Evaluation of Circulating Plasma Proteins in Breast Cancer: A Mendelian Randomization Analysis

Anders Mälarstig[1,2], Felix Grassmann[1,3], Leo Dahl[4], Marios Dimitriou[1,2], Dianna McLeod[1], Marike Gabrielson[1], Karl Smith-Byrne[5], Cecilia E. Thomas[4], Tzu-Hsuan Huang[6], Simon KG Forsberg[7], Per Eriksson[7], Mikael Ulfstedt[7], Mattias Johansson[8], Aleksandr V. Sokolov[9], Helgi B. Schiöth[9], Per Hall[1, 10], Jochen M. Schwenk[4], Kamila Czene[1], Åsa K. Hedman[1,2]

1. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
2. Pfizer Worldwide Research, Development and Medical, Stockholm, Sweden
3. Institute of Clinical Research and Systems Medicine, Health and Medical University, Potsdam, Germany
4. Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology, Solna, Sweden
5. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford
6. Cancer Immunology Discovery, Pfizer Inc., San Diego, California
7. Olink Proteomics AB, Uppsala, Sweden
8. Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France
9. Department of Surgical Sciences, Functional Pharmacology and Neuroscience, Uppsala University, Uppsala, Sweden.
10. Department of Oncology, Södersjukhuset, Stockholm, Sweden

Corresponding author:

Anders Malarstig, Ph.D.,

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet

Nobels väg 12A

171 65 Stockholm, Sweden

Phone: +46 (0) 8-55052514

E-mail: anders.malarstig@ki.se

Abstract

The blood proteome reflects homeostatic and dynamic cellular processes across human organs. However, few blood proteomics studies of sufficient depth and size have been reported in breast cancer. To comprehensively identify circulating proteins with a causal role in breast cancer we measured 2,929 unique proteins in plasma from 598 women selected from the Karolinska Mammography Project and explored associations between proteins levels, clinical characteristics, and gene variants. The analysis revealed 812 cis-acting protein quantitative trait loci (pQTL), which were used as instruments in Mendelian randomisation (MR) analysis of breast cancer. Five proteins ($P < 1.7x10-5$, Bonferroni-corrected) with a potential causal role in breast cancer risk were revealed (CD160, DNPH1, LAYN, LRRC37A2 and TLR1). Confirming the MR findings in independent cohorts (FinnGen R9 and the UK Biobank), our study suggests that these proteins should be further explored as potential drug targets in breast cancer.

Introduction

Breast cancer is globally the most common cancer in women and is associated with significant morbidity and mortality [1]. Genome-wide and exome-wide genetic association studies have successfully identified over 300 breast cancer susceptibility loci [2-4] but the mechanisms underpinning most loci and specific gene variants remain uncharacterized, which limits translation of genetic susceptibility loci to new therapies and precision medicine tools [4].

Mendelian randomisation (MR) offers an alternative approach to the mapping and understanding of etiologically important pathways in cancer risk and development. MR aims to elucidate causal relationships between modifiable risk factors and disease based on the analysis of genetic variants in observational data [5]. In comparison to genome-wide association studies (GWAS), MR exploits a more confined test space, which increases statistical power, and inherently supports causal gene identification. MR can be further supported by genetic colocalization analysis of exposure and outcome [6]. The relevance of MR has been evaluated and supported by retrospective analyses of drug targets with a proven aetiological or causal role in disease from randomised controlled trials (RCT) [7,8].

Circulating proteins possess many of the characteristics suitable for discovery of breast cancer biology using MR. Firstly, the plasma proteome has been shown to reflect both normal physiology and pathogenic biological processes in cancer [9]. Secondly, circulating proteins can be measured with high throughput and precision a variety of advanced methods [10 11]. Thirdly, recent studies have shown that a majority of circulating proteins are associated with cis-acting protein quantitative trait loci (pQTL) i.e. located within 1 Mbp from the protein-encoding gene [12,13]. Fourthly, individual cis-pQTL explain relatively large proportions of variance in the protein, making them statistically powerful instrumental variables for causal inference using MR [12 14]. Hundreds of pQTL for plasma proteins have been identified, but so far no studies have reported pQTL in an entirely female population [7,12,13,15-19].

Here, we measured a total of 2,929 unique proteins using the Olink PEA Explore assay in plasma samples taken from 598 women who were free of a breast cancer diagnosis at the time of sampling. We i) performed genetic association analysis of protein levels to identify cis-pQTL and ii) used the cis-pQTL as instrumental variables in MR analysis of breast cancer in the BCAC case-control meta- analysis of breast cancer risk, and iii), replicated MR findings in a second breast cancer case-control meta-analysis of FinnGen [20] and the UK Biobank [21]. Lastly, we followed up on significant proteins identified in the MR analysis by visualising and evaluating colocalization of the protein and breast

cancer genetic associations and evaluated potential causal relationships with established and emerging breast cancer risk factors, also using MR (figure 1).

Out of 737 plasma proteins evaluated using MR, genetically elevated levels of five proteins were associated with breast cancer risk, namely CD160, 2'-deoxynucleoside 5'-phosphate N-hydrolase 1 (DNPH1), layilin (LAYN), Leucine rich repeat containing 37 member A2 (LRRC37A2) and toll-like receptor 1 (TLR1), which were confirmed in an independent set of data. Our results suggest that these five proteins are aetiologically relevant for breast cancer development. Pending further validation, these findings may point to novel drug target opportunities or stratification biomarkers in breast cancer.

