circulating concentrations of FLT3LG.[37,38] FLT3LG is predominantly

expressed by lymphocytes, in particular natural killer cells, and has a high pLi. It also binds to FLT3, which is expressed on dendritic cells to enhance tumor antigen presentation to facilitate anti-tumor immune responses.[39] Prostate cancer cases carrying high-risk genetic variants in DNA repair pathway genes, such as CHEK2, have a greater risk of progression and are often early onset cases with a higher mutational burden.[40,41] Heightened mutation rates in the absence of effective tumor antigen presentation/immune surveillance would form a coherent biological explanation for higher cancer risk and shorter progression times. Therefore, lower FLT3LG may serve as a potential biomarker of early cancer processes leading to diagnosis among carriers of established prostate cancer risk variants.

Together these findings highlight the need for research into the potential role of blood proteins as circulating readouts that could indicate emerging early carcinogenic processes prior to diagnosis, and that may complement existing strategies that use germline genetics to identify and monitor *at-risk* populations.

We also identified protein-cancer associations with support from genetic analyses but with a discordant direction of effect. Using *cis*-pQTL, we identified an inverse association of TNFRSF14, a gene with high pLi, with NHL risk, while observational results suggested an association with higher risk, particularly within the initial three-years of follow-up. TNFRSF14 is known to acquire loss-of-function mutations early in the development of NHL, which may suggest that it has a protective role during NHL development.[30] TNFRSF14 may therefore be overexpressed as an anti-tumor response to the presence of disease, which could explain our findings. However, current protein assay technology limitations do not enable us to distinguish between multiple proteoforms that may contain higher levels of TNFRSF14 with loss of function variants in these samples.

Previous studies of proteins and cancer risk

While there have been multiple previous case-control and cross-sectional studies of circulating proteins and cancer risk (with blood taken at or after cancer diagnosis), there are limited published prospective data. We replicate some previously reported prospective associations for proteins and risk of cancer, which may serve as reassuring confirmation for the reproducibility of findings in this fast-emerging field of multiplex proteomics. We also identified many novel findings that may be due to the prospective study design and/or the large sample size. For example, we replicated the association of CDCP1 with lung cancer risk reported within the EPIC cohort, and also found concordant evidence for risk proteins, such as CEACAM5, identified within up to three years prior to diagnosis in the INTEGRAL project.[19,37,38] We additionally identified risk associations with lung cancer for multiple proteins that were either not previously investigated or that did not meet the significance criteria for multiple testing within previous studies. For colorectal cancer, we were not able to replicate the previously reported associations for several proteins identified in prospective studies using samples taken up to three years prior to diagnosis or in those studies with relatively modest numbers of incident cases ($n \leq 100$).[42,43] We also did not replicate protein risk associations previously reported for pancreatic cancer.[44] Nonetheless, our findings are in-line with some of those reported in a cross-cancer case-control study (with bloods collected at or after diagnosis) within the Uppsala-Umeå Comprehensive Cancer Consortium biobank; we replicated the reported association of GFAP with glioma and the associations of CNTN5, SLAMF7, MZB1, QPCT and TNFRSF13B with multiple myeloma.[45]

Our study has several notable strengths. We examined the role of over one thousand blood proteins in cancer development and report several hundred novel protein and cancer associations. The detailed information in the UK Biobank on a wide range of cohort characteristics (including cohort-wide exome data) has made it possible to assess the potential for cancer-specific confounders to influence our findings and to run complementary genetic analyses on the majority of candidate proteins identified in our observational analyses. Further, information on cancer diagnosis was obtained from data linkage, thus minimizing selective dropouts. The cross-cancer approach also reduced outcome selection bias and enabled us to find proteins associated with both multiple and specific cancers, and their subtypes.

Furthermore, the UK Biobank is a mature prospective cohort, which allowed us to assess whether protein-cancer associations were being driven by altered protein levels in individuals who were likely to have preclinical disease at blood draw and/or persisted with longer follow-up. Nonetheless, some hematological cancers can be present long before clinical diagnosis, such as chronic lymphocytic leukemia.[46,47] Further, liver and kidney disease both have risk factors, including cirrhosis and chronic kidney disease, respectively, that we may expect to perturb the blood proteome far in advance of diagnosis. It is therefore possible that associations with risk observed more than seven years prior to diagnosis may still be due to either reverse causality or be markers of established risk factors and not etiological. However, proteins associated with cancer risk long before diagnosis and that have support from complementary genetic analyses may warrant follow-up as potential cancer risk factors.

We also note that we only analyzed protein concentrations measured at baseline and therefore were not able to address potential regression dilution bias, which may have led to underestimates of relative risks. Also, while this is the largest cohort study of plasma proteins and cancer to date, we had relatively limited power to detect protein-cancer associations for less common cancer sites and subsites that nonetheless hold substantial public health importance. Finally, the UK Biobank predominantly consists of adults of White ethnicity and who have a more favorable risk profile compared to the national UK population.[15] Proteomics holds significant promise for developing future cancer prevention initiatives that are needed to address the predicted increase in cancer burden among diverse populations, and so further studies into the proteomics of cancer risk in non-White populations are necessary.[48]

There are several research priorities leading from our findings that are necessary to pursue to more fully understand the roles of proteins in cancer development and progression. Priorities are more large-scale prospective data from mature cohorts to replicate our findings and further complementary genetic studies, including Mendelian randomization analyses. As new GWAS data for cancers of the blood, liver, and kidney become available, further investigations into etiology using genetic epidemiology will be possible. Where protein associations prove replicable, it will be necessary to better understand their role at the tissue and cellular level. This is of particular interest given proteins are the target of 98% of all drugs and that 38 of our candidate etiological proteins are the target of existing drugs, of which nine had further directionally concordant evidence from genetic analyses supporting their role in cancer development .[49] Nonetheless, substantial additional research would be needed to assess any potential for therapeutic prevention, including functional and experimental studies, and those to assess potential toxicity.

