Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 18 January 2024
DOI 10.3389/fmolb.2023.1340917

Check for updates

OPEN ACCESS

EDITED BY
Li Chen,
Huazhong University of Science and Technology, China

REVIEWED BY
Min Li,
Western Theater General Hospital, China
Yawei Hua,
Henan Provincial Cancer Hospital, China
Zhiheng Lin,
Shandong University of Traditional Chinese Medicine, China
Dongdong Zhang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Bailin Zhang,
✉ zhangbl@cicams.ac.cn
Gao Jidong,
✉ ab168@cicams.ac.cn

RECEIVED 19 November 2023
ACCEPTED 29 December 2023
PUBLISHED 18 January 2024

CITATION
Wang Y, Yi K, Chen B, Zhang B and Jidong G (2024), Elucidating the susceptibility to breast cancer: an in-depth proteomic and transcriptomic investigation into novel potential plasma protein biomarkers.
Front. Mol. Biosci. 10:1340917.
doi: 10.3389/fmolb.2023.1340917

Elucidating the susceptibility to breast cancer: an in-depth proteomic and transcriptomic investigation into novel potential plasma protein biomarkers

Yang Wang[1], Kexin Yi[1], Baoyue Chen[2], Bailin Zhang[1]* and Gao Jidong[1,3]*

[1]Department of Breast Surgical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Department of General Surgery, Beijing Puren Hospital, Beijing, China, [3]Department of Breast Surgical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital and Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China

Objectives: This study aimed to identify plasma proteins that are associated with and causative of breast cancer through Proteome and Transcriptome-wide association studies combining Mendelian Randomization.

Methods: Utilizing high-throughput datasets, we designed a two-phase analytical framework aimed at identifying novel plasma proteins that are both associated with and causative of breast cancer. Initially, we conducted Proteome/Transcriptome-wide association studies (P/TWAS) to identify plasma proteins with significant associations. Subsequently, Mendelian Randomization was employed to ascertain the causation. The validity and robustness of our findings were further reinforced through external validation and various sensitivity analyses, including Bayesian colocalization, Steiger filtering, heterogeneity and pleiotropy. Additionally, we performed functional enrichment analysis of the identified proteins to better understand their roles in breast cancer and to assess their potential as druggable targets.

Results: We identified 5 plasma proteins demonstrating strong associations and causative links with breast cancer. Specifically, PEX14 (OR = 1.201, $p = 0.016$) and CTSF (OR = 1.114, $p < 0.001$) both displayed positive and causal association with breast cancer. In contrast, SNUPN (OR = 0.905, $p < 0.001$), CSK (OR = 0.962, $p = 0.038$), and PARK7 (OR = 0.954, $p < 0.001$) were negatively associated with the disease. For the ER-positive subtype, 3 plasma proteins were identified, with CSK and CTSF exhibiting consistent trends, while GDI2 (OR = 0.920, $p < 0.001$) was distinct to this subtype. In ER-negative subtype, PEX14 (OR = 1.645, $p < 0.001$) stood out as the sole protein, even showing a stronger causal effect compared to breast cancer. These associations were robustly supported by colocalization and sensitivity analyses.

Conclusion: Integrating multiple data dimensions, our study successfully pinpointed plasma proteins significantly associated with and causative of breast cancer, offering valuable insights for future research and potential new biomarkers and therapeutic targets.

KEYWORDS

proteome-wide association study, transcriptome-wide association study, plasma proteins, mendelian randomization, breast cancer

1 Introduction

In 2020, a concerning 2.3 million women were diagnosed with breast cancer, establishing it as the most common cancer among women worldwide (Sung et al., 2021). This high prevalence underscores the urgency for ongoing research; however, despite significant efforts, the precise causes of breast cancer remain elusive. The disease is marked by a wide range of biological characteristics, including diverse histological and molecular features (Prat and Perou, 2011). Among these, the estrogen receptor (ER) status stands out as a crucial biomarker, significantly influencing treatment strategies such as endocrine therapy for ER-positive breast cancers (Trayes and Cokenakes, 2021). In addition to tissue-specific protein markers, the study of proteins in circulating plasma, often found due to cellular leakage or active secretion (Anderson and Anderson, 2002), is increasingly important. Due to the ease of detection and reproducibility of plasma proteins, these proteins are suitable for biomarkers and potential therapeutic targets (Suhre et al., 2021). Recent studies have highlighted the significant relationship between a variety of circulating proteins and breast cancer, thereby providing crucial insights into the disease's prognosis (Key et al., 2010; Christopoulos et al., 2015; Rosendahl et al., 2021; Veyssière et al., 2022; Mälarstig et al., 2023). The identification of these proteins as potential biomarkers has opened new avenues for early detection and personalized medicine in breast cancer, emphasizing the importance of understanding the complex biological interactions and pathways involved in cancer progression.

Genome-wide association studies (GWAS) have been instrumental in identifying nearly 200 genetic loci associated with breast cancer, revealing insights into genetic predispositions (Michailidou et al., 2017; Shu et al., 2020; Zhang et al., 2020; Gudjonsson et al., 2022). These discoveries underscore the importance of genetic factors in breast cancer susceptibility. Particularly, SNPs located within a 500 Kb range of the transcription start sites of protein-coding genes, known as cis-acting quantitative trait loci (cis-QTLs). Among these, protein Quantity Trait Loci (pQTLs) are crucial for regulating protein levels and are valuable tools for research (Sun et al., 2018). Utilizing pQTL as genetic proxies allow us to make a deeper exploration of the role of plasma proteins in breast cancer susceptibility. Recently, Proteome-Wide Association Studies (PWAS) (Wingo et al., 2021) and Transcriptome-Wide Association Studies (TWAS) (Gusev et al., 2016) have been pivotal in understanding the functions of proteins and gene expression in disease onset and progression. Initial PWAS focused primarily on neurological contexts due to data limitations (Zhang et al., 2022a), However, recent advancements (Zhang et al., 2022a) have broadened the scope of these studies to include diverse health conditions, thereby enriching our understanding of the associations between plasma proteins and various diseases (Li et al., 2023).

Our first phase focused on identifying proteins that are inherently associated with breast cancer at both proteomic and transcriptomic levels. For PWAS analysis, we integrated plasma protein pQTL data from ARIC cohort (Zhang et al., 2022a) with breast cancer GWAS summary data, including its different ER subtypes. Additionally, we carried out a supplementary TWAS in whole blood and breast mammary tissues. This combined P/TWAS methodology revealed significant associations between plasma proteins and breast cancer. However, it is crucial to note that such associations do not automatically imply causations. To address this, in our second phase, we employed two-sample Mendelian Randomization (MR) analysis (Emdin et al., 2017), adding a causal dimension to the protein-breast cancer relationship. We further assessed shared causal variants between them by genetic Bayesian colocalization. To ensure the robustness and broader applicability of our findings, we further conducted external validations of the established causal link. These validations were achieved using 4 extensive large plasma protein pQTL datasets (Folkersen et al., 2017; Sun et al., 2018; Ferkingstad et al., 2021; Gudjonsson et al., 2022) and the eQTLGen dataset (Võsa et al., 2021).

In our study, we implemented a two-phase design that integrates P/TWAS with MR analyses. This comprehensive methodology, blending associative and causative analyses, provides valuable insights into breast cancer. Furthermore, the relative simplicity in detecting plasma proteins not only strengthens their role in development of diagnostic biomarkers but also suggests their potential value in the development of therapeutic targets for breast cancer.

2 Materials and methods

2.1 Research framework

The analysis flowchart for the study is presented in Figure 1. A two-phase analytical approach was employed in this study, merging P/TWAS for association and MR for causation. Additionally, to guarantee the validity and reliability of the findings, a discovery-confirmatory framework was implemented in both phases.

2.2 Breast cancer GWAS summary data source

The GWAS summary data from the Breast Cancer Association Consortium (BCAC), which specifically focused on individuals of European descent (https://bcac.ccge.medschl.cam.ac.uk/), was