Results

Sample characteristics

The KARMA study consented and recruited a total of 70,877 women during mammography screening from two Swedish regions (Stockholm and Skåne). The aim of the project is identification of risk factors for breast cancer [22]. The sample for the present substudy was selected for the purpose of evaluating plasma protein biomarkers in relation to incident breast cancer within 2 years from blood sampling, which is described in our companion paper by Grassmann et al. The selection included samples from 299 women in the Southern Sweden (Skåne) region who received a breast cancer diagnosis within 2 years after blood draw and 299 random controls from the same region, who, as of 2021, had remained breast cancer free. No difference between cases and controls was seen for median age, body mass index or percent women receiving hormone replacement therapy at time of blood draw. The proportion of smokers and women with a family history of breast cancer were more common among cases (Table 1).

Protein analysis, detectability, and quality control

We chose to analyse the plasma samples using an affinity proteomics approach. While targeted methods, such as the Olink PEA approach, are inherently biased towards the subset of proteins that are measured, we attempted to maximise the possibility for discovery by measuring as many proteins as possible. Hence, we used the recently launched version of Olink's Explore I and II panels, which includes 2,949 proteins (Supplementary table 5). Out of this set, 2,213 (75%) could be detected in > 50% of the samples when judging their normalized protein expression levels (NPX) above limit of detection (LOD) (Supplementary figure 1, Supplementary table 5). The ranges per

protein varied between 0.17 NPX and 9.27 NPX (Supplementary figure 2). The proportion of proteins above LOD were lower for the most recent addition to the panels (Explore II). However, it is worth noting that the set of proteins in Explore II are, on average, less abundant than those of the Explore I panel, as shown in a comparison of average levels across proteins overlapping with a mass spectrometry peptide-based analysis generated by the Human Protein Atlas effort (Supplementary table 3, Supplementary figure 3) [23].

Association between plasma protein levels and clinical characteristics

To examine observational relationships between protein levels and clinical characteristics of the KARMA women, we regressed each measured protein against seven factors (age, alcohol consumption, number of births, body mass index (BMI), hormone replacement therapy (HRT), peri- and post-menopause and current smoking. In these analyses we included both women who developed breast cancer and those who did not as there were no significant differences between both groups in our companion paper, indicating that the protein levels are similar between both groups at blood draw. All associations are shown in Supplementary table 6. A total of 684 proteins were associated with BMI and 459 proteins were associated with age (Figure 2). Several of the observed associations have previously been described such as higher plasma levels of leptin and fatty-acid binding protein 4 (FABP4) with increasing BMI [24], higher FSHB in post-menopausal women and higher PLAP levels in smokers [25]. Some less described correlations included lower plasma levels of glycodelin (PAEP) and chordin like 2 (CHRDL2) and higher levels of glycoprotein hormone alpha polypeptide (CGA) in post- and peri-menopausal women, and lower levels of osteomodulin (OMD) in women using hormone replacement therapy (HRT).

The replication of known trait-to-protein associations suggest that the data quality was satisfactory, and that additional trait-to-protein associations are enabled by expansion of the number of detectable proteins.

Identification of cis-pQTL

To identify genetic instruments for the downstream causality testing using MR, gene variants within a range of 1Mbp up and downstream of genes encoding each of the 2,929 unique proteins were tested for association with levels of the corresponding protein. Significant associations ($p<2.2x10^{-4}$) were observed for a total of 812 independent variants ($R^2>0.1$) and 737 proteins, henceforth referred to as cis-pQTL (supplementary table 1). Most of the pQTL were observed for proteins on Olink Explore I panel (n=523) but several pQTL were also observed for Explore II proteins (n=289). Some of the cis-

pQTL showed effect sizes well above 1 standard deviation, including the nucleotidase NT5C (missense, Pro68Leu, MAF 3 %), acylphosphatase (ACYP1) (~7 kbp upstream of gene, MAF 1.5 %) and carboxypeptidase Q (CPQ) (intron, MAF 1.7 %).

We conclude that pQTL are readily detected for proteins on both Explore I and II panels, providing potential MR instruments for 737 proteins.

Replication analysis

To investigate the validity of the cis-pQTL identified in KARMA, effect sizes were compared with cis-pQTL previously reported for a subset of 90 proteins measured using Olink PEA in the SCALLOP CVD-I study [7]. Measurements for all 90 proteins were available in the KARMA study. Of those 90, cis-pQTL for 33 of the proteins reported by the SCALLOP CVD-I study were associated in KARMA at $p<0.05$. The Pearson correlation coefficient between effect sizes for the 33 overlapping variants was 0.91 (supplementary figure 4).

To also investigate the generalisability of the identified cis-pQTL, the variants, or those in high linkage disequilibrium (LD) (>0.8), were looked up in previously published studies reporting cis-pQTL based on the Somascan proteomics platform [26,27]. The overlap of Olink proteins available after quality control in the KARMA study and proteins measured in previously published work based on the Somascan platform was 569 proteins (supplementary table 1). Of the 603 significant cis-pQTL observed in KARMA for the subset of overlapping proteins, we observed evidence of replication for 374 proteins at Bonferroni-corrected $p<6.1\times10^{-5}$ whereas a total of 229 cis-pQTL did not show evidence of replication at the aforementioned p-value threshold.