In conclusion, we discovered multiple associations between blood proteins and cancer risk. Many of these were detectable more than seven years before cancer diagnosis and had concordant evidence from genetic analyses, suggesting they may have a role in cancer development. We also identified proteins that may mark early cancer processes among carriers of established cancer risk variants, which may serve as potential biomarkers for risk stratification and early diagnosis.

List of tables and figures

Main

comparison, with the yellow dash line representing the conventional significance threshold. Odds Ratios (OR) are the relative risk per standard deviation increase. *Cis*-variants were adjusted to be on the same scale.

Figure 5. – Forrest plots for the prospective and genetic associations of SFPTA2 with lung cancer risk, CD74 and TNFRSF1B with risk of non-Hodgkin lymphoma, and ADAM8 with risk of leukemia

Forrest plots display the association of each of CD74 and TNFRSF1B with risk of non-Hodgkin lymphoma, and ADAM8 and SFTPA2, with risk of leukemia and lung cancer, respectively. For each protein-cancer association evidence for the association of concentrations with cancer risk is presented from minimally and fully adjusted models per SD, as well as models stratified by time-to-diagnosis, and from exome proteins score and *cis*-pQTL analyses.

Extended Methods and Figures

Extended figure 1 – Study design flow chart and results summary

Extended figure 2 – Percentage change in the log hazard ratios between fully and minimally adjusted models

Extended Tables

Extended Table 1 - Number of cases and age of diagnosis by cancer sub-site in all participants, and in men and women

Extended Table 2 - UK Biobank exome sequencing cohort numbers per cancer site

Extended Table 3 - All protein-cancer associations from minimally adjusted, multivariable-adjusted and by time-to-diagnosis analyses

Extended Table 4 - Protein-cancer associations that pass multiple testing correction stratified by sex

Extended Table 5 - Protein-cancer associations that pass multiple test correction mapped to currently available drugs

Extended Table 6 - Cis and trans single variant affecting protein-cancer associated proteins on cancer risk and exome score weights

Extended Table 7 - Association of exome score predicted protein concentration with cancer risk for protein-cancer associations that passed multiple testing correction

Table 1. Baseline characteristics of the UK Biobank analysis cohort, overall, by sex, and in those who developed any malignant cancer

Characteristics	All (N=44,645)	Women (n=23,274)	Men (n=21,371)	Developed a malignant cancer (n=4,921)
Sociodemographic				
Age (years)	57.0 (8.3)	57.0 (8.1)	57.1 (8.4)	60.6 (7.0)
Townsend deprivation, *n* (%)				
Most affluent	8,954 (20.1%)	4,598 (19.8%)	4,356 (20.4%)	992 (20.2%)
Most deprived	9,416 (21.1%)	4,797 (20.6%)	4,619 (21.6%)	1,066 (21.7%)
Unknown	53 (0.1%)	21 (0.1%)	32 (0.1%)	4 (0.1%)
Lifestyle				
Physical activity level, *n* (%)				
Low <10 METs	8,430 (18.9%)	4,205 (18.1%)	4,225 (19.8%)	959 (19.5%)
10 to <50 METs	18,281 (40.9%)	9,352 (40.2%)	8,929 (41.8%)	1,946 (39.5%)
High ≥50 METs	7,927 (17.8%)	3,644 (15.7%)	4,283 (20.0%)	852 (17.3%)
Unknown	10,007 (22.4%)	6,073 (26.1%)	3,934 (18.4%)	1,164 (23.7%)
Smoking, *n* (%)				
Never	24,481 (54.8%)	13,980 (60.1%)	10,501 (49.1%)	2,220 (45.1%)
Former	15,248 (34.2%)	7,158 (30.8%)	8,090 (37.9%)	2,007 (40.8%)
Current <15 cigarettes/day	1,368 (3.1%)	751 (3.2%)	617 (2.9%)	180 (3.7%)
Current ≥15 cigarettes/day	1,818 (4.1%)	752 (3.2%)	1,066 (5.0%)	315 (6.4%)
Current, amount unknown	1,502 (3.4%)	521 (2.2%)	981 (4.6%)	177 (3.6%)
Unknown	228 (0.5%)	112 (0.5%)	116 (0.5%)	22 (0.4%)
Alcohol intake, *n* (%)				
non-drinkers	3,586 (8.0%)	2,240 (9.6%)	1,346 (6.3%)	359 (7.3%)
<1g/day	4,821 (10.8%)	3,483 (15.0%)	1,338 (6.3%)	509 (10.3%)
1-9g/day	13,781 (30.9%)	9,129 (39.2%)	4,652 (21.8%)	1,368 (27.8%)
10-19g/day	9,594 (21.5%)	5,010 (21.5%)	4,584 (21.4%)	1,077 (21.9%)
≥20 g/day	12,563 (28.1%)	3,230 (13.9%)	9,333 (43.7%)	1,578 (32.1%)
Unknown	300 (0.7%)	182 (0.8%)	118 (0.6%)	30 (0.6%)
Anthropometric				
Standing height in cm	168.7 (9.3)	162.4 (6.4)	175.6 (6.9)	169.6 (9.1)
Body mass index (kg/m^2)	27.3 (4.6)	27.0 (5.1)	27.6 (4.0)	27.6 (4.6)
Family history of breast cancer, n (%)	3,251 (7.3%)	1,754 (7.5%)	1,497 (7.0%)	374 (7.6%)
Family history of prostate cancer, n (%)	2,887 (6.5%)	1,531 (6.6%)	1,356 (6.3%)	369 (7.5%)