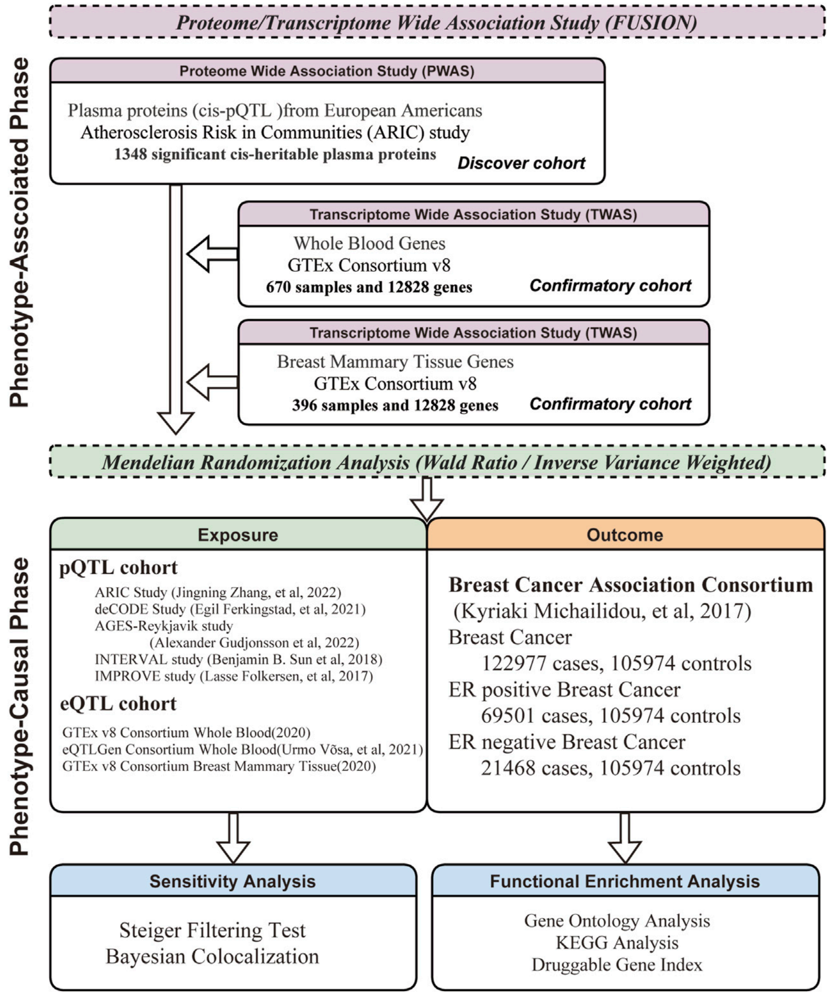

Proteome/Transcriptome Wide Association Study (FUSION)
Phenotype-Asscoiated Phase
Proteome Wide Association Study (PWAS)
Plasma proteins (cis-pQTL)from European Americans
Atherosclerosis Risk in Communities (ARIC) study
1348 significant cis-heritable plasma proteins
Discover cohort
Transcriptome Wide Association Study (TWAS)
Whole Blood Genes
GTEx Consortium v8
670 samples and 12828 genes
Confirmatory cohort
Transcriptome Wide Association Study (TWAS)
Breast Mammary Tissue Genes
GTEx Consortium v8
396 samples and 12828 genes
Confirmatory cohort
Mendelian Randomization Analysis (Wald Ratio / Inverse Variance Weighted)
Phenotype-Causal Phase
Exposure
pQTL cohort
ARIC Study (Jingning Zhang, et al, 2022)
deCODE Study (Egil Ferkingstad, et al, 2021)
AGES-Reykjavik study
(Alexander Gudjonsson et al, 2022)
INTERVAL study (Benjamin B. Sun et al, 2018)
IMPROVE study (Lasse Folkersen, et al, 2017)
eQTL cohort
GTEx v8 Consortium Whole Blood(2020)
eQTLGen Consortium Whole Blood(Urmo Võsa, et al, 2021)
GTEx v8 Consortium Breast Mammary Tissue(2020)
Outcome
Breast Cancer Association Consortium
(Kyriaki Michailidou, et al, 2017)
Breast Cancer
122977 cases, 105974 controls
ER positive Breast Cancer
69501 cases, 105974 controls
ER negative Breast Cancer
21468 cases, 105974 controls
Sensitivity Analysis
Steiger Filtering Test
Bayesian Colocalization
Functional Enrichment Analysis
Gene Ontology Analysis
KEGG Analysis
Druggable Gene Index

FIGURE 1
Framework of Comprehensive Research Methodology. This research methodology is divided into two phases: phenotype-association and phenotype-causation. Each phase follows a discovery-confirmatory approach.

utilized in our study. This dataset was comprised of 122,977 breast cancer cases and 105,974 controls. The same analytical approach was also applied to ER positive and negative breast cancer. The ER-positive subtype was found to consist of 69,501 cases and 105,974 controls, while the ER-negative subtype included 21,468 cases and 105,974 controls.

2.3 Quantity trait loci (QTL) dataset sources

Cis-pQTL data for European Americans' (EA) plasma proteins were obtained from the ARIC cohort (nilanjanchatterjeelab.org/pwas/), generated using PLINK2 software (Purcell et al., 2007). The SeqID file names correspond to the SOMAmers (Slow Off-rate Modified Aptamers), which are utilized for measuring protein levels in biological samples by leveraging their enhanced affinity and specificity for target proteins (Rohloff et al., 2014). For external validation, cis-pQTL data from 4 extensive plasma protein cohorts of European descent were used. Additionally, our study also explored expression quantitative trait loci (eQTLs), which influence gene expression at the transcriptome level (Zhu et al., 2016). We extracted eQTL data using the SMR toolkit (Wu et al., 2021), a tool specifically designed for genetic epidemiological research, from two major sources: the Genotype-Tissue Expression Project (GTEx Consortium, 2020) and the eQTLGen consortium (Võsa et al., 2021). Detailed descriptions of each dataset are provided in Table 1.

2.4 Proteome/transcriptome-wide association studies with fusion

FUSION (Boston, MA, United States) (Gusev et al., 2016), which is a software to establish associations between functional phenotype and GWAS phenotype, was used to conduct P/TWAS analysis. In our study, FUSION was implemented to identify associations between protein/gene expression levels and Breast Cancer susceptibility. Methodologically, FUSION takes two inputs: 1) Precomputed functional weights, and 2) GWAS summary statistics unified to a reference SNP panel. In PWAS,

TABLE 1 Detailed information about each GWAS summary data.

GWAS	Cohort	Paper title	Year	Author	PMID	Sample size	Protein/ Gene measured
pQTL							
Discovery	ARIC study	Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies	2022	Jingning Zhang, et al	35,501,419	7,213	4435 (1,318 in PWAS)
Confirmatory	Icelandic Cancer Project (52% of participants) and deCODE genetics (48% of participants)	Large-scale integration of the plasma proteome with genetics and disease	2021	Egil Ferkingstad, et al	34,857,953	35,559	4719
Confirmatory	AGES-Reykjavik study	A genome-wide association study of serum proteins reveals shared loci with common diseases	2022	Alexander Gudjonsson, et al	35,078,996	5,368	2091
Confirmatory	INTERVAL study	Genomic atlas of the human plasma proteome	2018	Benjamin B. Sun, et al	29,875,488	3,301	2,994
Confirmatory	IMPROVE study	Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease	2017	Lasse Folkersen, et al	28,369,058	3,394	83
eQTL							
Confirmatory	GTEx v8 Consortium Whole Blood	The GTEx Consortium atlas of genetic regulatory effects across human tissues	2020	GTEx Consortium	32,913,098	670	12,828
Confirmatory	eQTLGen Consortium Whole Blood	Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression	2021	Urmo Võsa, et al	34,475,573	31,684	16,987
Confirmatory	GTEx v8 Consortium Breast Mammary Tissue	The GTEx Consortium atlas of genetic regulatory effects across human tissues	2020	GTEx Consortium	32,913,098	396	12,828
Breast Cancer							
Overall Breast Cancer	Association analysis identifies 65 new breast cancer risk loci		2017	Kyriaki Michailidou, et al	29,059,683	228,951	
ER positive	Association analysis identifies 65 new breast cancer risk loci		2017	Kyriaki Michailidou, et al	29,059,683	175,475	
ER negative	Association analysis identifies 65 new breast cancer risk loci		2017	Kyriaki Michailidou, et al	29,059,683	127,442	

precomputed functional weights of plasma proteins were obtained from the ARIC study (Zhang et al., 2022b), and the reference SNP panel was derived from the European descent of the 1000G project (http://www.internationalgenome.org/faq/how-do-i-cite-1000-genomes-project). The primary outputs of FUSION are the Z-score and *p*-value, wherein Z-score quantifies the strength and direction of the associations between plasma proteins and breast cancer, while the *p*-value elucidates the statistical significance of this association. To enhance our findings at transcriptomic level, we incorporated TWAS for both whole blood and breast mammary tissues. The precomputed functional weights for TWAS, provided by Junghyun Jung from the Mancuso lab (http://gusevlab.org/projects/fusion/). A false discovery rate (FDR, Benjamini–Hochberg) threshold of 0.05 was applied to determine the statistical significance of the results.

2.5 Bayesian colocalization analysis and protein association classification

Bayesian colocalization analysis (Giambartolomei et al., 2014) was utilized to evaluate the probability that the same genetic variant affects both plasma protein and breast cancer. The default parameters set by the analysis were followed, including $p1$ = 10e−4 (the probability of a variant being a significant pQTL), $p2$ = 10e−4 (the probability of a variant associated with breast cancer), and $p12$ = 10e−5 (the probability of a variant being significant in both protein/gene and GWAS). This analysis involved five predefined hypotheses: H0, indicating no association with either trait; H1, signifying association with trait1 only; H2, implying association with trait2 only; H3, representing associations with both traits due to different SNPs;

and H4, indicating association with both traits due to a common SNP. A posterior probability of H4 (PPH4) exceeding 0.8, or in some cases 0.7, is generally interpreted as strong evidence of the same genetic variant being implicated in both traits (Giambartolomei et al., 2014).

Recent studies have investigated the causal associations between plasma proteins and diseases like colorectal cancer (Sun et al., 2023) and inflammatory bowel disease (Chen et al., 2023), utilizing a scoring system that integrates *p*-value and PPH4. Building on this approach, our research employs P/TWAS and Bayesian Colocalization analysis to systematically categorize the degrees of association between proteins. The scoring system was as follows: a significant adjusted *p*-value was awarded 1 point, and a PPH4 > 0.75 also earned 1 point. Based on the cumulative scores, associations were categorized as follows: a score between 1 and 2 indicated a “Weak” association, 3 to 4 suggested a “Moderate” association, and 5 to 6 signified a “Strong” association.