Mendelian randomization analysis

We performed two-sample inverse-variance weighted or Wald-scores MR analysis using protein exposures from the KARMA cis-pQTL to investigate potential causal effects on breast cancer risk using outcome data from BCAC and from the FinnGen R8-UK-biobank meta-analysis [5]. We were unable to identify genetic proxies for seven of the proteins with cis-pQTL in KARMA, resulting in the testing of 730 protein exposures. Of those, seven proteins surpassed the statistical threshold for significance ($p<7.5\times10^{-5}$) in the discovery study (Figure 3) of which five replicated in the independent breast cancer case control study from FinnGen [20] and UK-biobank [21] with consistent effect sizes and directions (Table 2). The replicated proteins, shown here by the names of their encoding genes, were CD160, DNPH1, LAYN, LRRC37A2 and TLR1. The full summary of MR results is provided in Supplementary table 4.

We further investigated whether the five proteins with replicated MR evidence for all breast cancers were equally associated in estrogen-receptor (ER) positive compared to ER negative breast cancer (Table 3). However, the effect sizes were similar across ER+ and ER- breast cancer risk, suggesting these five proteins associate equally with ER+ and ER- breast cancer risk.

It was also hypothesised that proteins with MR evidence for an etiologically important role in breast cancer might influence breast cancer risk via a breast cancer risk factor. To test this, further MR analysis was performed using GWAS of potential breast cancer risk factors as outcomes, including age at menarche, age at menopause, waist-hip ratio, mammographic density, sex hormone binding globulin and insulin growth factor 1 levels (IGF-1) [28]. LRRC37A2 showed MR evidence for later age at menarche and earlier age at menopause in two independent outcome datasets, and also for higher IGF-1 levels (Supplementary table 2). CD160 showed nominal MR evidence for an etiological role lower age at menarche.

To summarise, the MR analysis showed that genetic elevation of CD160, DNPH1, LAYN, LRRC37A2 and TLR1 associate with breast cancer risk, and with similar effects on ER+ and ER- cancer.

Colocalisation analysis

All imputed variants in proximity to the cis-pQTL for proteins with significant MR evidence were visually inspected with the corresponding genomic region for breast cancer risk using mirror plots. The cis-regions around DNPH1 and LRRC37A2 showed the strongest degree of concordance between lead variants for protein levels and breast cancer risk (Supplementary figure 7 and 8). Lead pQTL in cis-regions for CD160, LAYN and TLR1 were not the variants with the lowest p-values for breast cancer risk but were localised in the same, size limited, genomic region. We considered the cis-pQTL to be colocalised with breast cancer risk (Supplementary figure 6, 8 and 10).

Systematic search for drugs targeting CD160, DNPH1, LAYN, LRRC37A2 and TLR1

To investigate if any of the five proteins identified in the present investigation had been previously explored as drug targets, we performed a systematic search across several databases, including NIH Pharos Consortium, IUPHAR/BPS Guide to Pharmacology, DrugBank and ClinicalTrials.gov. With the exception of LAYN, targeted by Hyaluronic acid, none of the proteins were registered as known drug targets [29].

Discussion

We measured 2,949 circulating proteins in plasma from 598 women to identify 812 independent cis-pQTL which were applied in MR to investigate associations between genetically predicted protein levels and breast cancer risk. We found that genetically lower levels of CD160 and LRRC37A2 and genetically higher levels of DNPH1, LAYN and TLR1 were associated with increased risk of breast cancer. In addition, genetically higher levels of LRRC37A2 associated with age at menarche, which adds to previous knowledge of its modest MR evidence for breast cancer risk [28]. MR using cis-pQTL instruments allowed us to model life-long genetic exposure to higher/lower protein levels, which implies an aetiologically important role of associated proteins in disease. In our companion paper by Grassmann et al., we found no circulating proteins associated with 2-year risk of incident breast cancer. Indeed, none of the five proteins identified in the present investigation were significantly associated with incident breast cancer. This indicates that genetically predicted protein levels did not capture this short-term risk.

Among the five proteins identified in our study, DNPH1, also described as Rcl, encodes the enzyme 2'-deoxynucleoside 5'-phosphate N-hydrolase, which plays a role in nucleotide metabolism and is a target of ETV1 -a transcription factor expressed in breast tumours [30]. Two independent CRISPR screens for modulators of BRCA-associated breast tumour sensitivity to PARP inhibitors, an established treatment in BRCA-deficient breast cancer, have shown that genomic inhibition DNPH1 sensitizes BRCA-deficient cells to treatment with PARP inhibitors [31,32]. The lead pQTL identified in KARMA, rs75591122, is located ~18.2 kbp upstream from the DNPH1 gene on chromosome 6 and is one of several variants proximal to the DNPH1 gene associated with DNPH1 gene expression levels across multiple tissues [33]. Genetically increased circulating protein levels of DNPH1 was in our study associated with increased breast cancer risk, which is concordant with experimental studies suggesting that DNPH1 inhibition in breast cancer may be promising avenue for drug development.