Family history of lung cancer, n (%)	4,971 (11.1%)	2,597 (11.2%)	2,374 (11.1%)	602 (12.2%)
Family history of colorectal cancer, n (%)	4,162 (9.3%)	2,140 (9.2%)	2,022 (9.5%)	536 (10.9%)
Women's health				
Parity in women, *n* (%)				
Nulliparous	-	4,257 (9.5)	-	416 (19.1%)
1-2 births	-	13196 (29.6)	-	1225 (56.3)
>3 births	-	5768 (47.9)	-	530 (24.3)
Unknown	-	53 (0.1)	-	5 (0.2)
Age at first menarche in women, *n* (%)				
<12 years	-	4,467 (19.2%)	-	441 (20.3%)
Unknown	-	750 (3.2%)	-	66 (3.0%)
Menopausal status in women, n (%)				
Premenopausal	-	5,564 (23.9%)	-	356 (16.4%)
Postmenopausal	-	16,580 (71.2%)	-	1,760 (80.9%)
Unknown	-	1,130 (4.9%)	-	60 (2.8%)
Hormone replacement therapy use in women, *n* (%)				
Never	-	15,036 (64.6%)	-	1,256 (57.7%)
Past	-	8,102 (34.8%)	-	910 (41.8%)
Unknown	-	136 (0.6%)	-	10 (0.5%)
Oral contraceptive pill use in women, *n* (%)				
Never	-	4683 (20.1)	-	490 (22.5)
Past	-	18481 (79.4)	-	1674 (76.9)
Unknown	-	110 (0.5)	-	12 (0.6)

Values are presented as mean (standard deviation) unless otherwise specified.

Table 2. Summary of protein-cancer associations* that have support from one or more of long lagtime, *cis*-pQTL, or exome protein score analyses

Evidence/Cancer site	n/N**	Directionally concordant support from:						
Lagtime > 7 yrs								
cis-pQTL								
exGS								
Head and neck (overall, oral, lip and oral cavity)	1/1						**TPP1**	
Oesophagus (overall, oesophageal adenocarcinoma)	13/22						**ANG**,**CCL14**,CCL22,**EGFL7**,EPS8L2,**FABP1**,PIGR, **REG3A**,SPINK1, ST6GAL1,TFF2,TNFRSF10B	**RARRES2**
Stomach	6/8						**ANXA10**,CXCL17,GDF15,GGH,TFF1,TFF2	
Colorectum (overall, colon, rectal)	10/15			AREG, **RBP2**, **SPINK4**		MMP12	**KRT19**, **PDGFC**,**PREB**,REG4,TFF2	**AGR2**
Liver	81/131		ANGPT2[D], CD74,CXADR, EPHA2[D], PIGR	**CDH2**,**CHI3L1**, **KRT18**,SIGLEC1,**SPON2**,**SULT2A1**			**ACE2**,**ACP5**,**ACY1**,ADA2,ADGRE2,**ADGRG1**,**BST2**,**C19ORF12**,**CCL15**,**CD163**,CDCP1,CDH6,**CDHR2**,**CLSTN2**,**CNDP1**,**COL4A1**[D],CTSD,**CTSL**,**DDR1**,DPP10,**EFEMP1**,**ENG**,**ENPP2**,**ERBB2**[D],**FGFR2**[B],FSTL3,FUT3_FUT5,**GGT1**,GRN,HAVCR1,**IGFBP3**,IGFBP7,IL10RB,IL18BP,**IL18R1**,IL4R[D],**IL6ST**[D],**ITGA5**,**ITGB2**[D],ITGB7[D],**KRT14**,LAG3[D],LGALS9,**LTBP2**,**MME**[D],**MSR1**,NFASC,**NOMO1**,**NRCAM**,**NRP2**,**NT5E**,**PCDH17**,**PDGFRA**[B],PGF[D],**PLXNB2**,**PVR**,SDC1,SEMA7A,**SEZ6L2**,SLAMF1,SPINT1,SPP1, **TGFBI**,TIMP1,**TNFRSF11B**,**TNFRSF21**,TNFSF13B[D],VCAM1,**VTCN1**	ALCAM
Lung (overall, squamous, small cell, adenocarcinoma)	18/28	**SFTPA2**	PIGR	**ALPP**, **MMP9**, PLAUR, **TNR**			**BAMBI**, CDCP1, CXCL17,CEACAM5,**CLEC5A**,FUT3_FUT5,**ITIH3**,**MSLN**,**PRSS8**,**SPARCL1**,TNFRSF10B	MMP12
Breast	5/5						**ANGPTL4**,**CRLF1**,**GAL**, AREG	**STC2**
Prostate	2/3			**FLT3LG**			**GP2**	
Kidney	30/51		HAVCR1	**CLEC14A**, CXADR, **FOLR1**, **IGFBP6**			BTN2A1,**CA12**[D],**CD38**[D],**CD300C**,**CD302**,CRIM1, **EPHB4**[D],ESM1,**IFNGR1**[I],**JAM**,**LAYN**,**LRP11**,**MMP7**[D], **NBL1**,**NECTIN4**[D],**RTN4R**,**TGFBR2**,**THBD**,**TNFRSF1A**, TNFRSF14,**TNFRSF19**,FSTL3,HYOU1,KLRB1,LGALS9	
Bladder	1/1						**WAS**	

