frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 14 February 2023
DOI 10.3389/fimmu.2023.1069379

OPEN ACCESS

EDITED BY
Pietro Ghezzi,
University of Urbino Carlo Bo, Italy

REVIEWED BY
Bor-Sen Chen,
National Tsing Hua University, Taiwan
Mario Marco Müller,
University Hospital Jena, Germany

*CORRESPONDENCE
Vinod Kumar
v.kumar@radboudumc.nl

SPECIALTY SECTION
This article was submitted to Inflammation, a section of the journal Frontiers in Immunology

RECEIVED 13 October 2022
ACCEPTED 23 January 2023
PUBLISHED 14 February 2023

CITATION
Boahen CK, Oelen R, Le K, Netea MG, Franke L, Wijst MGPvd and Kumar V (2023) Integration of *Candida albicans*-induced single-cell gene expression data and secretory protein concentrations reveal genetic regulators of inflammation. *Front. Immunol.* 14:1069379. doi: 10.3389/fimmu.2023.1069379

Integration of *Candida albicans*-induced single-cell gene expression data and secretory protein concentrations reveal genetic regulators of inflammation

Collins K. Boahen[1,2], Roy Oelen[3], Kieu Le[3], Mihai G. Netea[2,4], Lude Franke[3], Monique G.P. van der Wijst[3] and Vinod Kumar[1,2,3,5]*

[1]Department of Internal Medicine and Radboud Institute of Molecular Life Sciences (RIMLS), Radboud University Medical Center, Nijmegen, Netherlands, [2]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, Netherlands, [3]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, Netherlands, [4]Department for Immunology and Metabolism, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany, [5]Nitte University Centre for Science Education and Research (NUCSER), Nitte (Deemed to be University), Mangalore, India

Both gene expression and protein concentrations are regulated by genetic variants. Exploring the regulation of both eQTLs and pQTLs simultaneously in a context- and cell-type dependent manner may help to unravel mechanistic basis for genetic regulation of pQTLs. Here, we performed meta-analysis of *Candida albicans*-induced pQTLs from two population-based cohorts and intersected the results with *Candida*-induced cell-type specific expression association data (eQTL). This revealed systematic differences between the pQTLs and eQTL, where only 35% of the pQTLs significantly correlated with mRNA expressions at single cell level, indicating the limitation of eQTLs use as a proxy for pQTLs. By taking advantage of the tightly co-regulated pattern of the proteins, we also identified SNPs affecting protein network upon *Candida* stimulations. Colocalization of pQTLs and eQTLs signals implicated several genomic loci including MMP-1 and AMZ1. Analysis of *Candida*-induced single cell gene expression data implicated specific cell types that exhibit significant expression QTLs upon stimulation. By highlighting the role of trans-regulatory networks in determining the abundance of secretory proteins, our study serve as a framework to gain insights into the mechanisms of genetic regulation of protein levels in a context-dependent manner.

KEYWORDS

***Candida albicans*, proteomics, pQTL, single-cell eQTL, colocalization**

Introduction

Genome-wide association studies (GWAS) have successfully identified tens of thousands of associations between single nucleotide polymorphisms (SNPs) and human diseases. Correlating these GWAS SNPs with molecular traits (QTLs) such as gene expression (eQTLs) is a commonly used strategy to prioritize causal genes. This is partly due to the robustness of RNA-sequencing technologies and the feasibility of eQTLs to provide insights into the molecular mechanisms of genetic variants associated with complex diseases (1). The majority of the eQTL studies have made use of RNA extracted from whole blood and analyzed using bulk RNA-sequencing approaches to unravel disease biology. However, this approach is limited in identifying a genetic variant's cell-type-specific and context-dependent impact on gene expression levels (2, 3) or causal cell types of a particular disease.