Another of the five proteins was CD160, which is a receptor expressed in immune cells that has been described to play important roles in NK cell biology, predominantly functioning as an activating NK-cell receptor [34]. CD160 is predominantly expressed on healthy NK cells and is one of the driver genes for a specific NK subset related to higher cytokine production [35]. Reduction in CD160 expression led to impaired NK cells and poor outcomes in Hepatocellular carcinoma patients [36] and since dysfunctional NK cells also correlate with breast cancer progression [37] it can be hypothesized that CD160 could have a similar protective role in breast cancer. Indeed, in our study, genetically elevated circulating protein levels of CD160 associated with a protective effect in breast cancer, suggesting

that a drug activating CD160 specifically on NK cells may enhance anti-tumour immune responses in breast cancer.

Our search for drug targets highlighted the connection between LAYN and Hyaluronic Acid. LAYN encodes Layilin, which is a talin-binding transmembrane and integral membrane protein functioning as a receptor for Hyaluronic acid (HA), with a role in cell adhesion and motility [38,39]. HA is an extracellular matrix component that impacts tumor microenvironment where elevated HA levels has been reported in multiple cancer types including breast cancer [40]. Interestingly, targeted depletion of HA controlled the breast cancer tumor growth in xenotransplant mouse models of immunocompetent mice but not of immunodeficient mice, which indicates a potential tumor-immunity role for its receptors i.e. Layilin [41]. Accordingly, high LAYN expression belongs to transcriptomic signatures specific for regulatory T cells (Tregs) and exhausted CD8+ T cells for several cancer types including breast cancer [42 43]. In our study, genetic elevation of LAYN protein levels associated with increased breast cancer risk, suggesting a LAYN inhibitor would be desired for treatment of breast cancer. However, mechanistic studies will be required to confirm the direction of effect proposed by the MR evidence and to validate LAYN as drug target in breast cancer.

Several other studies have investigated genetic elevation of circulating proteins to identify potential aetiological or causal factors for breast cancer risk. Murphy et al. reported that genetically elevated circulating insulin growth factor levels (IGF-1) were associated with a weak but significantly increased risk of breast cancer whereas IGF-binding protein-3 was unassociated [44]. Zhu et al. demonstrated absence of association with breast cancer for genetically elevated levels of C-reactive protein [45] and Shu et al. reported a wider MR analysis, instrumenting 1,469 proteins using Somascan-based pQTL in the INTERVAL cohort, of which genetic instruments for 26 proteins were found to be associated [45,46]. Bouras et al. instrumented 47 inflammatory cytokines and reported that genetically increased levels of CXCL1 and decreased levels of MIF associated with breast cancer [47]. Our study included 10 of the 28 proteins previously reported in breast cancer MR studies, and while none of the reported proteins surpassed statistical significance in our study, SCG3 and TFPI showed nominal significance in our discovery MR (Supplementary table 4).

Our study has both strengths and limitations. One of the strengths is the large number of proteins tested for cis-pQTL and that the cis-pQTL used to instrument genetic elevation using MR were identified in women only, which should provide better estimates in MR for female breast cancer. Another strength is that the protein exposures meeting statistical significance in our discovery MR, using data from the BCAC consortium as outcome, were replicated in the independent case-control analysis that combined breast cancer cases and controls in FinnGen and the UK-Biobank.

However, our study had limited sample size for discovering cis-pQTL with smaller effect sizes. Therefore, we cannot exclude that additional proteins on the Olink Explore II panels harbour significant cis-pQTL but remained undetected in the KARMA sample. To decrease the false-negative error rate we only included variants in cis to decrease the multiple-test burden and corrected the p-value threshold for significant for the number of independent variants in each cis-region. Effect-sizes observed in KARMA were highly concordant with an overlapping set of 33 cis-pQTL for proteins measured with Olink PEA that were previously reported. To evaluate the robustness of cis-pQTL identified in KARMA, we sought replication for an overlapping set of 569 proteins measured with Somascan. Of those, 2/3 (374/569) were replicated, which is on par with the expected replication rate given differences in protein analysis methods [16].

In conclusion, by applying an MR approach for a broad range of circulating proteins we found that genetically elevated CD160, DNPH1, LAYN, LRRC37A2 and TLR1 associate with breast cancer. This suggests that these five proteins play an aetiological or causal role in breast cancer, providing a basis for further functional evaluation of their potential as drug targets.

Materials and methods

KARMA study collection

We included 299 breast cancer cases and 299 breast cancer free controls from the Swedish KARMA study in the analysis. The cohorts are thoroughly described elsewhere and previously analysed in several BCAC studies. Briefly, the KARMA Cohort consists of 70,877 women performing a screening or clinical mammogram at 4 hospitals in Sweden during the period October 2010–March 2013.

Plasma protein measurements on Olink Explore

Plasma proteomics was performed in samples from 299 BC cases and 299 BC free controls from the Swedish KARMA study using the Olink Explore I and II panels (Olink Proteomics AB, Uppsala, Sweden) according to the manufacturer's protocol. Explore combines the Proximity Extension Assay (PEA) technology with Next generation sequencing (NGS).