Brain	1/1					**GFAP**		
Non-Hodgkin lymphoma (overall, diffuse lymphoma)	60/125	CD74, TNFRSF1B	FCRL3, SLAMF8		TNFRSF1B ,BTN2A1	**LRIG1, SEMA4D**	CD6,TNFRSF10A,ADGRE5,**CCL21**,CCL22,CD22^{D},CD 27,CD28,CD48,CD70,CD79B^{D},**CEACAM21**,CRTAM, **CXCL13**,**DCTPP1**,EFNA4,FCRL2,FCRLB,**GALNT3**,**GC NT1**,**HLA_E**,IFNLR1,IL12RB1,IL4R^{D},**JCHAIN**,KLRB1,**L AIR2**,LILRB1, LTA,LY9,**LYPD8**,**MARCO**,MILR1,MMP12,**NOS1**,**NO S3**,**RBP5**, SEMA7A,**SERPINA9**,**SH2D1A**,SIGLEC10,SLAMF6,**T NF**,TNFRSF13C,	CSF1, **CXCL9,** IL2RA^{B},IL1 2A_IL12B, IL12B^{D}, IL18BP, SIGLEC1, VCAM1
Multiple myeloma	36/59		CD48, TNFRSF13, IGFBP7, LY9,**TFPI2**,T NFRSF10	CD79B^{D}, TNFSF13B^{I}			**BMP6**,CD274^{D},**CNTN5**,**CRELD2**,FCRL5,**FRZB**,**GDNF**, **GOLM2**,**IDS**,**IL5RA**^{D},**IL6R**,IL10RB,**MDK**,**PCOLCE**,**PR ELP**,**PTPRS**,RNASET2,**SLAMF7**^{D},TNFSF13,**ARSA**, FCRL2, FCRLB, HYOU1, ICAM3, IFNLR1, SDC1, ST6GAL1	**ITM2A**
Leukaemia	48/87	ADAM8	CD22^{D}, DSC2	ADGRE5, CD74, CD83, IL18BP, PDCD1^{D}, SEMA7A, TNFRSF1B, **TREML2**,		FCRL5, **PIK3AP1**	CD6,LTA,**B4GALT1**,CD200,**CD200R1**,CD27,CD70,C D79B^{D},CRTAM,EFNA4,**EZR**,FCRL1,FCRL3, GRN,ICAM3,IGSF3,IL2RA^{B},LY9,**PARP1**, **ROR1**,**SDC4**,SIGLEC10,SIGLEC6,SLAMF6,**TCL1A**,TN FRSF13B,TNFRSF13C,TNFRSF4,TNFRSF8,TNFRSF9	**APEX1,** EPHA2^{D}, IL4R^{D}, **PSIP1, SRP14**

*protein-cancer associations in the prospective cohort analyses that were significant after corrections for the effective number of tests.

**n/N represents the number of protein-cancer associations that have support from one or more of long lagtime, *cis*-pQTL, or exome protein score analyses/ the total protein-cancer site associations

For protein-cancer associations with >1 association with a cancer and i's subsites we only present the one with the highest tier

Bold proteins represent proteins-cancer associations that are specific to one cancer site after corrections for the effective number of tests.

Proteins that are underlined are targets of approved drugs with ^{I} indicating the action of its approved drug increases the risk of the corresponding cancer and ^{D} indicating the action of its approved drug decreases the risk of the corresponding cancer ^{B} indicating the presence of approved drugs with both an increased and decreased risk.