2.6 Mendelian Randomization and sensitivity analysis

In the causal analysis, we primarily conducted further analysis on proteins with strong and moderate associations. MR analysis were based on 3 essential assumptions for genetic instrumental variables: relevance, independence, and exclusion-restriction (Davies et al., 2018). We implemented a stringent selection process for SNPs to be used as instrumental variables, requiring a p < 5e-8, or p < 5e-6 in cases when SNP was absent. Clump was applied in accordance with the default parameters. The Wald Ratio (WR) method was employed when a single SNP was used as the instrumental variable, whereas the inverse-variance weighted (IVW) method was predominant when the instrumental variables involved multiple SNPs (Burgess et al., 2019). To reinforce the robustness of our findings, we conducted several sensitivity analyses. The Steiger filtering test (Deng et al., 2022) was utilized to eliminate the possibility of reverse causal associations. Additionally, heterogeneity and pleiotropy sensitivity analyses were conducted for proteins that met the criteria (Bowden et al., 2015; Greco et al., 2015). Furthermore, to improve the reliability and applicability of our results, external validation was carried out on pQTL data derived from 4 extensive plasma protein cohorts in European populations.

2.7 Enrichment analysis and potential druggable targets

To delve deeper into the intricate relationships and biological functions of significant proteins identified in our PWAS, gene ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses was performed. Given the emerging role of plasma proteins as potential therapeutic targets (Sun et al., 2018), we matched P/TWAS-MR significant proteins with the druggable genome database (Finan et al., 2017), which categorizes 4,479 genes into three druggability tiers: Tier 1 includes approved drugs and candidates in clinical trials, Tier 2 encompasses targets of biologically active molecules and those similar to approved drug targets, and Tier 3 comprises genes for secreted or extracellular proteins and other key druggable gene family members. Additionally, the significant proteins were annotated using the Therapeutic Target Database (http://db.idrblab.net/ttd/) (Zhou et al., 2022).

2.8 Statistical methods

In this study, data analysis was executed using R software (version 4.3.1). The P/TWAS analysis followed the analytical process previously described. The Benjamini–Hochberg method was employed for multiple testing correction, with adjusted *p*-values <0.05 considered statistically significant. Causations were investigated using the “TwoSampleMR” package, while Bayesian colocalization analysis was carried out using the “COLOC” package. The “ClusterProfiler” package (Wu et al., 2021) was utilized for functional enrichment analysis. Data visualization was achieved through the “Forestploter” and “ggplot2” packages, and data cleaning was performed using the “tidyverse” package.

3 Results

3.1 Identification of associations at the proteomic level

In our study, a total of 25 plasma proteins were significantly associated with breast cancer (Table 2; Figure 2A, and Supplementary Table S1). Of these proteins, 14 showed a Z-score greater than 0, denoting a positive association with breast cancer. Conversely, the remaining 11 proteins suggested an inverse association with the disease. When duplicate SOMAmers are present, we select the protein corresponding to the smallest *p*-value for subsequent analysis, such as RSPO3 (Supplementary Table S1). In ER subtypes analysis, 16 proteins were found to be significantly associated with ER-positive breast cancer and 6 with ER-negative breast cancer (Supplementary Table S2, S3). The PWAS Manhattan plot illustrates the distribution of significant genes across different chromosomes and their respective *p*-value (Figure 3A, Supplementary Figure S2A, B).

3.2 Identification of associations at the transcriptomic level

For the 25 proteins identified by PWAS, 12 showed significant associations in the whole blood TWAS analysis ($P.adj$ < 0.05) (Supplementary Table S4). While, in the breast mammary tissue TWAS, 10 of these proteins were further validated ($P.adj$ < 0.05) (Table 2, Supplementary Table S5). Among the 16 significant proteins in ER-positive breast cancer, with 7 were confirmed in both whole blood and breast mammary tissue analyses. Meanwhile, in ER-negative breast cancer, 2 out of the 6 significant proteins were validated (Figures 2B,C, Supplementary Table S6). It should be noted that MST1 exhibited contradictory associations in PWAS (Z = 4.194, $P.adj$ = 0.004) and TWAS (Z = −2.547, $P.adj$ = 0.014). This pattern was also observed in ER-positive and ER-negative subtypes. Due to the complex nature and potential biological

TABLE 2 Integrative analysis and stratification of proteome and transcriptome associations in breast cancer.

Gene	CHR	Plasma protein PWAS (discovery cohort)			Whole blood TWAS (confirmatory cohort)			Breast tissue TWAS (confirmatory cohort)			Score	Association power
		Zscore	*P*_FDR	PPH4	Zscore	*P*_FDR	PPH4	Zscore	*P*_FDR	PPH4		
PGD	1	−9.152	7.39E-17	0.994	-	-	-	-	-	-	2	Weak
TLR1	4	6.225	3.18E-07	1	-	-	-	-	-	-	2	Weak
FBLN5	14	5.226	7.60E-05	0	0.460	0.6460	0	-	-	-	1	Weak
PEX14	1	4.839	0.0004	0.989	6.843	1.09E-10	0.195	5.341	8.14E-07	0.79	5	Strong
LAYN	11	4.499	0.0018	0.936	-	-	-	-	-	-	2	Weak
SNUPN	15	−4.413	0.0022	0.952	−5.285	8.82E-07	0.939	−5.256	8.14E-07	0.372	5	Strong
GSTM4	1	−4.273	0.0036	0.618	−3.451	0.0011	0.001	−3.736	0.0004	0.175	3	Moderate
MST1	3	4.194	0.0045	0.904	−2.547	0.0139	0.148	−3.266	0.0015	0.584	4	Moderate (inconsistent)
CSK	15	−4.147	0.0049	0.779	−3.979	0.0002	0.843	−4.613	1.46E-05	0.863	6	Strong
NTN4	12	3.938	0.0108	0	-	-	-	-	-	-	1	Weak
PAPPA	9	−3.692	0.0255	0.137	-	-	-	-	-	-	1	Weak
CTSF	11	3.681	0.0255	0.777	2.895	0.0059	0.944	4.394	0.0000	0.939	6	Strong
PARK7	1	−3.648	0.0269	0.978	−4.644	1.20E-05	0.965	−2.391	0.0185	0.014	5	Strong
NCF1	7	3.601	0.0296	0.002	4.419	2.78E-05	0.958	-	-	-	3	Moderate
COL6A3	2	−3.585	0.0296	0.026	-	-	-	-	-	-	1	Weak
RSPO3	6	−3.541	0.0319	0.272	-	-	-	-	-	-	1	Weak
HEBP1	12	3.533	0.0319	0.058	2.358	0.0215	0.015	-	-	-	2	Weak
NRP1	10	−3.508	0.0328	0.038	-	-	-	-	-	-	1	Weak
ABO	9	−3.442	0.0369	0.298	−4.953	3.41E-06	0.169	2.127	0.0334	0.994	4	Moderate
PRDX1	1	3.436	0.0369	0.004	1.993	0.0498	0.005	3.392	0.0011	0.03	3	Moderate
EMILIN3	20	3.410	0.0369	0.072	-	-	-	-	-	-	1	Weak
ANXA4	2	3.404	0.0369	0.681	2.820	0.0067	0.008	-	-	-	2	Weak
POSTN	13	3.401	0.0369	0.01	-	-	-	-	-	-	1	Weak
LDHA	11	−3.352	0.0424	0.404	−3.307	0.0016	0.416	−3.625	0.0005	0.291	3	Moderate
UROD	1	3.302	0.0487	0.254	-	-	-	2.834	0.0056	0.113	2	Weak

implications of MST1’s contrasting results, we did not conduct further analysis on this protein.

3.3 Bayesian Colocalization analysis

Among 25 significant proteins, 9 exhibited strong genetic colocalization evidence. Additionally, 4 proteins–SNUPN (PPH4 = 93.9%), CSK (PPH4 = 84.3%), CTSF (PPH4 = 94.4%), and PARK7 (PPH4 = 96.5%)–also demonstrated the same strong genetic evidence at the whole blood transcriptomic level. Remarkably, CSK (PPH4 = 86.3%) and CTSF (PPH4 = 93.9%) were further validated in the breast mammary tissue transcriptomic level (Table 2). In ER-positive breast cancer, 5 proteins showed strong evidence of genetic colocalization. Notably, 2 of these proteins, CSK (PPH4 = 85.8%, 86.5%) and GDI2 (PPH4 = 97%, 97.2%), demonstrated the same strong genetic colocalization evidence in both whole blood and breast mammary tissues. In the ER-negative breast cancer, PEX14 showed strong genetic colocalization evidence in protein (PPH4 = 99.9%) and breast mammary tissue (PPH4 = 88.8%), but this pattern was not replicated at the whole blood transcriptomic level (PP4 = 8%, Supplementary Table S6).