In brief, the PEA technology uses matching pairs of oligonucleotide-labelled antibody probes. The PEA probes bind to target antigens producing a binding complex where the complimentary oligonucleotides exist in close proximity to each other, enabling the formation of a target sequence. The dual targeting of probes has been proven to produce outstanding specificity enabling for a high degree of multiplexing while maintaining sensitivity and a broad dynamic range. In the Olink Explore

protocol, target sequence is amplified in a double PCR reaction and purified before the NGS. The sequence data is processed and normalized to produce Olinks relative quantification unit Normalized Protein eXpression (NPX). The produced DNA signal functionally works as a proxy for the protein levels present in the sample. Further details on the Olink Explore protocol and internal quality control are available in the Supplementary methods 1 document.

Olink analysis quality control

The Olink QC-system includes negative controls, used to monitor the background noise and to set the limit of detection (LOD). Supplementary figure 1 and Supplementary table 5 show the percentage of samples with NPX above LOD.

Association with clinical characteristics

For each of the 2,949 measured protein levels, the following linear regression model was fitted: *NPX ~ age + bmi + menopause_preVSperi + menopause_preVSpost + birth_times + hrt_status + alcohol_gram_week + smoking_status* where menopause_preVSperi contrasts pre- versus peri-menopausal patients, menopause_pre VS post contrasts pre- versus post-menopausal patients, hrt_status contrasts current users of hormone replacement therapy versus patients who have never used it or who have used it in the past, and smoking_status contrast current smokers versus those who have never smoked or smoked in the past. All p-values were FDR corrected for the 2,949 x 7 performed tests.

Protein QTL mapping

Genome-wide genotyping in the KARMA study was performed using the Illumina iSelect or Oncoarray arrays, followed by imputation using the Wellcome Trust Sanger Institute imputation service using the 1000 genomes phase 3 as reference. Standard quality control was applied as previously described. Variants with a minor allele frequency < 0.01 were filtered out prior to analysis. The final dataset included 9,087 million variants.

Proteins >75 % of NPX values below LOD were filtered out before the pQTL analysis, yielding a total of 2,476 proteins in the analysis. Values below LOD were included. The pQTL discovery analysis was performed using an additive model with adjustments for age, BMI and 10 genetic PCs in PLINK 2.0 . To preserve statistical power for pQTL identification, only variants within a 1 mega-base pair window of the protein coding gene were tested for association with respective circulating protein level. To manage multiple test correction, while limiting false-negatives, the total number of variants per cis-

region were calculated as well as the number of independent variants (R2<0.1). The average number of variants per cis-region was 6,249 (Supplementary Figure 5) and 180 independent variants (min,max 12-511). Statistical significance was therefore defined as an alpha of 0.05 divided by 180 to account for average number of independent variants tested per cis-region (p=2.77E-04). A false-discovery rate (FDR) at 5 % provided a similar estimate (p< 5.54E-04).

Mendelian Randomization analysis

We performed Two-sample MR using the R package Two-Sample MR to test for proteins with a potential causal role in breast cancer. Independent cis-pQTL (r2 < 0.001) were used as instrumental variables (IV), and GWAS of breast cancer risk from the BCAC consortium were used as outcome, which included data from 122,977 breast cancer cases and 105,974 controls. In the case of a single independent IV Wald Ratio was applied, otherwise inverse-variance weighted estimates were reported. The threshold for statistical significance was defined as ($7.5x10^{-5}$) to account for multiple testing. The replication analysis was performed in a meta-analysis of FinnGen R9 and the UK-biobank, which included 25,807 cases and 355,307 controls. Only the seven proteins that met statistical significance in the BCAC discovery analysis were included in the replication analysis, and hence a nominal p-value of 0.05 was considered statistically significant.

Acknowledgements

We thank all the participants in the Karma study and the study personnel for their devoted work during data collection. We also want to acknowledge the participants and investigators of the FinnGen study. The data handling and analysis were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Conflicts of interest

AM, AH and TH are employees of Pfizer Inc. SKF, PE and MU are employees of Olink Proteomics AB.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Funding

This work was financed by the Swedish Research Council (Grant 2022-00584), the Swedish Cancer Society (Grants 22 2207, 19 0267 and 20 0990), the Stockholm County Council (Grant 20200102) and the Karolinska Institutet's Research Foundation (Grant 2018-02146). This work was also supported by a grant from the Stockholm County Council (FoU-954555), Olink Proteomics AB and Pfizer Inc.

Data availability

Access to phenotypes, biospecimen and genotypes from the KARMA study can be requested from https://karmastudy.org/contact/data-access/ . Access to scripts and pipelines will be provided through GitHub.

Tables

Table 1

Variable	Controls (BC negative)	Cases (incident BC)
Number of individuals	299	299
Age at baseline (S.D) [years]	58.83 (9.26)	58.11 (9.49)
Body mass index at interview (S.D) [kg/m2]	25.20 (4.16)	25.73 (4.14)
Hormone replacement therapy ever [%]	35.66	37.76
Current smoker at interview [%]	11.23	16.32
Family history of BC [%]	11.27	20.92

Table 2

Exposures	BCAC, all breast cancer			FinnGen and UK-Biobank		
Protein	**nsnp**	**beta**	**pval**	**nsnp**	**beta**	**pval**
CD160	1	-0.09	1.70E-06	1	-0.07	1.50E-02
DNPH1	1	0.08	3.80E-07	1	0.05	3.50E-02
LAYN	1	0.13	1.40E-05	1	0.12	8.40E-03
LRRC37A2	1	-0.05	5.70E-10	1	-0.05	6.80E-05
MST1	1	0.03	7.20E-05	1	0.02	6.60E-02
TLR1	1	0.07	6.40E-06	1	0.11	7.40E-05
TXK	1	0.07	3.10E-06	1	0.03	3.40E-01