References

1. Knuppel, A. *et al.* Circulating Insulin-like Growth Factor-I Concentrations and Risk of 30 Cancers: Prospective Analyses in UK Biobank. *Cancer Res.* **80**, 4014–4021 (2020).
2. Watts, E. L. *et al.* Circulating insulin-like growth factors and risks of overall, aggressive and early-onset prostate cancer: a collaborative analysis of 20 prospective studies and Mendelian randomization analysis. *Int. J. Epidemiol.* **52**, 71–86 (2023).
3. Smith Byrne, K. *et al.* The role of plasma microseminoprotein-beta in prostate cancer: an observational nested case-control and Mendelian randomization study in the European prospective investigation into cancer and nutrition. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **30**, 983–989 (2019).
4. Menon, U. *et al.* Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *The Lancet* **397**, 2182–2193 (2021).
5. Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer. Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. *JAMA Oncol.* **4**, e182078 (2018).
6. Ballehaninna, U. K. & Chamberlain, R. S. The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal. *J. Gastrointest. Oncol.* **3**, 105–119 (2012).
7. Lane, J. A. *et al.* Latest results from the UK trials evaluating prostate cancer screening and treatment: the CAP and ProtecT studies. *Eur. J. Cancer Oxf. Engl. 1990* **46**, 3095–3101 (2010).
8. Johansson, Å. *et al.* Identification of genetic variants influencing the human plasma proteome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4673–4678 (2013).
9. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

10. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).
11. DeBoever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
12. Gkatzionis, A., Burgess, S. & Newcombe, P. J. Statistical methods for cis-Mendelian randomization with two-sample summary-level data. *Genet. Epidemiol.* **47**, 3–25 (2023).
13. Zhong, W. *et al.* Whole-genome sequence association analysis of blood proteins in a longitudinal wellness cohort. *Genome Med.* **12**, 53 (2020).
14. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* dyw314 (2017) doi:10.1093/ije/dyw314.
15. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
16. Collins, R. What makes UK Biobank special? *The Lancet* **379**, 1173–1174 (2012).
17. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. 2022.06.17.496443 Preprint at https://doi.org/10.1101/2022.06.17.496443 (2022).
18. Wik, L. *et al.* Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol. Cell. Proteomics MCP* **20**, 100168 (2021).
19. Albanes, D. *et al.* The blood proteome of imminent lung cancer diagnosis. *Nat. Commun.* **14**, 3042 (2023).
20. Protocol for Processing UKB Whole Exome Sequencing Data Sets. https://dnanexus.gitbook.io/uk-biobank-rap/science-corner/whole-exome-sequencing-oqfe-protocol/protocol-for-processing-ukb-whole-exome-sequencing-data-sets.

21. Dhindsa, R. S. *et al.* Influences of rare protein-coding genetic variants on the human plasma proteome in 50,829 UK Biobank participants. 2022.10.09.511476 Preprint at https://doi.org/10.1101/2022.10.09.511476 (2022).
22. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
23. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci. Publ. Protein Soc.* **27**, 233–244 (2018).
24. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).
25. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
26. DailyMed. https://www.dailymed.nlm.nih.gov/dailymed/.
27. Electronic medicines compendium (emc). https://www.medicines.org.uk/emc#gref.
28. Chen, Z.-L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
29. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
30. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
31. Kim, K. H. & Lee, M.-S. GDF15 as a central mediator for integrated stress response and a promising therapeutic molecule for metabolic disorders and NASH. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1865**, 129834 (2021).
32. Floros, J., Thorenoor, N., Tsotakos, N. & Phelps, D. S. Human Surfactant Protein SP-A1 and SP-A2 Variants Differentially Affect the Alveolar Microenvironment, Surfactant Structure, Regulation and

Function of the Alveolar Macrophage, and Animal and Human Survival Under Various Conditions. *Front. Immunol.* **12**, (2021).

33. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).
34. Beer, P. A. *et al.* Two routes to leukemic transformation after a JAK2 mutation–positive myeloproliferative neoplasm. *Blood* **115**, 2891–2900 (2010).
35. Gnanasambandan, K., Magis, A. & Sayeski, P. P. The Constitutive Activation of Jak2-V617F is Mediated by a π Stacking Mechanism Involving Phe 595 and Phe 617. *Biochemistry* **49**, 9972–9984 (2010).
36. Benton, C. B. *et al.* Janus kinase 2 variants associated with the transformation of myeloproliferative neoplasms into acute myeloid leukemia. *Cancer* **125**, 1855–1866 (2019).
37. Dagnino, S. *et al.* Prospective Identification of Elevated Circulating CDCP1 in Patients Years before Onset of Lung Cancer. *Cancer Res.* **81**, 3738–3748 (2021).
38. Robbins, H. A. *et al.* Design and methodological considerations for biomarker discovery and validation in the Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Program. *Ann. Epidemiol.* **77**, 1–12 (2023).
39. Wilson, K. R., Villadangos, J. A. & Mintern, J. D. Dendritic cell Flt3 – regulation, roles and repercussions for immunotherapy. *Immunol. Cell Biol.* **99**, 962–971 (2021).
40. Herberts, C., Wyatt, A. W., Nguyen, P. L. & Cheng, H. H. Genetic and Genomic Testing for Prostate Cancer: Beyond DNA Repair. *Am. Soc. Clin. Oncol. Educ. Book* e390384 (2023) doi:10.1200/EDBK_390384.
41. Gerhauser, C. *et al.* Molecular evolution of early onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996-1011.e8 (2018).
42. Sun, X. *et al.* Prospective Proteomic Study Identifies Potential Circulating Protein Biomarkers for Colorectal Cancer Risk. *Cancers* **14**, 3261 (2022).

43. Harlid, S., Myte, R. & Van Guelpen, B. The Metabolic Syndrome, Inflammation, and Colorectal Cancer Risk: An Evaluation of Large Panels of Plasma Protein Markers Using Repeated, Prediagnostic Samples. *Mediators Inflamm.* **2017**, 4803156 (2017).