3.4 Stratification of plasma protein association strengths

In breast cancer, 25 proteins were classified: 5 as “Strong” association (red), 6 as “Moderate” association (blue), and the

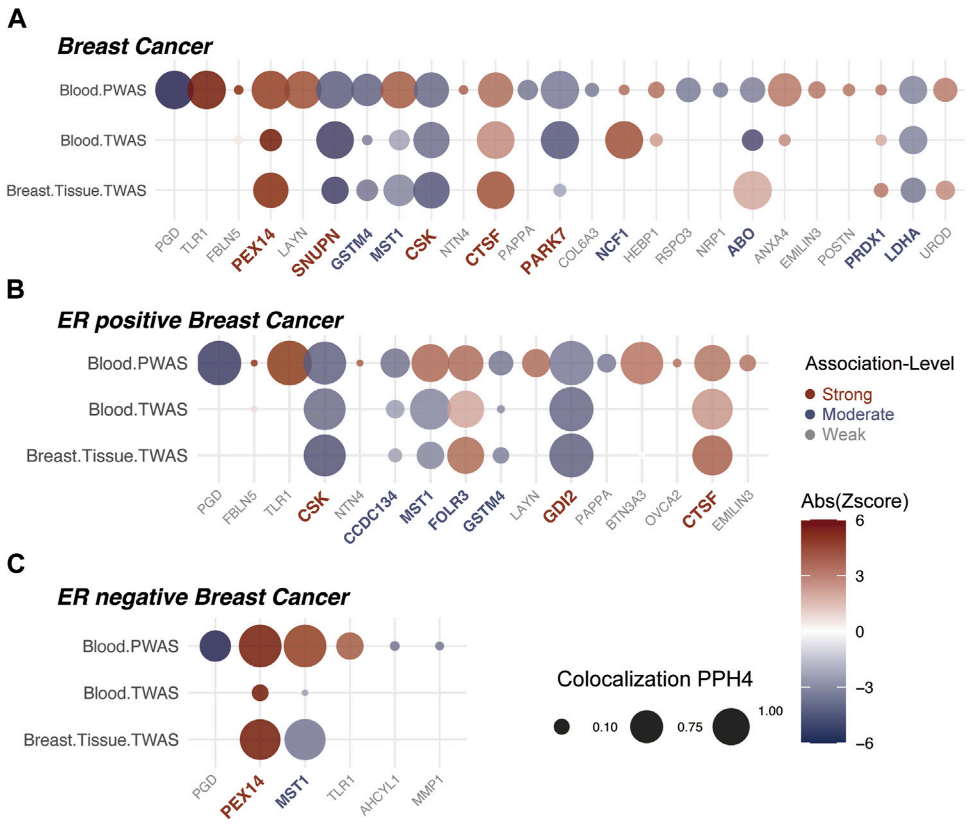


FIGURE 2
Insights from Proteome/Transcriptome-Wide Association Analyses and Bayesian Colocalization. The association strength of proteins with breast cancer is denoted by colors: red for strong, blue for moderate, and grey for weak associations. **(A)** Comprehensive P/TWAS for plasma proteins in breast cancer susceptibility. Dot size signifies results from Bayesian Colocalization analysis, with color gradient reflecting the Z-value. Proteins are sequentially arranged based on ascending p-value significance from left to right. **(B)** Comprehensive P/TWAS for plasma proteins in ER positive breast cancer susceptibility. **(C)** Comprehensive P/TWAS for plasma proteins in ER negative breast cancer susceptibility.

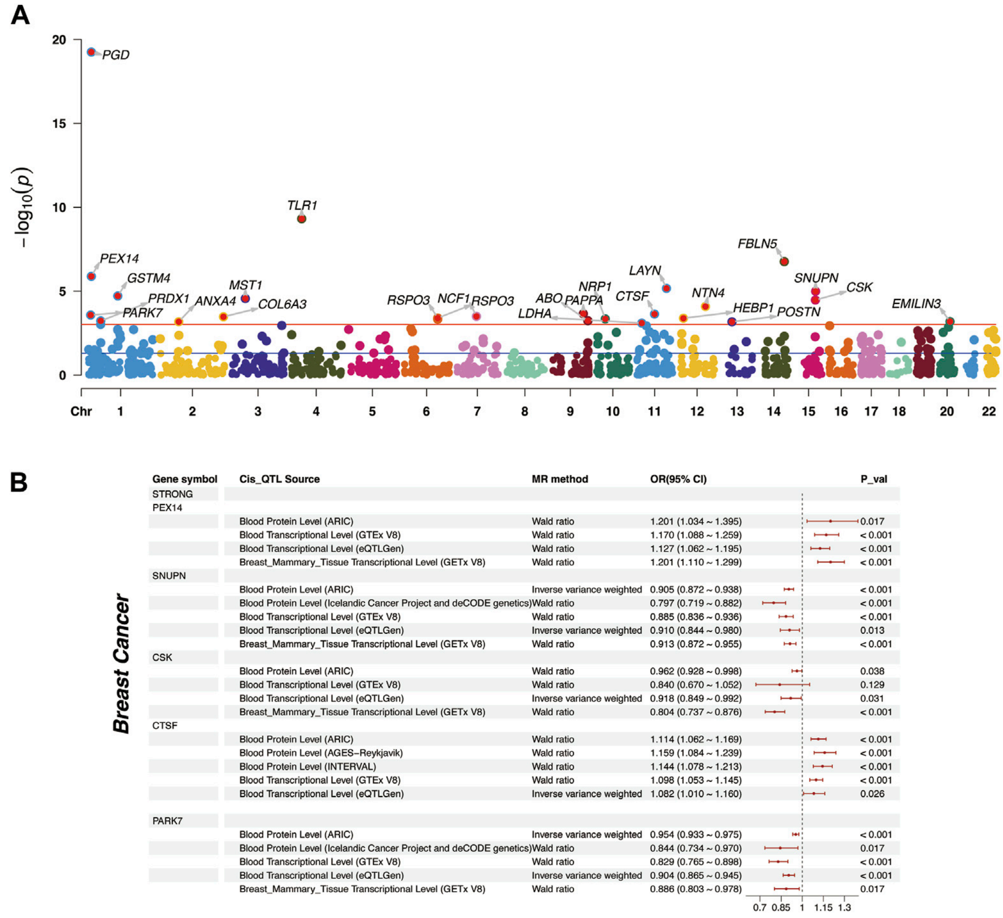
remaining as "Weak" association (grey) (Figure 2A). Among the "Strong" Tiers, PEX14 (Z = 4.839) and CTSF (Z = 3.681) had a positive association with breast cancer. Whereas, SNUPN (Z = −4.413), CSK (Z=−4.417), and PARK7 (Z = −3.648) showed negative associations (Table 2).

In ER-positive breast cancer, 16 proteins were classified: 3 proteins showed "Strong" association (red), with GDI2 (Z = −3.652) newly identified and negatively associated with ER-positive breast cancer. Additionally, CSK and CTSF followed the same trends with the findings from breast cancer. Besides, 4 proteins were "Moderate" (blue), and 9 proteins were "Weak" associations (grey) (Figure 2B). In ER-negative breast cancer, 6 proteins were classified: PEX14 and MST1 showed "Strong" associations. Notably, PEX14 not only showed the same trend as observed in breast cancer (Z = 4.839, p = 0.0004) but also exhibited a notably stronger effect (Z = 5.929, p = 2.02E-6). MST1 was not further analyzed due to inconsistent trends in P/TWAS. The other 4 proteins were categorized as "Weak" association (grey) (Figure 2C, Supplementary Table S6).

It is crucial to highlight that, although PGD and TLR1 were significant across all three outcomes in PWAS analyses (Figure 3A and Supplementary Figure S2A, B), their absence from the corresponding TWAS analysis relegated them to the "Weak" association. Moreover, the results of these two proteins were not sufficiently reliable in MR Analysis (Supplementary Figure S2C, Supplementary Table S7).

3.5 Mendelian Randomization analyses

Upon determining the strength of associations, we supplemented the causations with MR analysis (Supplementary Table S8). We primarily focused on the causal effects of "Strong" associated proteins. Among the 5 "Strong" associated proteins, PEX14 was found to have a positive causation at the proteomic (OR = 1.201, p = 0.017) and transcriptomic level (OR = 1.17, p < 0.001). CTSF demonstrated a positive causation in the ARIC cohort (OR = 1.114, p < 0.001), and the consistent trends were also external validated in INTERVAL cohort (OR = 1.144, p < 0.001) (Sun et al., 2018) and AGES-Reykjavik cohort (OR = 1.159, p < 0.001) (Gudjonsson et al., 2022) (Figure 3B). The remaining 3 proteins, SNUPN (OR = 0.905, p < 0.001), CSK (OR = 0.962, p = 0.038), and PARK7 (OR = 0.954, p < 0.001), all exhibited negative causations with breast cancer. External validations from the deCODE cohort further confirmed the causations for SNUPN (OR = 0.797, p < 0.001) and PARK7 (OR = 0.844, p = 0.017). However, CSK's causation at the whole blood transcriptomic level was somewhat unsignificant (OR = 0.84, p = 0.129) (Figure 3B, Supplementary Table S9).