Table 3

	ER+ breast cancer				ER- breast cancer			
Exposures	**BCAC**		**FinnGen**		**BCAC**		**FinnGen**	
Protein	**beta**	**pval**	**beta**	**pval**	**beta**	**pval**	**beta**	**pval**
CD160	-0.08	5.10E-04	-0.14	6.90E-03	-0.06	9.30E-02	-0.07	2.80E-01
DNPH1	0.08	6.20E-06	0.07	8.80E-02	0.09	6.00E-04	0.05	3.40E-01
LAYN	0.12	5.50E-04	0.13	1.20E-01	0.12	2.60E-02	0.17	1.00E-01
LRRC37A2	-0.04	1.80E-06	-0.06	3.50E-02	-0.04	7.90E-03	-0.01	8.30E-01
TLR1	0.07	1.60E-04	0.11	4.10E-02	0.09	2.30E-03	0.11	9.40E-02

References

1. Allahqoli, L. *et al.* The Global Incidence, Mortality, and Burden of Breast Cancer in 2019: Correlation With Smoking, Drinking, and Drug Use. *Front Oncol* **12**, 921015 (2022).

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

3. Dumont, M. *et al.* Uncovering the Contribution of Moderate-Penetrance Susceptibility Genes to Breast Cancer by Whole-Exome Sequencing and Targeted Enrichment Sequencing of Candidate Genes in Women of European Ancestry. *Cancers (Basel)* **14**(2022).

4. Romualdo Cardoso, S., Gillespie, A., Haider, S. & Fletcher, O. Functional annotation of breast cancer risk loci: current progress and future directions. *Br J Cancer* **126**, 981-993 (2022).

5. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-63 (2008).

6. Schmidt, A.F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nat Commun* **11**, 3255 (2020).

7. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135-1148 (2020).

8. Henry, A. *et al.* Therapeutic Targets for Heart Failure Identified Using Proteomics and Mendelian Randomization. *Circulation* **145**, 1205-1217 (2022).

9. Hanash, S.M., Pitteri, S.J. & Faca, V.M. Mining the plasma proteome for cancer biomarkers. *Nature* **452**, 571-9 (2008).

10. Deutsch, E.W. *et al.* Advances and Utility of the Human Plasma Proteome. *J Proteome Res* **20**, 5241-5263 (2021).

11. Suhre, K., McCarthy, M.I. & Schwenk, J.M. Genetics meets proteomics: perspectives for large population-based studies. *Nat Rev Genet* **22**, 19-37 (2021).

12. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet* **13**, e1006706 (2017).

13. Sun, B.B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv*, 2022.06.17.496443 (2022).

14. Macdonald-Dunlop, E. *et al.* Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases. *medRxiv*, 2021.08.03.21261494 (2021).

15. Yang, Z. *et al.* Genetic Landscape of the ACE2 Coronavirus Receptor. *Circulation* **145**, 1398-1411 (2022).

16. Katz, D.H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci Adv* **8**, eabm5164 (2022).

17. Png, G. *et al.* Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nat Commun* **12**, 7042 (2021).

18. Zhernakova, D.V. *et al.* Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat Genet* **50**, 1524-1532 (2018).

19. Enroth, S., Johansson, A., Enroth, S.B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat Commun* **5**, 4684 (2014).

20. Kurki, M.I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).

21. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).

22. Gabrielson, M. *et al.* Cohort Profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). *Int J Epidemiol* **46**, 1740-1741g (2017).

23. Uhlen, M. *et al.* The human secretome. *Sci Signal* **12**(2019).

24. Lind, L. *et al.* Changes in Proteomic Profiles are Related to Changes in BMI and Fat Distribution During 10 Years of Aging. *Obesity (Silver Spring)* **28**, 178-186 (2020).

25. Rasmuson, T. *et al.* Tumor markers in mammary carcinoma. An evaluation of carcinoembryonic antigen, placental alkaline phosphatase, pseudouridine and CA-50. *Acta Oncol* **26**, 261-7 (1987).

26. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat Commun* **12**, 6822 (2021).

27. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

28. Chen, F. *et al.* Mendelian randomization analyses of 23 known and suspected risk factors and biomarkers for breast cancer overall and by molecular subtypes. *Int J Cancer* **151**, 372-380 (2022).

29. Bono, P., Rubin, K., Higgins, J.M. & Hynes, R.O. Layilin, a novel integral membrane protein, is a hyaluronan receptor. *Mol Biol Cell* **12**, 891-900 (2001).

30. Shin, S., Bosc, D.G., Ingle, J.N., Spelsberg, T.C. & Janknecht, R. Rcl is a novel ETV1/ER81 target gene upregulated in breast tumors. *J Cell Biochem* **105**, 866-74 (2008).

31. Fugger, K. *et al.* Targeting the nucleotide salvage factor DNPH1 sensitizes BRCA-deficient cells to PARP inhibitors. *Science* **372**, 156-165 (2021).