In addition to eQTLs, protein quantitative trait loci (pQTLs) are also important molecular traits to understand GWAS findings. Circulating plasma proteins play essential roles in various biological process such as signaling and defense against infections but also, the dysregulation of proteins themselves lead to various diseases and are mostly the targets for therapeutic interventions (4). Proteins provide the closest link to phenotypic traits as being the ultimate product of transcripts. Given that not all RNA alterations lead to functional changes, studies of genetics regulation at protein level are warranted. By making use of secretory protein and genotype data from 30,931 samples, a recent pQTL study (5) showed that approximately 29% of the pQTLs are also GWAS SNPs. Also, pQTLs can be tissue- and context-specific. For example, SNPs affecting cytokine production upon *ex-vivo* blood stimulation were shown to overlap with SNPs associated with infectious and inflammatory diseases (6). However, which specific cell type is contributing to the production of these secretory proteins is unclear. In addition, by examining the relationship between pQTLs and eQTLs, previous studies have demonstrated the disparity of genetic variants associated with mRNA expressions and protein abundances as summarized by a minireview (7), and a recently published study affirms this observation with substantial difference between pQTLs and eQTLs where only 32% of the index eQTLs variants were replicated in pQTLs (8). While previous studies examining the extent of overlap between pQTLs and eQTLs extensively explored steady-state conditions, the degree to which these findings are replicated in stimulated conditions remain elusive.

Proteins constitute the largest class of drug targets and thus, the identification of disease-mediating candidate proteins is crucial to bridging the gap between human diseases and the genome (9). However, the proportion of pQTLs mostly overlapping with known disease-associated loci is very limited with percentages as low as 29% (5) and 12% (10) observed in previous studies. Studying pQTLs in the right context could provide more explanation of the link between genetics and diseases as well potential targets for treatment. To tackle these challenges, in this study, we first aimed to identify pQTL variants upon *ex-vivo Candida albicans* stimulation in two independent European cohorts. Secondly, we combined the pQTLs identified with cell-type-specific eQTLs upon *Candida albicans* stimulation to prioritize genes at the genomic regions regulating the protein abundances. This study provides deeper insights into the genetic basis underlying variations in *Candida*-stimulated protein abundance as we identified pQTLs through meta-analysis which colocalized with cell-type-specific *cis*-eQTLs.

Materials and methods

Study populations

The 500FG cohort of healthy individuals of Western European ancestry comprises of 237 males and 296 females with age range of 18 to 75 years, being part of the Human Functional Genomics Project, HFGP (www.humanfunctionalgenomics.org).

The 1M-scBloodNL cohort comprises of 120 individuals, 53 males and 67 females between 27 to 78 years. This cohort is part of the Lifelines DEEP cohort, a prospective population cohort of participants from the northern Netherlands (11).

PBMC isolation and Candida albicans stimulation experiments

500FG cohort

Peripheral blood mononuclear cells (PBMCs) collection has been previously described (6). With informed consent, venous blood was drawn from the cubital vein of study participants into 10mL EDTA Monoject tubes (Medtronic, Dublin). The fraction of PBMC was obtained by density centrifugation of EDTA blood diluted 1:1 in pyrogen-free saline over Ficoll-Paque (Pharmacia Biotech, Uppsala). Cells were washed twice in saline and suspended in medium (RPMI 1640) supplemented with gentamicin (10 mg/mL), L-glutamine (10 nM) and pyruvate (10mM). The cells were counted in a Coulter counter (Beckman Coulter, Pasadena) and the number of was adjusted to 5 x 10^6 cells/mL. A total of 5 x 10^5 PBMCs were added in 100 ul to round-bottom 96-well plated (Greiner) and incubated with 100 uL of stimulus (heat-killed *Candida albicans* yeast of strain ATCC MYA-3573,UC 820, 1 X10^6/mL or RPMI 1640 as previously described (12).