Gene symbol	Cis_QTL Source	MR method	OR(95% CI)	P_val
STRONG				
PEX14				
	Blood Protein Level (ARIC)	Wald ratio	1.201 (1.034 ~ 1.395)	0.017
	Blood Transcriptional Level (GTEx V8)	Wald ratio	1.170 (1.088 ~ 1.259)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Wald ratio	1.127 (1.062 ~ 1.195)	< 0.001
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	1.201 (1.110 ~ 1.299)	< 0.001
SNUPN				
	Blood Protein Level (ARIC)	Inverse variance weighted	0.905 (0.872 ~ 0.938)	< 0.001
	Blood Protein Level (Icelandic Cancer Project and deCODE genetics)	Wald ratio	0.797 (0.719 ~ 0.882)	< 0.001
	Blood Transcriptional Level (GTEx V8)	Wald ratio	0.885 (0.836 ~ 0.936)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	0.910 (0.844 ~ 0.980)	0.013
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	0.913 (0.872 ~ 0.955)	< 0.001
CSK				
	Blood Protein Level (ARIC)	Wald ratio	0.962 (0.928 ~ 0.998)	0.038
	Blood Transcriptional Level (GTEx V8)	Wald ratio	0.840 (0.670 ~ 1.052)	0.129
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	0.918 (0.849 ~ 0.992)	0.031
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	0.804 (0.737 ~ 0.876)	< 0.001
CTSF				
	Blood Protein Level (ARIC)	Wald ratio	1.114 (1.062 ~ 1.169)	< 0.001
	Blood Protein Level (AGES–Reykjavik)	Wald ratio	1.159 (1.084 ~ 1.239)	< 0.001
	Blood Protein Level (INTERVAL)	Wald ratio	1.144 (1.078 ~ 1.213)	< 0.001
	Blood Transcriptional Level (GTEx V8)	Wald ratio	1.098 (1.053 ~ 1.145)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	1.082 (1.010 ~ 1.160)	0.026
PARK7				
	Blood Protein Level (ARIC)	Inverse variance weighted	0.954 (0.933 ~ 0.975)	< 0.001
	Blood Protein Level (Icelandic Cancer Project and deCODE genetics)	Wald ratio	0.844 (0.734 ~ 0.970)	0.017
	Blood Transcriptional Level (GTEx V8)	Wald ratio	0.829 (0.765 ~ 0.898)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	0.904 (0.865 ~ 0.945)	< 0.001
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	0.886 (0.803 ~ 0.978)	0.017

FIGURE 3
Distribution of plasma proteins and Mendelian Randomization analysis of “Strong” plasma proteins. **(A)** The Manhattan plot represented plasma proteins with significant affiliations to breast cancer. The red horizontal line indicates the FDR corrected *p*-value threshold for significance. Chromosomal designations populate the horizontal axis, contrasted with respective -log10 *p*-values on the vertical spectrum. **(B)** Two-sample Mendelian Randomization analysis for “Strong” plasma proteins to breast cancer, including external validation at proteomic and transcriptomic levels.

In ER-positive breast cancer, CSK (OR = 0.955, $p = 0.038$) and CTSF (OR = 1.125, $p < 0.001$) maintained the same causal trends as observed in breast cancer (Supplementary Table S10). Additionally, GDI2 was identified as a newly negatively significant protein (OR = 0.92, $p < 0.001$). However, its causal effect was not significant at the transcriptomic level (OR = 1.001, $p < 0.981$, Figure 4A). In ER-negative breast cancer, PEX14 stood out as the sole “Strong” protein. Notably, its causal effect in this subtype (OR = 1.645, $p < 0.001$, Figure 4B) was further pronounced compared to breast cancer (OR = 1.201, $p = 0.017$). Meanwhile, we expanded our MR analyses to include “Moderate” proteins. The results revealed that their causal effects were generally less consistent and of reduced significance compared to those of the “Strong” proteins (Supplementary Figure S1, Supplementary Table S11).

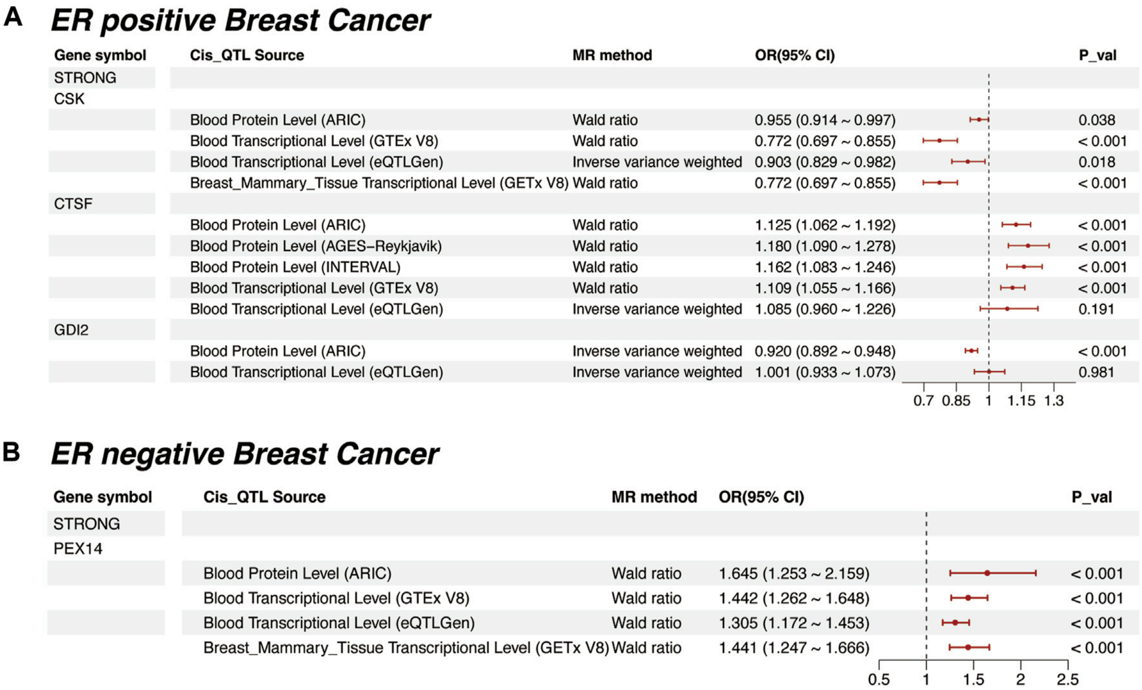
3.6 Sensitivity analysis and functional enrichment analysis

Considering that the pQTLs of most plasma proteins was a single SNP, conducting sensitivity analyses for heterogeneity and pleiotropy is typically not required. As result, in ER-positive breast cancer, BTN3A3, EMILIN3, FOLR3, and NTN4 showed heterogeneity, while in ER-negative cases, this was not observed (Supplementary Table S10). BTN3A3 in ER-positive breast cancer also displayed pleiotropy. The Steiger filtering test confirmed that MR effects were due to plasma proteins affecting breast cancer outcomes (Supplementary Table S8, S10). Importantly, our “Strong” proteins exhibited neither heterogeneity nor pleiotropy.

Furthermore, the plasma proteins identified by PWAS were subjected to Gene Ontology (GO) cluster analysis. This analysis revealed a predominant association with biological processes related to oxidative stress, such as “reactive oxygen species metabolic” and “response to reactive oxygen species” terms. Additionally, for cellular components, we observed a significant enrichment in the “collagen-containing extracellular matrix” term (Supplementary Figure S3). Besides, KEGG pathway enrichment did not reveal any significantly enriched pathways (Supplementary Table S12).

3.7 Druggable target propensity for significant proteins

Plasma proteins are not only crucial as diagnostic biomarkers but also serve as potential drug targets. In our study, we evaluated



A ER positive Breast Cancer

Gene symbol	Cis_QTL Source	MR method	OR(95% CI)	P_val
STRONG				
CSK				
	Blood Protein Level (ARIC)	Wald ratio	0.955 (0.914 ~ 0.997)	0.038
	Blood Transcriptional Level (GTEx V8)	Wald ratio	0.772 (0.697 ~ 0.855)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	0.903 (0.829 ~ 0.982)	0.018
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	0.772 (0.697 ~ 0.855)	< 0.001
CTSF				
	Blood Protein Level (ARIC)	Wald ratio	1.125 (1.062 ~ 1.192)	< 0.001
	Blood Protein Level (AGES−Reykjavik)	Wald ratio	1.180 (1.090 ~ 1.278)	< 0.001
	Blood Protein Level (INTERVAL)	Wald ratio	1.162 (1.083 ~ 1.246)	< 0.001
	Blood Transcriptional Level (GTEx V8)	Wald ratio	1.109 (1.055 ~ 1.166)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	1.085 (0.960 ~ 1.226)	0.191
GDI2				
	Blood Protein Level (ARIC)	Inverse variance weighted	0.920 (0.892 ~ 0.948)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Inverse variance weighted	1.001 (0.933 ~ 1.073)	0.981

B ER negative Breast Cancer

Gene symbol	Cis_QTL Source	MR method	OR(95% CI)	P_val
STRONG				
PEX14				
	Blood Protein Level (ARIC)	Wald ratio	1.645 (1.253 ~ 2.159)	< 0.001
	Blood Transcriptional Level (GTEx V8)	Wald ratio	1.442 (1.262 ~ 1.648)	< 0.001
	Blood Transcriptional Level (eQTLGen)	Wald ratio	1.305 (1.172 ~ 1.453)	< 0.001
	Breast_Mammary_Tissue Transcriptional Level (GETx V8)	Wald ratio	1.441 (1.247 ~ 1.666)	< 0.001

FIGURE 4
Mendelian Randomization for "Strong" plasma proteins in different ER Breast Cancer Subtypes. **(A)** Mendelian randomization results for ER-positive breast cancer, including external validation at proteomic and transcriptomic levels. **(B)** Mendelian randomization results for ER-negative breast cancer, including external validation at proteomic and transcriptomic levels.

the significant proteins for their potential as drug targets. By aligning our findings with the druggable genome database (Finan et al., 2017), we determined that 16 of the 25 proteins have druggable targets. These include 3 proteins in Tier 1; 3 in Tier 2, and 10 in Tier 3 (Supplementary Table S13, Left column). Furthermore, we compared our results with the Therapeutic Target Database (Zhou et al., 2022), 11 of these 16 proteins were identified as targets of existing or potential drugs. This group comprised 3 Successful targets, 3 Patented-recorded Targets, 1 in clinical trials, and 4 documented in literature (Supplementary Table S13 Right column). Among the "Strong" proteins, CSK and CTSF were found to be drug targets with patent records, categorized under Tiers 1 and 2 respectively. CTSF has been documented to be used in the treatment of bone cancer and chronic obstructive pulmonary disease (Li et al., 2017) (Table 3). However, the remaining "Strong" proteins have not yet been reported.