32. Zimmermann, M. *et al.* CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature* **559**, 285-289 (2018).

33. GTEx_Consortium. GTEx. (2023).

34. Le Bouteiller, P. *et al.* CD160: a unique activating NK cell receptor. *Immunol Lett* **138**, 93-6 (2011).

35. Crinier, A. *et al.* Single-cell profiling reveals the trajectories of natural killer cell differentiation in bone marrow and a stress signature induced by acute myeloid leukemia. *Cell Mol Immunol* **18**, 1290-1304 (2021).

36. Sun, H. *et al.* Reduced CD160 Expression Contributes to Impaired NK-cell Function and Poor Clinical Outcomes in Patients with HCC. *Cancer Res* **78**, 6581-6593 (2018).

37. Mamessier, E. *et al.* Human breast cancer cells enhance self tolerance by promoting evasion from NK cell antitumor immunity. *J Clin Invest* **121**, 3609-22 (2011).

38. Borowsky, M.L. & Hynes, R.O. Layilin, a novel talin-binding transmembrane protein homologous with C-type lectins, is localized in membrane ruffles. *J Cell Biol* **143**, 429-42 (1998).

39. Sedy, J.R. *et al.* CD160 activation by herpesvirus entry mediator augments inflammatory cytokine production and cytolytic function by NK cells. *J Immunol* **191**, 828-36 (2013).

40. Henke, E., Nandigama, R. & Ergun, S. Extracellular Matrix in the Tumor Microenvironment and Its Impact on Cancer Therapy. *Front Mol Biosci* **6**, 160 (2019).

41. Zamloot, V., Ebelt, N.D., Soo, C., Jinka, S. & Manuel, E.R. Targeted Depletion of Hyaluronic Acid Mitigates Murine Breast Cancer Growth. *Cancers (Basel)* **14**(2022).

42. De Simone, M. *et al.* Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. *Immunity* **45**, 1135-1147 (2016).

43. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169**, 1342-1356 e16 (2017).

44. Murphy, N. *et al.* Insulin-like growth factor-1, insulin-like growth factor-binding protein-3, and breast cancer risk: observational and Mendelian randomization analyses with approximately 430 000 women. *Ann Oncol* **31**, 641-649 (2020).

45. Zhu, M. *et al.* C-reactive protein and cancer risk: a pan-cancer study of prospective cohort and Mendelian randomization analysis. *BMC Med* **20**, 301 (2022).

46. Shu, X. *et al.* Evaluation of associations between genetically predicted circulating protein biomarkers and breast cancer risk. *Int J Cancer* **146**, 2130-2138 (2020).

47. Bouras, E. *et al.* Circulating inflammatory cytokines and risk of five cancers: a Mendelian randomization analysis. *BMC Med* **20**, 3 (2022).