44. Kartsonaki, C. *et al.* Circulating proteins and risk of pancreatic cancer: a case-subcohort study among Chinese adults. *Int. J. Epidemiol.* **51**, 817–829 (2022).

45. Álvez, M. B. *et al.* Next generation pan-cancer blood proteome profiling using proximity extension assay. *Nat. Commun.* **14**, 4308 (2023).

46. Kolijn, P. M. *et al.* High-risk subtypes of chronic lymphocytic leukemia are detectable as early as 16 years prior to diagnosis. *Blood* **139**, 1557–1563 (2022).

47. Kaaks, R. *et al.* Lag times between lymphoproliferative disorder and clinical diagnosis of chronic lymphocytic leukemia: a prospective analysis using plasma soluble CD23. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **24**, 538–545 (2015).

48. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

49. Langenberg, C., Hingorani, A. D. & Whitty, C. J. M. Biological and functional multimorbidity—from mechanisms to management. *Nat. Med.* **29**, 1649–1657 (2023).

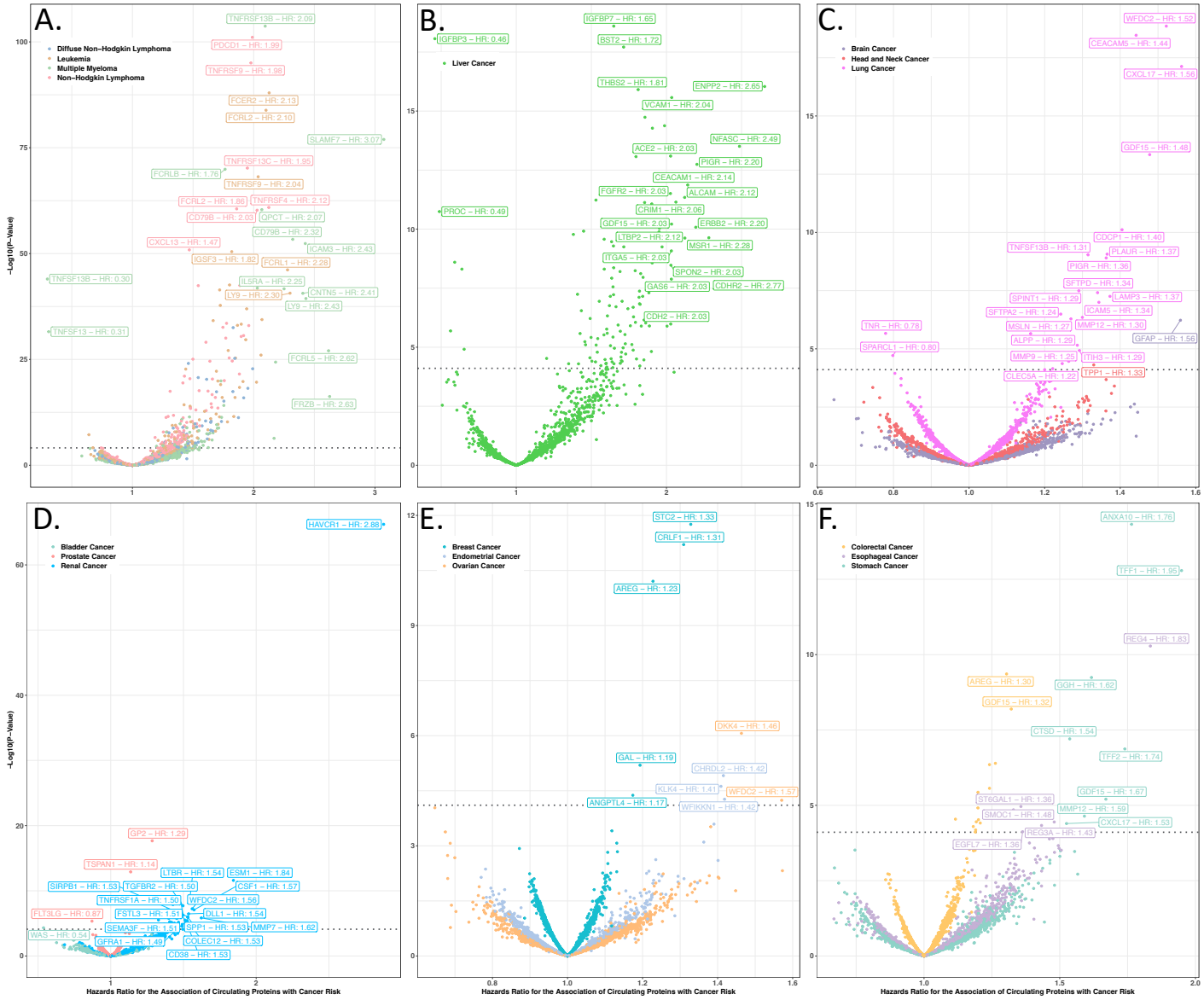

Figure 1 – Volcano plots for the prospective association of circulating proteins with risk of cancer