4 Discussion

Plasma proteins, due to their ease of detection and reproducibility, are increasingly utilized to distinguish between cancer patients and healthy individuals, enhancing the effectiveness of screening programs (Huijbers et al., 2010). Recent advancements in molecular technologies and techniques have shown significant potential in utilizing plasma protein biomarkers such as Adipsin and CA15-3 for early detection and quantification for diagnostic and therapeutic applications in breast cancer (Afzal et al., 2022; Rajkumar et al., 2022; Veyssière et al., 2022). A recent high-throughput study identified 61 proteins associated with various cancers (Gregga et al., 2023). While this study provided valuable insights into pan-cancer associations, it did not explore causation. Furthermore, research specifically targeting plasma protein biomarkers for breast cancer is still limited. Currently, Mendelian Randomization has emerged as an effective method to establish causation in various diseases (Emdin et al., 2017), including cholesterol-related cardiovascular disease (Kathiresan et al., 2008), inflammatory diseases (Swerdlow et al., 2012), metabolic disorders (Fall et al., 2015), and specific cancers such as small cell lung cancer and colorectal cancer (Sun et al., 2023; Wu et al., 2023).

Despite, the application of MR in identifying plasma proteins as drug targets in breast cancer is still sporadic. For instance, one study performed MR analysis on a single cohort of 732 plasma proteins, where GDI2 and CTSF were identified as potential targets for breast cancer (Ren et al., 2023), aligning with our research. However, it is important to note that this study also focused on pan-cancer research and lacked association analysis. Additionally, another study focused on the causation found a causal link between TLR1 and breast cancer (Mälarstig et al., 2023). This protein was ranked significantly in our analysis, but it is noteworthy that TLR1 lacks external cohort validation, and the study also did not perform association analyses. Therefore, current research on plasma proteins typically focuses on either association or causation, rarely addressing both. Our study bridges this gap by integrating these two approaches. We employed P/TWAS to identify associations and used MR to establish causation. This approach successfully pinpointed significant proteins related to breast cancer risk from

TABLE 3 Comprehensive evaluation of strong associated proteins as potential druggable targets or existing therapeutics.

Gene	UniProt	Description	Finan et al	Therapeutic target database		
			Tier	Target type	Drug name	Disease
Breast Cancer						
PEX14	O75381	Peroxisomal Biogenesis Factor 14	-	-	-	-
SNUPN	O95149	Snurportin 1	-	-	-	-
CSK	P41240	Tyrosine-protein kinase CSK	Tier 1	Patented-recorded Target	936,563-93-8	Not Available
CTSF	Q9UBX1	cathepsin F	Tier 2	Patented-recorded Target	PMID27998201-Compound-5	Bone cancer; Chronic obstructive pulmonary disease
PARK7	Q99497	Parkinsonism Associated Deglycase	-	-	-	-
ER pos Breast Cancer						
CSK	P41240	Tyrosine-protein kinase CSK	Tier 1	Patented-recorded Target	936,563-93-8	Not Available
CTSF	Q9UBX1	cathepsin F	Tier 2	Patented-recorded Target	PMID27998201-Compound-5	Bone cancer; Chronic obstructive pulmonary disease
GDI2	P50395	Rab GDP dissociation inhibitor beta	-	-	-	-
ER neg Breast Cancer						
PEX14	O75381	Peroxisomal Biogenesis Factor 14	-	-	-	-

thousands of candidates in 5 large proteomics cohorts. To ensure the robustness and generalizability of our findings, we adopted a "discovery-confirmatory" analytical framework at both the association and causation phases. Overall, we found 5 proteins (PEX14, CTSF, SNUPN, CSK, PARK7) with strong causal links to breast cancer. While, in ER-positive breast cancer, 3 proteins (CSK, CTSF, GDI2) were identified. In contrast, only PEX14 was linked to ER-negative breast cancer.

Among the 5 plasma proteins, SNUPN, CSK, and PARK7 emerged as "Strong" negatively causative associated proteins, indicating a protective effect against breast cancer development. A study has highlighted the potential clinical applications of SNUPN in acute lymphoblastic leukemia (Mata-Rocha et al., 2019); however, research exploring its role in solid tumors, including breast cancer, is currently limited. Despite the current research limitations, SNUPN's potential as a biomarker or tumor suppressor is promising and warrants further exploration. PARK7 is recognized for its neuroprotective role in Parkinson's disease (Kochmanski et al., 2022) and has been reported to significantly regulate cell survival and cancer progression in various cancers (Jin, 2020). It negatively regulates PTEN and PKB/Akt phosphorylation, thus influencing cell survival and death (Kim et al., 2005). In breast cancer, low PARK7 expression was correlated with pathological complete response in 79.6% of cases following neoadjuvant therapy (Kawate et al., 2013), and loss of PARK7 function is associated with increased sensitivity to doxorubicin in breast cancer cells (Zhang et al., 2015). The effect of PARK7 in balancing tumor cell survival and normal cell physiology merits further research. Lastly, as a key member of the Src family kinases (SFKs), CSK plays a vital role in combating cancer progression in various cancers (Sabe et al., 1994). Recent study indicates that CSK maintains negative regulation of Src through Tyr527 phosphorylation, inhibiting breast cancer cells growth and spread (Dias et al., 2022). Additionally, another study on ER-positive breast cancer found that in cases of endocrine therapy resistance, reduced CSK leads to enhanced PAK2 activity and subsequent non-estrogen-dependent cancer growth (Xiao et al., 2018). The dual effect of CSK in both tumor suppression and inducing endocrine treatment resistance positions it as a notable target for research.

The other two "Strong" proteins are positively associated and represent a risk factor in breast cancer onset. CTSF (cathepsin F) plays a key role in the lysosomal protein degradation pathway (Wex et al., 1999). Currently, it is reported as an effective diagnostic biomarker in cervical cancer (Vazquez-Ortiz et al., 2005), gastric cancer (Ji et al., 2018), and non-small cell lung cancer (Wei et al., 2022). A recent study reported that CTSF may act as an independent poor prognostic factor for basal-like breast cancer (Huang et al., 2021). PEX14 (Peroxisomal Biogenesis Factor 14) is essential for peroxisomal biogenesis (Neufeld et al., 2009). Our research reveals a significant causal risk association of PEX14 with breast cancer (OR = 1.201), particularly in ER-negative subtype (OR = 1.645). Notably, PEX14 has been identified as a key risk factor in triple-negative breast cancer (TNBC) (Purrington et al., 2014) and is one of the top five genes influencing adaptive anti-tumor immunity, as shown in a TNBC model study using a whole-genome RNAi screening platform

(Shuptrine et al., 2017). These insights emphasize PEX14's importance in TNBC immunotherapy and drug target research. Furthermore, PEX14 plays a crucial role in maintaining peroxisomal functions, and its deficiency leads to ROS accumulation, lipid peroxidation, and consequent cell death (Guo et al., 2023). Our functional enrichment analysis corroborates this, highlighting numerous pathways related to reactive oxygen species (ROS), which are instrumental in promoting cell growth, cancer progression, immune responses, and poorer survival outcomes in breast cancer (Oshi et al., 2022). Additionally, studies have shown that PEX14 knockdown increases intracellular H_2O_2 levels, triggering ferroptosis and cell death (Guan et al., 2022). This further underscores PEX14's pivotal role in managing oxidative stress and cell viability, marking its significance in breast cancer research. Additionally, GDI2 was identified as a protein with a "Strong" negative causal association in the ER-positive breast cancer. A study suggested that GDI2 is associated with aggressive features and poor patient survival in hepatocellular carcinoma (Zhang et al., 2021). However, the inability to confirm its role through at additional transcriptomic levels and the absence of external validation has diminished our confidence in the significance of this protein.

Given the proven effectiveness of MR in identifying drug targets (Folkersen et al., 2020), we performed a drug-target evaluation on these plasma proteins (Supplementary Table S13). Notably, CSK and CTSF emerged as Tier1 and Tier2 proteins, respectively. CSK is crucial in regulating cellular processes such as apoptosis, survival, and proliferation. Its pivotal role in cancer cell signaling earmarks CSK as a promising target for cancer therapy (Fortner et al., 2022). Similarly, CTSF, known for its significant involvement in the progression of various cancers (Wei et al., 2022), neurodegenerative diseases (van der Zee et al., 2016), and skin aging (Takaya et al., 2023), garners attention. Research on inhibitors and modulators targeting CTSF is underway. Although other strongly associated proteins currently lack clear therapeutic applications, given their strong causal relationship with breast cancer, it is worthwhile to further explore them for drug target development.

This study is currently subject to several limitations yet. First, the study only involves individuals of European descent, which necessitates caution when applying these findings to more diverse populations. Second, the precomputed functional weights for plasma proteins are currently only available from the ARIC cohort, future datasets expansion are expected to enhance the precision and breadth of such analyses. In addition, as the current BCAC molecular subtype data lacks rsID, matching chromosomes and base pair positions results in significant information loss. However, with the continuous expansion and updating of the molecular subtype database, we anticipate a deeper understanding of this content. Lastly, our analysis is primarily data-based, hence we will design related basic scientific research in the future to further investigate the etiological association between plasma proteins and breast cancer.