Figures

Figure 1

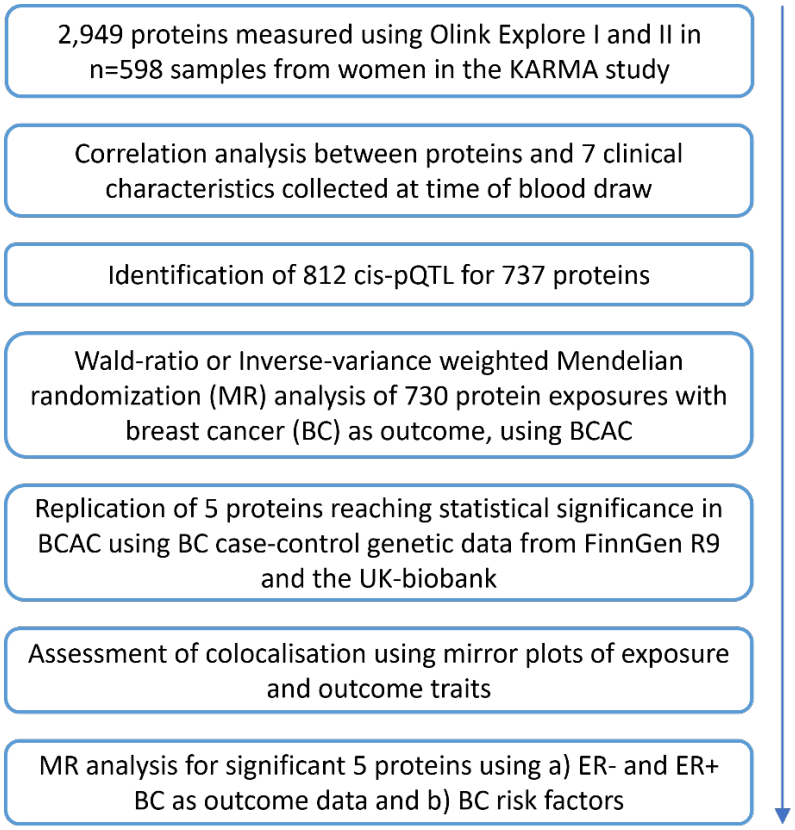

Figure 1. Flow chart of study design, analyses and main results

Figure 2

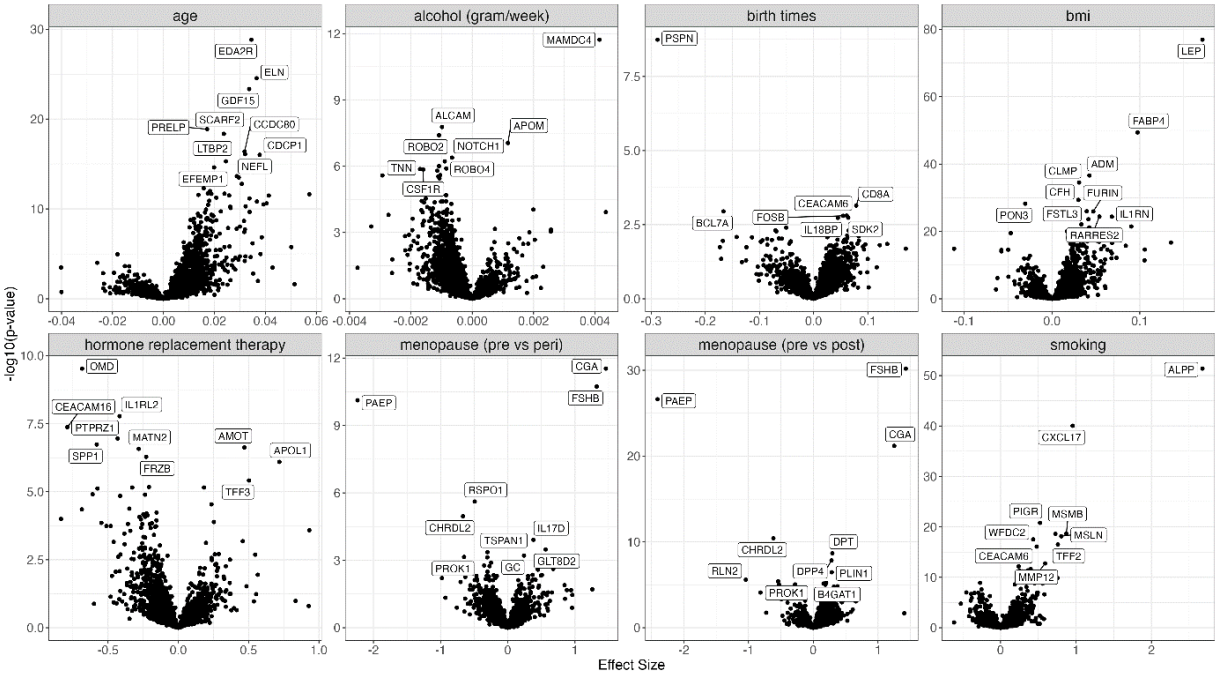

Figure 2*:* Volcano plots showing estimated effect sizes (x-axis) and the corresponding non-adjusted – log10(p-value) (y-axis). Effect sizes were given by a linear regression model per protein, including all 7 traits. Each panel shows one of the investigated baseline traits, corresponding to one term in the regression model. The names of the topmost significant proteins per trait are indicated in each panel. The number of proteins reaching FDR corrected statistical significance were for age:459, Alcohol consumption:172, Birth times:7, BMI:684, HRT:93, Menopause pre vs. peri:18, Menopause pre vs post:127, Current smoking:213.

Figure 3

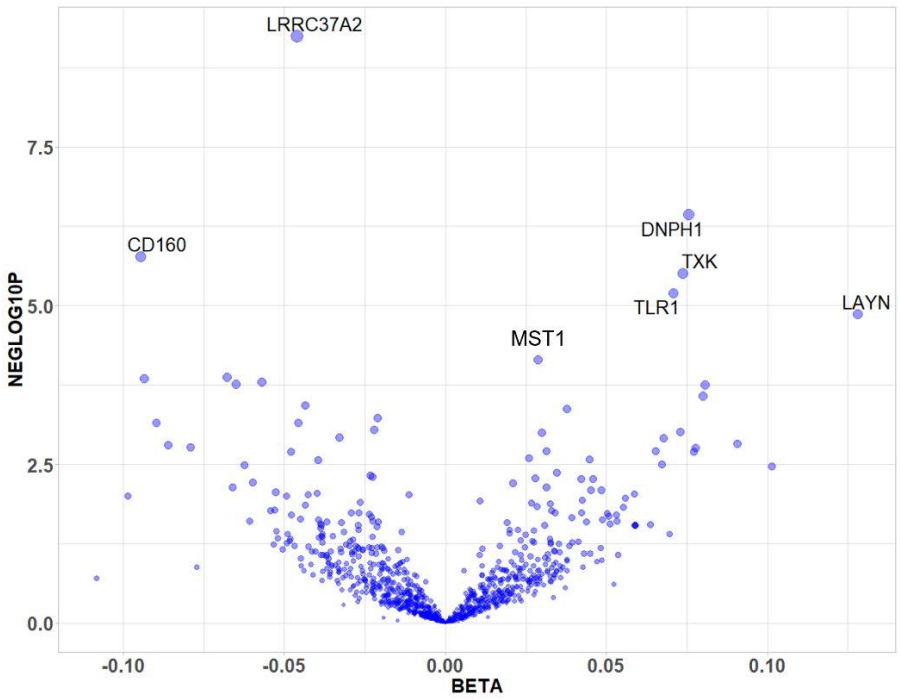

Figure 3: Mendelian randomization analysis on breast cancer risk in the BCAC study was performed by modelling exposure to genetically higher plasma levels of 730 proteins with at least one cis-pQTL. The Y-axis shows the -log10 p-value of the Wald-score or IVW and the X-axis shows the beta-estimates of the MR result for each protein that was tested.