1M-scBloodNL cohort

PBMCs from 120 volunteers were isolated and stimulated as previously described (13). Briefly, we used Cell Preparation Tubes with sodium heparin (BD) to isolate PBMCs, which were cryopreserved until use in RPMI1640 containing 40% FCS and 10% DMSO. After thawing and a 1h resting period, unstimulated cells were washed twice in medium supplemented with 0.04% BSA and directly processed for scRNA-seq. On the other hand, for stimulation experiments, 5x10^5 cells were seeded in a nucleon sphere 96-well round bottom plate in 200 ul RPMI1640 (supplemented with 50 ug/mL gentamicin, 2 mM L-glutamine and 1 mM L-glutamine and 1mM pyruvate). The cells were stimulated with 1x10^6 CFU/ml heat-killed *Candida albicans* blastoconidia (strain ATCC MYA-3573, UC 820), 50 ug/ml heat-killed *M. tuberculosis* (strain H37Ra, *In vivo*gen) or 1x10^7 heat-killed *P. aeruginosa* (*In vivo*gen) for 3h or 24h, at 37˚C in a 5% CO2 incubator. After stimulations, cells were washed twice in medium supplemented with 0.04% BSA. Cells were then counted using a haemocytometer, and cell viability was assessed by Trypan

Blue. While scRNA-seq data was generated for all the above-mentioned stimulations, Olink data was generated for only *Candida albicans stimulation*.

Measurement of inflammatory proteins

Inflammatory protein concentrations were measured using Olink® proteomics platform. In fact, supernatants were collected after 24h of *Candida albicans* stimulation and submitted to Olink Proteomics for analysis using the inflammatory panel assay of 92 analytes. Olink data are presented as Normalized Protein eXpression values (NPX, based on log2 scale). Immunoassays utilized by Olink are based on the Proximity Extension Assay (PEA) technology (14), which makes use of oligonucleotide-labeled antibodies binding to their respective protein. When the two antibodies are brought in proximity, a DNA polymerase target sequence is formed, which is subsequently quantified by quantitative real-time polymerase chain reaction (qPCR).

Preprocessing/filtering of protein data and normalization

Filtering of Olink generated data was restricted to only Proteins as the samples passed Olink internal quality control across all proteins. We excluded all proteins which failed to be quantified in at least 85% of the samples, meaning all proteins with more than 15% samples missingness (NPX value below the protein-specific limit of detection (LOD) value) were excluded from downstream analysis. The remaining Proteins with NPX values below the LOD were replaced with protein-specific LOD values.

Following per-protein cleaning, 42 and 35 proteins were available for the 1M-scBlodNL and 500FG cohorts respectively (Figure S1A). Out of these proteins, 26 were common between both cohorts.

The protein distributions on log2 scale were not normally distributed (Figure S1B). We applied rank-based inverse normal transformation as implemented in the GenABEL R package (15), to transform the data to mimic Gaussian distribution (Figure S1C).

SNP genotyping, quality control and imputation

The procedures for genotyping, genetic data filtering and genotype imputation of the 500FG cohort had been previously described (6). Extracted DNA was genotyped using the commercially available SNP chip, Illumina HumanOmniExpressExome-8 v1.0. Following pre-imputation filtering steps for both markers and individuals, the remaining dataset SNP genotypes were imputed with GoNL as reference panel (16).

For Lifelines Deep cohort, genotyping and imputation was performed as previously described (17). Both the HumanCytoSNP-12 BeadChip and the ImmunoChip platforms (Illumina, San Diego, CA, USA) were used to genotype the isolated DNA. Independent markers quality control was performed for both platforms and subsequently merged into one dataset. After merging, genotype SNPs were imputed using IMPUTE2 (18) against the GoNL reference panel.

Correlation analysis

Pairwise Pearson correlation analysis using the 'corrr" R package was performed on the normalized protein abundances after adjusting for age and sex. Based on Pearson's correlations for each pair of proteins, co-expression protein networks were reconstructed (using significant correlation coefficient threshold of 0.7, (absolute(r) > 0.70), a cut-off denoting a very strong strength of association.