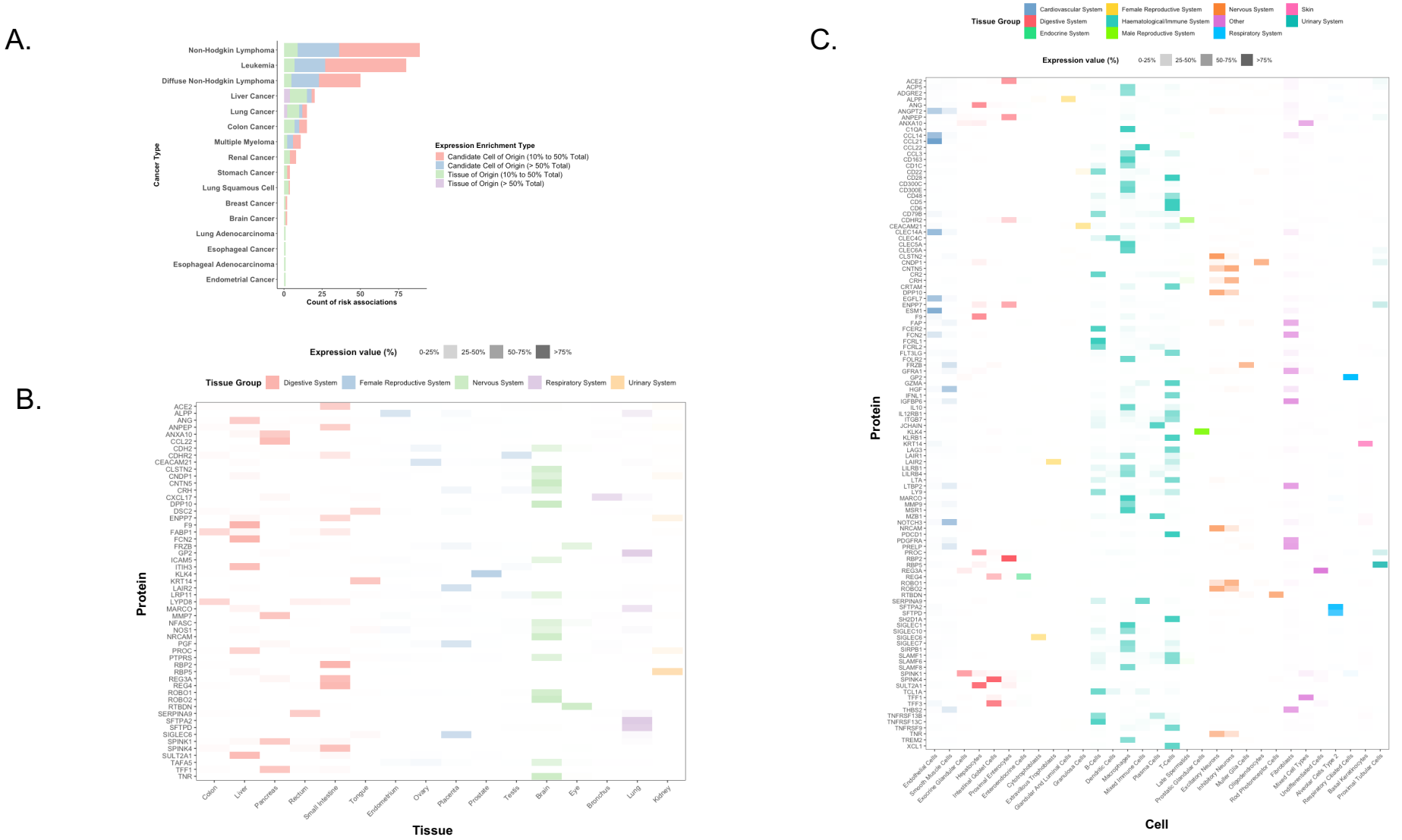

Figure 2 – Evidence for cellular and tissue enrichment of mRNA expression for cancer risk proteins

Figure 3 – Volcano plots for the prospective association of circulating proteins with risk of cancer by time to diagnosis