In summary, our study successfully identified several plasma proteins with strong association and causation to breast cancer and its distinct ER subtypes. As non-invasive and dynamic monitoring tools, plasma proteins hold significant potential as diagnostic biomarkers and therapeutic targets. They offer a comprehensive perspective on systemic health, which is crucial for early tumor detection, assessing treatment responses, and continuous disease monitoring. While these advancements are still in the early stages, they hold valuable promise for future research and practical applications in real-world scenarios.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding authors.

Author contributions

YW: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. KY: Data curation, Formal Analysis, Investigation, Methodology, Writing–original draft. BC: Data curation, Formal Analysis, Investigation, Methodology, Validation, Writing–review and editing. BZ: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing–review and editing.GJ: Funding acquisition, Project administration, Resources, Supervision, Validation, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by Beijing Municipal Natural Science Foundation (No.7222145) and Shenzhen Key Medical Discipline Construction Fund (SZXK095).

Acknowledgments

We would like to extend our heartfelt thanks to all the data contributors of the plasma protein cohorts referenced in our study. The pertinent literature has been cited within the article. Our sincere appreciation goes out to the participants for their commitment and to the researchers for their altruistic contributions. The breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research United Kingdom (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). All studies and funders are listed in Michailidou et al. (2017). We extend our profound gratitude to Junghyun Jung and the Mancuso Lab for their invaluable contribution to our study through the provision of precomputed TWAS models. Their contributions have been pivotal in the success of our research endeavors.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1340917/full#supplementary-material

References

Afzal, S., Hassan, M., Ullah, S., Abbas, H., Tawakkal, F., and Khan, M. A. (2022). Breast cancer; discovery of novel diagnostic biomarkers, drug resistance, and therapeutic implications. *Front. Mol. Biosci.* 9, 783450. doi:10.3389/fmolb.2022.783450

Anderson, N. L., and Anderson, N. G. (2002). The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell Proteomics* 1 (11), 845–867. doi:10.1074/mcp.r200007-mcp200

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., et al. (2019). Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Res.* 4, 186. doi:10.12688/wellcomeopenres.15555.2

Chen, J., Xu, F., Ruan, X., Sun, J., Zhang, Y., Zhang, H., et al. (2023). Therapeutic targets for inflammatory bowel disease: proteome-wide Mendelian randomization and colocalization analyses. *EBioMedicine* 89, 104494. doi:10.1016/j.ebiom.2023.104494

Christopoulos, P. F., Msaouel, P., and Koutsilieris, M. (2015). The role of the insulin-like growth factor-1 system in breast cancer. *Mol. Cancer* 14, 43. doi:10.1186/s12943-015-0291-7

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *Bmj* 362, k601. doi:10.1136/bmj.k601

Deng, Y. T., Ou, Y. N., Wu, B. S., Yang, Y. X., Jiang, Y., Huang, Y. Y., et al. (2022). Identifying causal genes for depression via integration of the proteome and transcriptome from brain and blood. *Mol. Psychiatry* 27 (6), 2849–2857. doi:10.1038/s41380-022-01507-9

Dias, R. V. R., Ferreira, C. T. A., Jennings, P. A., Whitford, P. C., and Oliveira, L. C. (2022). Csk αC helix: a computational analysis of an essential region for conformational transitions. *J. Phys. Chem. B* 126 (50), 10587–10596. doi:10.1021/acs.jpcb.2c05408

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *Jama* 318 (19), 1925–1926. doi:10.1001/jama.2017.17219

Fall, T., Hägg, S., Ploner, A., Mägi, R., Fischer, K., Draisma, H. H., et al. (2015). Age- and sex-specific causal effects of adiposity on cardiovascular risk factors. *Diabetes* 64 (5), 1841–1852. doi:10.2337/db14-0988

Ferkingstad, E., Sulem, P., Atlason, B. A., Sveinbjornsson, G., Magnusson, M. I., Styrmisdottir, E. L., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* 53 (12), 1712–1721. doi:10.1038/s41588-021-00978-w

Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., et al. (2017). The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9 (383), eaag1166. doi:10.1126/scitranslmed.aag1166

Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R. J., Frånberg, M., Sennblad, B., et al. (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* 13 (4), e1006706. doi:10.1371/journal.pgen.1006706

Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D. H., Hedman Å, K., Schork, A., et al. (2020). Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* 2 (10), 1135–1148. doi:10.1038/s42255-020-00287-2

Fortner, A., Chera, A., Tanca, A., and Bucur, O. (2022). Apoptosis regulation by the tyrosine-protein kinase CSK. *Front. Cell Dev. Biol.* 10, 1078180. doi:10.3389/fcell.2022.1078180

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10 (5), e1004383. doi:10.1371/journal.pgen.1004383

Greco, M. F., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat. Med.* 34 (21), 2926–2940. doi:10.1002/sim.6522

Gregga, I., Pharoah, P. D. P., Gayther, S. A., Manichaikul, A., Im, H. K., Kar, S. P., et al. (2023). Predicted proteome association studies of breast, prostate, ovarian, and endometrial cancers implicate plasma protein regulation in cancer susceptibility. *Cancer Epidemiol. Biomarkers Prev.* 32 (9), 1198–1207. doi:10.1158/1055-9965.Epi-23-0309

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369 (6509), 1318–1330. doi:10.1126/science.aaz1776

Guan, H., Zhou, P., Qi, Y., Huang, H., Wang, J., and Liu, X. (2022). Cigarette smoke-induced trophoblast cell ferroptosis in rat placenta and the effects of L-arginine intervention. *Ecotoxicol. Environ. Saf.* 243, 114015. doi:10.1016/j.ecoenv.2022.114015

Gudjonsson, A., Gudmundsdottir, V., Axelsson, G. T., Gudmundsson, E. F., Jonsson, B. G., Launer, L. J., et al. (2022). A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat. Commun.* 13 (1), 480. doi:10.1038/s41467-021-27850-z

Guo, Y., Zhou, P., Qiao, L., Guan, H., Gou, J., and Liu, X. (2023). Maternal protein deficiency impairs peroxisome biogenesis and leads to oxidative stress and ferroptosis in liver of fetal growth restriction offspring. *J. Nutr. Biochem.* 121, 109432. doi:10.1016/j.jnutbio.2023.109432

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48 (3), 245–252. doi:10.1038/ng.3506

Huang, K., Wu, Y., Xie, Y., Huang, L., and Liu, H. (2021). Analyzing mRNAsi-related genes identifies novel prognostic markers and potential drug combination for patients with basal breast cancer. *Dis. Markers* 2021, 4731349. doi:10.1155/2021/4731349

Huijbers, A., Velstra, B., Dekker, T. J., Mesker, W. E., van der Burgt, Y. E., Mertens, B. J., et al. (2010). Proteomic serum biomarkers and their potential application in cancer screening programs. *Int. J. Mol. Sci.* 11 (11), 4175–4193. doi:10.3390/ijms11114175

Ji, C., Zhao, Y., Kou, Y. W., Shao, H., Guo, L., Bao, C. H., et al. (2018). Cathepsin F knockdown induces proliferation and inhibits apoptosis in gastric cancer cells. *Oncol. Res.* 26 (1), 83–93. doi:10.3727/096504017x14928634401204

Jin, W. (2020). Novel insights into PARK7 (DJ-1), a potential anti-cancer therapeutic target, and implications for cancer progression. *J. Clin. Med.* 9 (5), 1256. doi:10.3390/jcm9051256

Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N. P., Roos, C., et al. (2008). Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.* 358 (12), 1240–1249. doi:10.1056/NEJMoa0706728

Kawate, T., Iwaya, K., Kikuchi, R., Kaise, H., Oda, M., Sato, E., et al. (2013). DJ-1 protein expression as a predictor of pathological complete remission after neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Res. Treat.* 139 (1), 51–59. doi:10.1007/s10549-013-2523-0

Key, T. J., Appleby, P. N., Reeves, G. K., and Roddam, A. W. (2010). Insulin-like growth factor 1 (IGF1), IGF binding protein 3 (IGFBP3), and breast cancer risk: pooled individual data analysis of 17 prospective studies. *Lancet Oncol.* 11 (6), 530–542. doi:10.1016/s1470-2045(10)70095-4

Kim, R. H., Peters, M., Jang, Y., Shi, W., Pintilie, M., Fletcher, G. C., et al. (2005). DJ-1, a novel regulator of the tumor suppressor PTEN. *Cancer Cell* 7 (3), 263–273. doi:10.1016/j.ccr.2005.02.010

Kochmanski, J., Kuhn, N. C., and Bernstein, A. I. (2022). Parkinson's disease-associated, sex-specific changes in DNA methylation at PARK7 (DJ-1), SLC17A6 (VGLUT2), PTPRN2 (IA-2β), and NR4A2 (NURR1) in cortical neurons. *NPJ Park. Dis.* 8 (1), 120. doi:10.1038/s41531-022-00355-2

Li, H., Zhang, Z., Qiu, Y., Weng, H., Yuan, S., Zhang, Y., et al. (2023). Proteome-wide mendelian randomization identifies causal plasma proteins in venous thromboembolism development. *J. Hum. Genet.* 68, 805–812. doi:10.1038/s10038-023-01186-6