Identification of pQTLs after Candida albicans stimulation

The association analysis of genotype-phenotype correlation was carried out using two main approaches. In the first method, the univariate approach was performed using the linear regression in PLINK (19). The pQTL analysis was conducted independently for both cohorts, that is one analysis for the 1M-scBloodNL and another for 500FG cohorts. In the second method, multivariate test of association based on canonical correlation analysis (CCA) (20), was conducted. CCA extracts the linear combination of highly corrected traits that explain the largest possible amount of the covariation between genetic variants and all traits (Proteins in this case). To control for potential confounding factors, we adjusted for covariates such as age and sex on normalized protein abundances and, regressed the residuals of each protein and protein network on SNP genotypes.

Meta-analyses

Summary statistics from both primary studies were utilized to perform meta-analyses.

Association results for both the 500FG and 1M-scBloodNL cohorts obtained from the univariate approach were combined using the weighted sum fixed-effect model as implemented in the METAL software program (21). The multivariate approach implemented in PLINK does not compute beta estimates and standard errors. Therefore, the meta-analysis of the multivariate P-values was carried out using sum of z (Stouffer's) method as implemented in the metap R package (22).

Genetic colocalization analyses

Colocalization analysis of cell-type-specific *cis*-eQTL and pQTL signals was conducted using Bayesian colocalization method which is implemented in the coloc package in R (23). We retrieved the genome-wide cell-type-specific *cis*-eQTL summary statistics from a previously published study using the 1M-scBloodNL cohort (13). Additionally, we performed trans-eQTL mapping only for the top pQTLs to ascertain their trans-eQTL effect. The eQTLs from this study were identified using PBMCs in an unstimulated condition as well as after 3h and 24h *in vitro*-stimulations with three different

pathogen stimulations, namely *Candida albicans*, *M. tuberculosis* and *P. aeruginosa*. *Cis*-eQTL was defined as SNP-gene distance of 100kb window and FDR < 0.05. However, for overlap comparison with pQTLs, we re-tested the relevant SNPs to ensure that SNPs outside the defined *cis*-region are included.

From our pQTL mapping results, we selected the index variants (SNPs with the smallest P value) for each protein and extracted all variants within a window size of 1Mb around the index pQTL variant for further analysis. The default prior probability ($1x10^{-6}$) that a random variant in the region is causal to both traits was applied. A posterior probability (PP4 >= 0.75) is considered as strong evidence of colocalization. LocusCompareR, being an R package was used for the visualization of results (24).

Results

Identification of protein quantitative trait loci (pQTL) in univariate manner

To identify SNPs associated with the concentration of specific proteins induced upon 24h *Candida albicans* stimulations, pQTL analysis was performed in two independent population-based cohorts. A total of 445 participants (500FG (N= 325) and 1M-scBloodNL (N=120) cohort) were studied for whom both genotype and Olink protein abundances were measured upon *ex vivo* 24h *Candida albicans* stimulation of their PBMCs (Figure 1). After genotype imputation and quality control, 4,095,761 SNPs and 26 inflammatory proteins remained that were common between both cohorts and that were used as input for the pQTL analysis. The protein data for both cohorts were generated in different batches. Therefore, pQTL analysis was run in each cohort separately, after which results (Figure S2) were integrated using meta-analysis to increase statistical power. We identified a genome-wide significant *cis*-acting pQTL variant rs484915 (P value = $1.81x10^{-8}$) on chromosome 11 correlating with MMP-1 production (Figure 2A; Table 1). In addition, the other top SNPs correlating with the remaining 25 proteins were *trans*-acting pQTLs with strong suggestive associations (P value > $5.0\ x10^{-8}$ to $5\ x10^{-6}$). For example, the second most significant hit aside the genome-wide significant cQTL, is an intronic SNP rs62205465 residing in the *ZNF133* locus and exhibited an association strength closed to the genome-wide significant threshold (P value = $5.80\ x10^{-8}$) with MCP-3 concentrations. Figure 2A illustrates the association results of all the proteins analyzed and their corresponding top SNPs (lowest P-value).