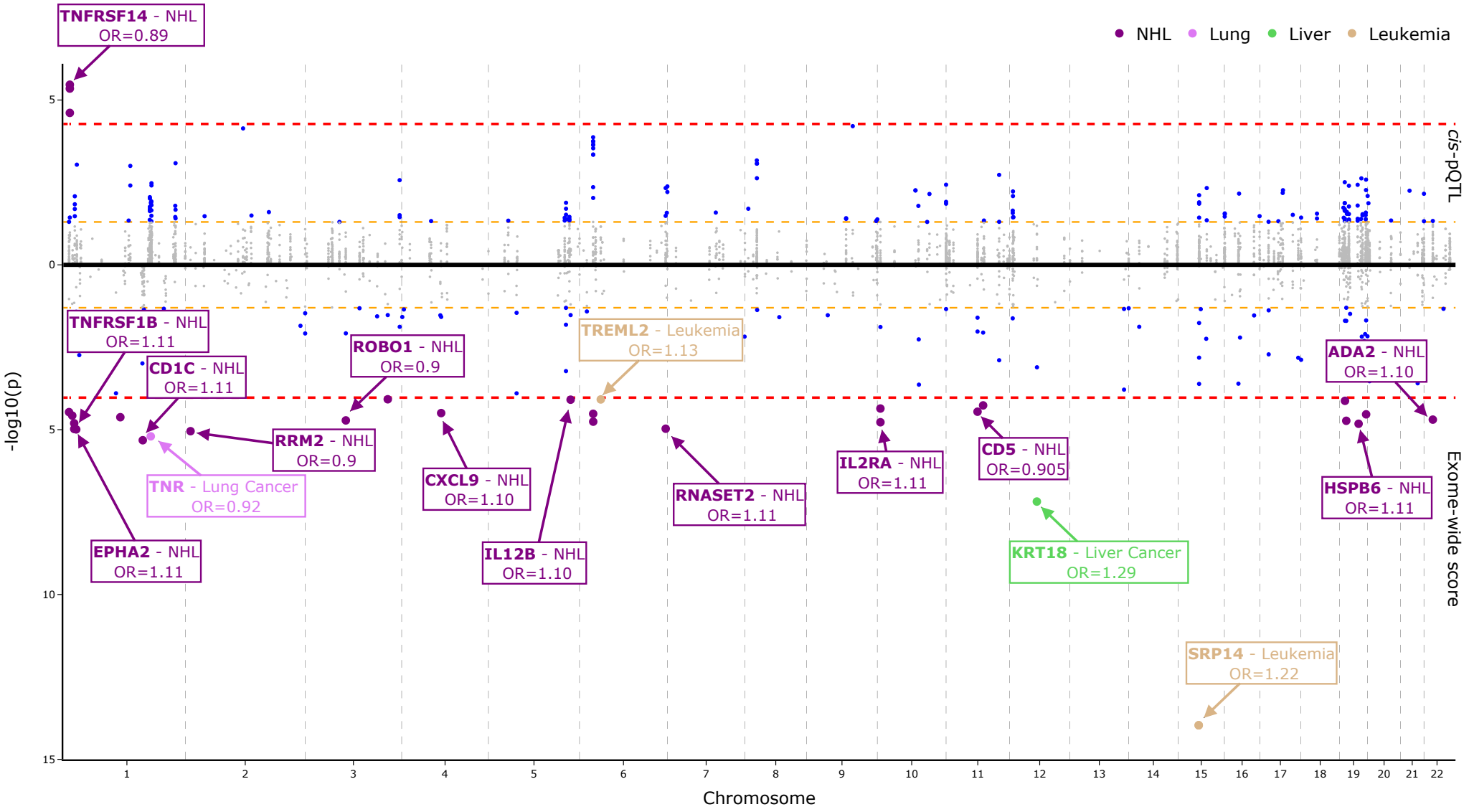

Figure 4 – Mirror Manhattan plot for the association of genetically predicted protein concentrations and cancer risk using cis-pQTL and exome scores

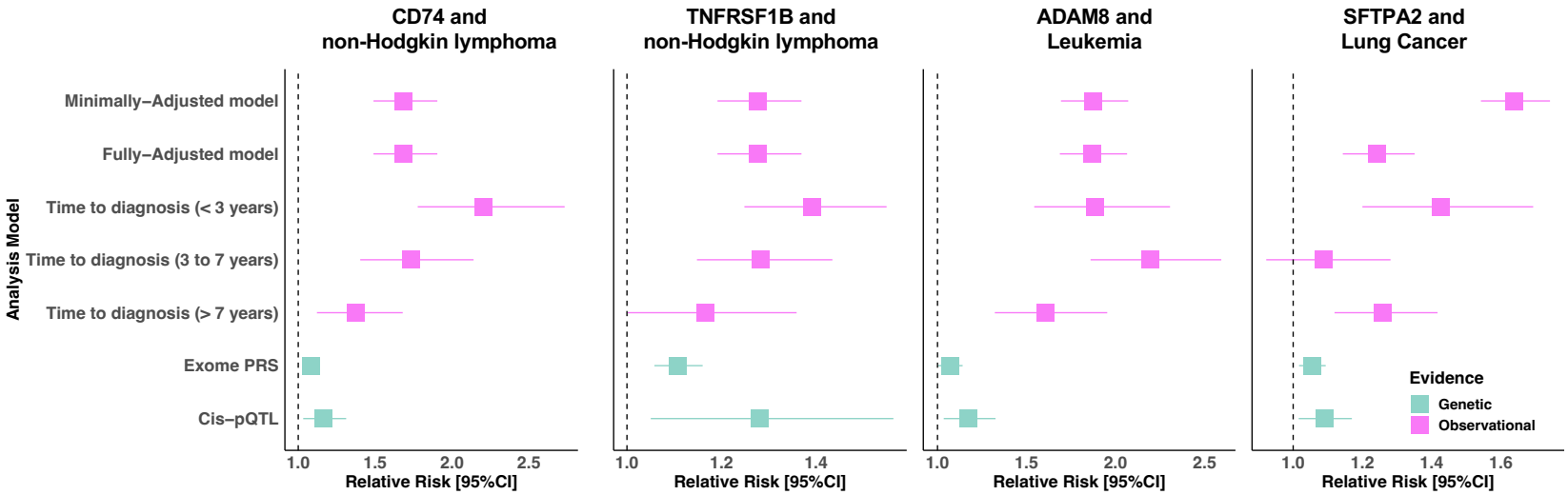

Figure 5. – Forrest plots for the prospective and genetic associations of SFPTA2 with lung cancer risk, CD74 and TNFRSF1B with risk of non-Hodgkin lymphoma, and ADAM8 with risk of leukemia