Li, Y. Y., Fang, J., and Ao, G. Z. (2017). Cathepsin B and L inhibitors: a patent review (2010 - present). *Expert Opin. Ther. Pat.* 27 (6), 643–656. doi:10.1080/13543776.2017.1272572

Mälarstig, A., Grassmann, F., Dahl, L., Dimitriou, M., McLeod, D., Gabrielson, M., et al. (2023). Evaluation of circulating plasma proteins in breast cancer using Mendelian randomisation. *Nat. Commun.* 14 (1), 7680. doi:10.1038/s41467-023-43485-8

Mata-Rocha, M., Rangel-López, A., Jiménez-Hernández, E., Morales-Castillo, B. A., González-Torres, C., Gaytan-Cervantes, J., et al. (2019). Identification and characterization of novel fusion genes with potential clinical applications in Mexican children with acute lymphoblastic leukemia. *Int. J. Mol. Sci.* 20 (10), 2394. doi:10.3390/ijms20102394

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551 (7678), 92–94. doi:10.1038/nature24284

Neufeld, C., Filipp, F. V., Simon, B., Neuhaus, A., Schüller, N., David, C., et al. (2009). Structural basis for competitive interactions of Pex14 with the import receptors Pex5 and Pex19. *Embo J.* 28 (6), 745–754. doi:10.1038/emboj.2009.7

Oshi, M., Gandhi, S., Yan, L., Tokumaru, Y., Wu, R., Yamada, A., et al. (2022). Abundance of reactive oxygen species (ROS) is associated with tumor aggressiveness, immune response, and worse survival in breast cancer. *Breast Cancer Res. Treat.* 194 (2), 231–241. doi:10.1007/s10549-022-06633-0

Prat, A., and Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* 5 (1), 5–23. doi:10.1016/j.molonc.2010.11.003

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Purrington, K. S., Slager, S., Eccles, D., Yannoukakos, D., Fasching, P. A., Miron, P., et al. (2014). Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* 35 (5), 1012–1019. doi:10.1093/carcin/bgt404

Rajkumar, T., Amritha, S., Sridevi, V., Gopal, G., Sabitha, K., Shirley, S., et al. (2022). Identification and validation of plasma biomarkers for diagnosis of breast cancer in South Asian women. *Sci. Rep.* 12 (1), 100. doi:10.1038/s41598-021-04176-w

Ren, F., Jin, Q., Liu, T., Ren, X., and Zhan, Y. (2023). Proteome-wide mendelian randomization study implicates therapeutic targets in common cancers. *J. Transl. Med.* 21 (1), 646. doi:10.1186/s12967-023-04525-5

Rohloff, J. C., Gelinas, A. D., Jarvis, T. C., Ochsner, U. A., Schneider, D. J., Gold, L., et al. (2014). Nucleic acid ligands with protein-like side chains: modified Aptamers and their use as diagnostic and therapeutic agents. *Mol. Ther. Nucleic Acids* 3 (10), e201. doi:10.1038/mtna.2014.49

Rosendahl, A. H., Björner, S., Ygland Rödström, M., Jirström, K., Borgquist, S., Ingvar, C., et al. (2021). Pre- and postoperative circulating igf-I, IGFBP-3, and IGFBP-7 levels in relation to endocrine treatment and breast cancer recurrence: a nested case-control study. *Front. Oncol.* 11, 626058. doi:10.3389/fonc.2021.626058

Sabe, H., Hata, A., Okada, M., Nakagawa, H., and Hanafusa, H. (1994). Analysis of the binding of the Src homology 2 domain of Csk to tyrosine-phosphorylated proteins in the suppression and mitotic activation of c-Src. *Proc. Natl. Acad. Sci. U. S. A.* 91 (9), 3984–3988. doi:10.1073/pnas.91.9.3984

Shu, X., Long, J., Cai, Q., Kweon, S. S., Choi, J. Y., Kubo, M., et al. (2020). Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat. Commun.* 11 (1), 1217. doi:10.1038/s41467-020-15046-w

Shuptrine, C. W., Ajina, R., Fertig, E. J., Jablonski, S. A., Kim Lyerly, H., Hartman, Z. C., et al. (2017). An unbiased *in vivo* functional genomics screening approach in mice identifies novel tumor cell-based regulators of immune rejection. *Cancer Immunol. Immunother.* 66 (12), 1529–1544. doi:10.1007/s00262-017-2047-2

Suhre, K., McCarthy, M. I., and Schwenk, J. M. (2021). Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* 22 (1), 19–37. doi:10.1038/s41576-020-0268-2

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., et al. (2018). Genomic atlas of the human plasma proteome. *Nature* 558 (7708), 73–79. doi:10.1038/s41586-018-0175-2

Sun, J., Zhao, J., Jiang, F., Wang, L., Xiao, Q., Han, F., et al. (2023). Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. *Genome Med.* 15 (1), 75. doi:10.1186/s13073-023-01229-9

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Swerdlow, D. I., Holmes, M. V., Kuchenbaecker, K. B., Engmann, J. E., Shah, T., Sofat, R., et al. (2012). The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* 379 (9822), 1214–1224. doi:10.1016/s0140-6736(12)60110-x

Takaya, K., Asou, T., and Kishi, K. (2023). Cathepsin F is a potential marker for senescent human skin fibroblasts and keratinocytes associated with skin aging. *Geroscience* 45 (1), 427–437. doi:10.1007/s11357-022-00648-7

Trayes, K. P., and Cokenakes, S. E. H. (2021). Breast cancer treatment. *Am. Fam. Physician* 104 (2), 171–178.

van der Zee, J., Mariën, P., Crols, R., Van Mossevelde, S., Dillen, L., Perrone, F., et al. (2016). Mutated CTSF in adult-onset neuronal ceroid lipofuscinosis and FTD. *Neurol. Genet.* 2 (5), e102. doi:10.1212/nxg.0000000000000102

Vazquez-Ortiz, G., Pina-Sanchez, P., Vazquez, K., Duenas, A., Taja, L., Mendoza, P., et al. (2005). Overexpression of cathepsin F, matrix metalloproteinases 11 and 12 in cervical cancer. *BMC Cancer* 5, 68. doi:10.1186/1471-2407-5-68

Veyssière, H., Bidet, Y., Penault-Llorca, F., Radosevic-Robin, N., and Durando, X. (2022). Circulating proteins as predictive and prognostic biomarkers in breast cancer. *Clin. Proteomics* 19 (1), 25. doi:10.1186/s12014-022-09362-0

Võsa, U., Claringbould, A., Westra, H. J., Bonder, M. J., Deelen, P., Zeng, B., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* 53 (9), 1300–1310. doi:10.1038/s41588-021-00913-z

Wei, S., Liu, W., Xu, M., Qin, H., Liu, C., Zhang, R., et al. (2022). Cathepsin F and Fibulin-1 as novel diagnostic biomarkers for brain metastasis of non-small cell lung cancer. *Br. J. Cancer* 126 (12), 1795–1805. doi:10.1038/s41416-022-01744-3

Wex, T., Levy, B., Wex, H., and Brömme, D. (1999). Human cathepsins F and W: a new subgroup of cathepsins. *Biochem. Biophys. Res. Commun.* 259 (2), 401–407. doi:10.1006/bbrc.1999.0700

Wingo, T. S., Liu, Y., Gerasimov, E. S., Gockley, J., Logsdon, B. A., Duong, D. M., et al. (2021). Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat. Neurosci.* 24 (6), 810–817. doi:10.1038/s41593-021-00832-6

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov. (Camb)* 2 (3), 100141. doi:10.1016/j.xinn.2021.100141

Wu, Y., Wang, Z., Yang, Y., Han, C., Wang, L., Kang, K., et al. (2023). Exploration of potential novel drug targets and biomarkers for small cell lung cancer by plasma proteome screening. *Front. Pharmacol.* 14, 1266782. doi:10.3389/fphar.2023.1266782

Xiao, T., Li, W., Wang, X., Xu, H., Yang, J., Wu, Q., et al. (2018). Estrogen-regulated feedback loop limits the efficacy of estrogen receptor-targeted breast cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.* 115 (31), 7869–7878. doi:10.1073/pnas.1722617115

Zhang, C., Qin, F., Li, X., Du, X., and Li, T. (2022a). Identification of novel proteins for lacunar stroke by integrating genome-wide association data and human brain proteomes. *BMC Med.* 20 (1), 211. doi:10.1186/s12916-022-02408-y

Zhang, G. Q., He, C., Tao, L., and Liu, F. (2015). Role of DJ-1 siRNA in reverse sensitivity of breast cancer cells to chemotherapy and its possible mechanism. *Int. J. Clin. Exp. Pathol.* 8 (6), 6944–6951.

Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., et al. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* 52 (6), 572–581. doi:10.1038/s41588-020-0609-2

Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M. E., et al. (2022b). Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* 54 (5), 593–602. doi:10.1038/s41588-022-01051-w

Zhang, W., Liu, Z., Xia, S., Yao, L., Li, L., Gan, Z., et al. (2021). GDI2 is a novel diagnostic and prognostic biomarker in hepatocellular carcinoma. *Aging (Albany NY)* 13 (23), 25304–25324. doi:10.18632/aging.203748

Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., et al. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res.* 50 (D1), D1398–d1407. doi:10.1093/nar/gkab953

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48 (5), 481–487. doi:10.1038/ng.3538