Comparison between pQTL and cis-eQTL upon Candida albicans stimulation

Previously, we had conducted a genome-wide eQTL analysis per major cell type using scRNA-seq data from unstimulated and 3h and 24h pathogen-stimulated (*Candida albicans, M. tuberculosis, P. aeruginosa*) PBMCs in the 120 individuals from the 1M-scBloodNL cohort (13). This data was used to interrogate whether the same pQTL SNP could also affect gene expression levels to explore the degree of association between pQTLs and eQTLs. Note that the pQTL data captures the bulk secretion of proteins coming from PBMCs that were stimulated for 24h with *Candida albicans*, whereas the eQTL data was collected for each cell type separately. Out of the 24 unique top pQTL variants with a suggestive or genome-wide significant association (P value from $1.81\ x\ 10^{-8}$ to $5.89\ x\ 10^{-6}$) identified from meta-analyzed results of the univariate approach (Table 1), we could only overlay 20 with the scRNA-seq derived eQTL data of the 1M-scBloodNL cohort, as eQTL data for the remaining 4 SNPs (rs36067904, rs13033376, rs62129298, and rs7018706) were not available. Seven out of the 20 tested SNPs showed an eQTL effect upon 24h *Candida albicans* stimulation in at least one cell type (FDR < 0.05) (Figure S3). Among these seven, two pQTLs showed genome-wide significant association with gene expression levels specifically in monocytes. The first *cis*-acting pQTL, rs484915 showed association with both MMP-1 protein concentrations and gene expression (Figure 2B). Of note, only

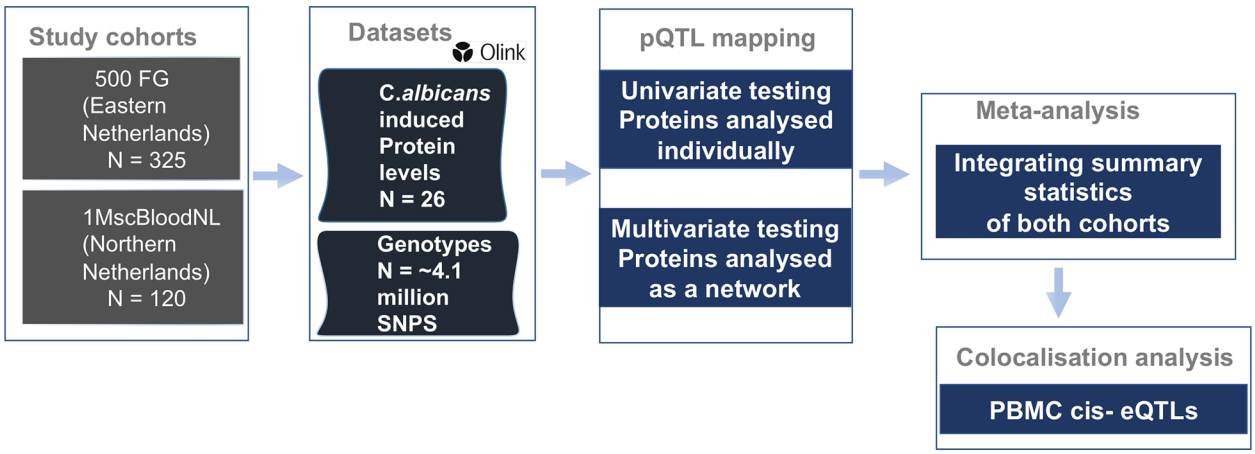

FIGURE 1
Overview of study cohorts, study design and analyses conducted. We performed protein quantitative trait loci (pQTL) mapping utilizing protein abundances and imputed genotypes in two population-based cohorts of individuals of European ancestry. We used two approaches to identify pQTL signals which were meta-analyzed and subsequently colocalized with cell-type specific eQTLs and candidemia GWAS signals to uncover genomic regions of shared association.

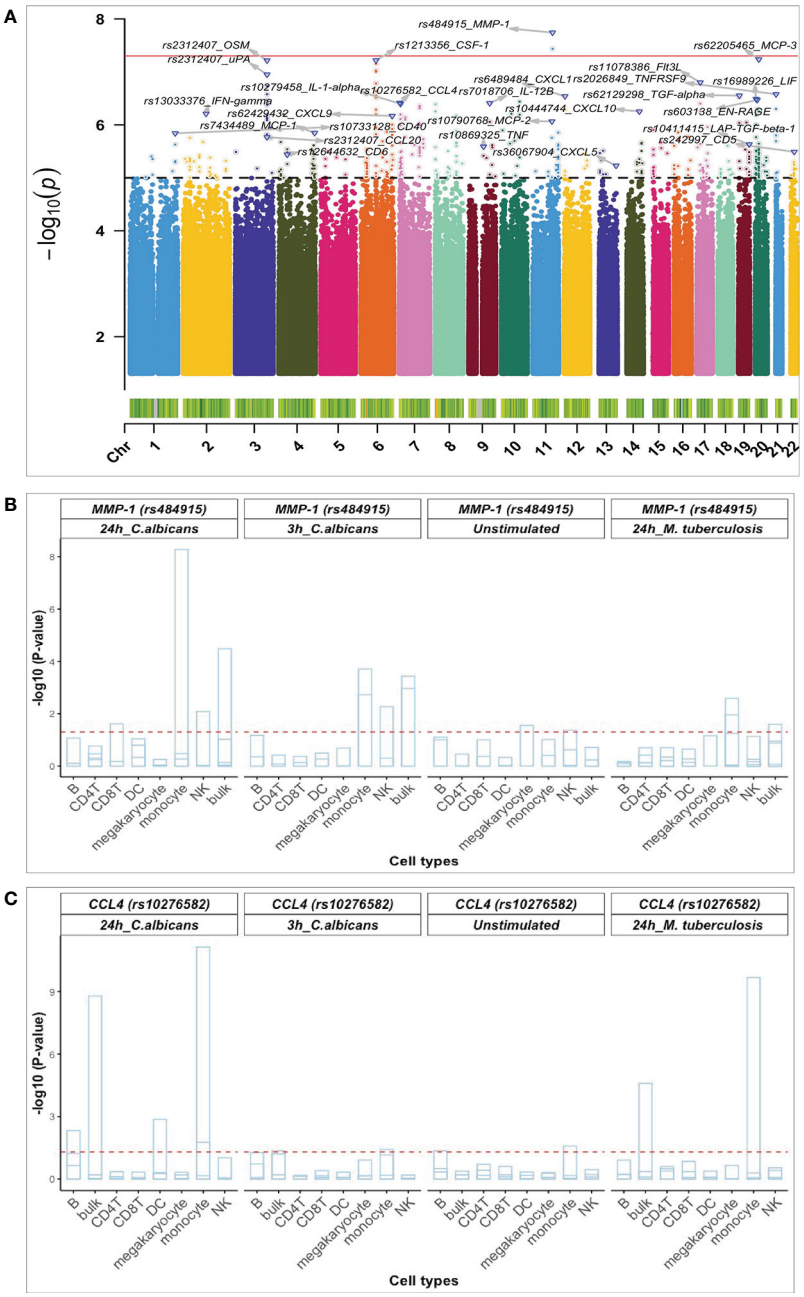

FIGURE 2
Identified pQTL SNPs and their association with cell-type specific cis-eQTLs. **(A)** Manhattan plot of pQTL genetic variants associated with 26 proteins identified using a univariate approach and after meta-analysis. The red bold horizontal line depicts the genome-wide significant threshold (p-value < 1 x 10^{-8}) and the black dashed denotes the suggestive association threshold (p-value = 1 x10^{-5}) Top pQTL variants and their correlated proteins are displayed on the plot. Barplots of pQTL variants correlating with - MMP-1 **(B)** and CCL4 **(C)** with genome-wide significant eQTLs in monocytes. The horizontal axis shows all the cell types considered and the vertical axis represents the negative log10 p-values for eQTLs. Each strip in the barplot corresponds to gene correlating with the pQTL variants (rs484915 and rs10276582). The barplot is also grouped per stimulation and timepoints together with pQTL variants. The horizontal red dashed line corresponds to 0.05 P-value, FDR corrected.

this pQTL among the seven was associated with the corresponding gene of the protein. The second is trans-acting SNP rs10276582 associated with CCL4 protein production affected *AMZ1* gene expression levels (Figure 2C). This observation suggests *AMZ1* may regulate CCL4 protein levels depending on an individual's genotype at SNP rs10276582 (or any other SNP in high LD).

To elucidate whether the effects of the significant eQTLs were detectable only in *Candida* stimulations after 24h, we further assessed their impact in unstimulated condition, 3h *Candida* stimulations, as well as in *Mycobacterium tuberculosis* stimulations after 24h. We observed that the effects of the *cis*-eQTLs were less pronounced before stimulation and after 3h *Candida* stimulation, which suggests temporal regulation of gene expression following *Candida* stimulation. Nearly equal strength of association was identified in the case of *M. tuberculosis* stimulations after 24h for only the *trans*-acting variant SNP rs10276582 (Figure 2C).

10.3389/fimmu.2023.1069379

TABLE 1 Summary of the top pQTLs from meta-analysis of 26 proteins common in 500FG and 1M-scBloodNL.

SNP	CHR	A1	A2	P-value	Zscore	Weight	D	Protein	Cell type	Distance	Closet gene
rs484915	11	A	T	$1.81x10^{-8}$	-5.629	445	–	MMP-1	Monocyte	*cis*	*MMP1*
rs62205465	20	T	C	$5.80x10^{-8}$	-5.425	445	–	MCP-3	Monocyte	*trans*	*ZNF133*
rs1213356	6	T	C	$6.04x10^{-8}$	-5.417	445	–	CSF-1	Monocyte	*trans*	*HTR1B*
rs2312407	3	A	G	$6.07x10^{-8}$	-5.417	445	–	OSM	Monocyte	*trans*	*BCHE*
rs2312407	3	A	G	$1.12x10^{-7}$	-5.306	445	–	uPA	Monocyte	*trans*	*BCHE*
rs11078386	17	A	C	$1.58x10^{-7}$	5.243	445	++	Flt3L	CD4T	*trans*	*NT5M*
rs2026849	21	A	G	$2.64x10^{-7}$	-5.148	445	–	TNFRSF9	Dendritic	*trans*	*NRIP1*
rs62129298	19	C	G	$2.80x10^{-7}$	5.136	445	++	TGF-alpha	Monocyte	*trans*	*TMIGD2*
rs6489484	12	T	C	$2.88x10^{-7}$	5.131	445	++	CXCL1	Monocyte	*trans*	*EFCAB4B*
rs16989226	20	A	G	$3.25x10^{-7}$	-5.108	445	–	LIF	Monocyte	*trans*	*SMOX*
rs603138	20	C	G	$3.39x10^{-7}$	-5.1	445	–	EN-RAGE	Monocyte	*trans*	*ANKRD5*
rs10276582	7	C	G	$3.83x10^{-7}$	-5.077	445	–	CCL4	Monocyte	*trans*	*AMZ1*
rs7018706	9	A	T	$3.89x10^{-7}$	-5.074	445	–	IL-12B	Dendritic	*trans*	*TMEM38B*
rs10279458	7	T	C	$4.02x10^{-7}$	5.068	445	++	IL-1-alpha	Monocyte	*trans*	*GRID2IP*
rs10444744	14	A	G	$5.59x10^{-7}$	5.005	445	++	CXCL10	Monocyte	*trans*	*FLRT2*
rs13033376	2	A	T	$6.18x10^{-7}$	-4.986	445	–	IFN-gamma	CD4T	*trans*	*INSIG2*
rs62429432	6	A	C	$6.77x10^{-7}$	4.968	445	++	CXCL9	Dendritic	*trans*	*PARK2*
rs10790768	11	A	T	$8.61x10^{-7}$	4.921	445	++	MCP-2	Monocyte	*trans*	*CNTN5*
rs7434489	4	A	G	$1.42x10^{-6}$	4.822	445	++	MCP-1	Monocyte	*trans*	*FAM149A*
rs10733128	1	T	C	$1.44x10^{-6}$	4.82	445	++	CD40	Dendritic	*trans*	*U6*
rs2312407	3	A	G	$1.71x10^{-6}$	-4.786	445	–	CCL20	Monocyte	*trans*	*BCHE*
rs10411415	19	C	G	$2.33x10^{-6}$	4.722	445	++	LAP-TGF-beta-1	CD4T	*trans*	*SGK110*
rs10869325	9	A	G	$2.57x10^{-6}$	4.702	445	++	TNF	Monocyte	*trans*	*RORB*
rs242997	22	A	G	$3.23x10^{-6}$	4.656	445	++	CD5	CD4T	*trans*	*LARGE*
rs12644632	4	A	G	$3.64x10^{-6}$	4.631	445	++	CD6	CD4T	*trans*	*KCTD8*
rs36067904	13	A	T	$5.89x10^{-6}$	-4.53	445	–	CXCL5	Monocyte	*trans*	*FAM155A*

D, Direction of effect size; Bold, SNP affecting 3 proteins; A1, Effect allele; CHR, chromosome; Cell type represents the cells in which the pQTL proteins are mostly expressed in the 1M-scBloodNL cohort. ++ means positive effect size and – means negative effect size.

Colocalization analysis identifies causal genes at Candida albicans-induced pQTLs

Next, to uncover the potential mechanisms underlying the observed pQTLs, we tested whether SNPs impacting protein concentrations at specific loci are also the same causal regulatory eQTLs through colocalization analysis. We identified strong evidence of colocalization for both pQTLs and eQTLs. For instance, in the MMP-1 locus (Figure 3A), the posterior probability (PP.H4) was 0.998 (Figure 3B). Also, the *trans*-acting pQTL SNP rs10276582 located in the *AMZ1* locus on chromosome 7 (Figure 3C), showed significant colocalization (PP.H4 = 0.996) with the eQTL effect. eQTL analysis revealed association of this variant with three different *cis* genes, *GNA12* ($P = 7.1 \times 10^{-1}$), *TTYH3* ($P = 1.7 \times 10^{-2}$) and *AMZ1* ($P = 7.31 \times 10^{-12}$) (Figure 3D), indicating the presence of multiple causal genes at this locus. However, both *GNA12* and *TTYH3* exhibited weaker strength of association than AMZ1, which showed statistically significant association. Thus, pinpointing AMZ1- which is predicted to belong to a large metallopeptidase protein family and interacts with cell receptors and growth factors, as the potential causal gene in this genomic region. This observation demonstrates that *cis*-regulation of gene expression levels maybe involved in the mechanisms by which distal variants impact protein expression.

Exploring other mechanisms of pQTLs function on protein levels

To further understand the mechanisms or regulation of the trans-pQTL, we first looked for evidence of the genes near or at the trans loci encoding for any of the proteins interacting with our tested protein. To do this, we used an annotation database – Human

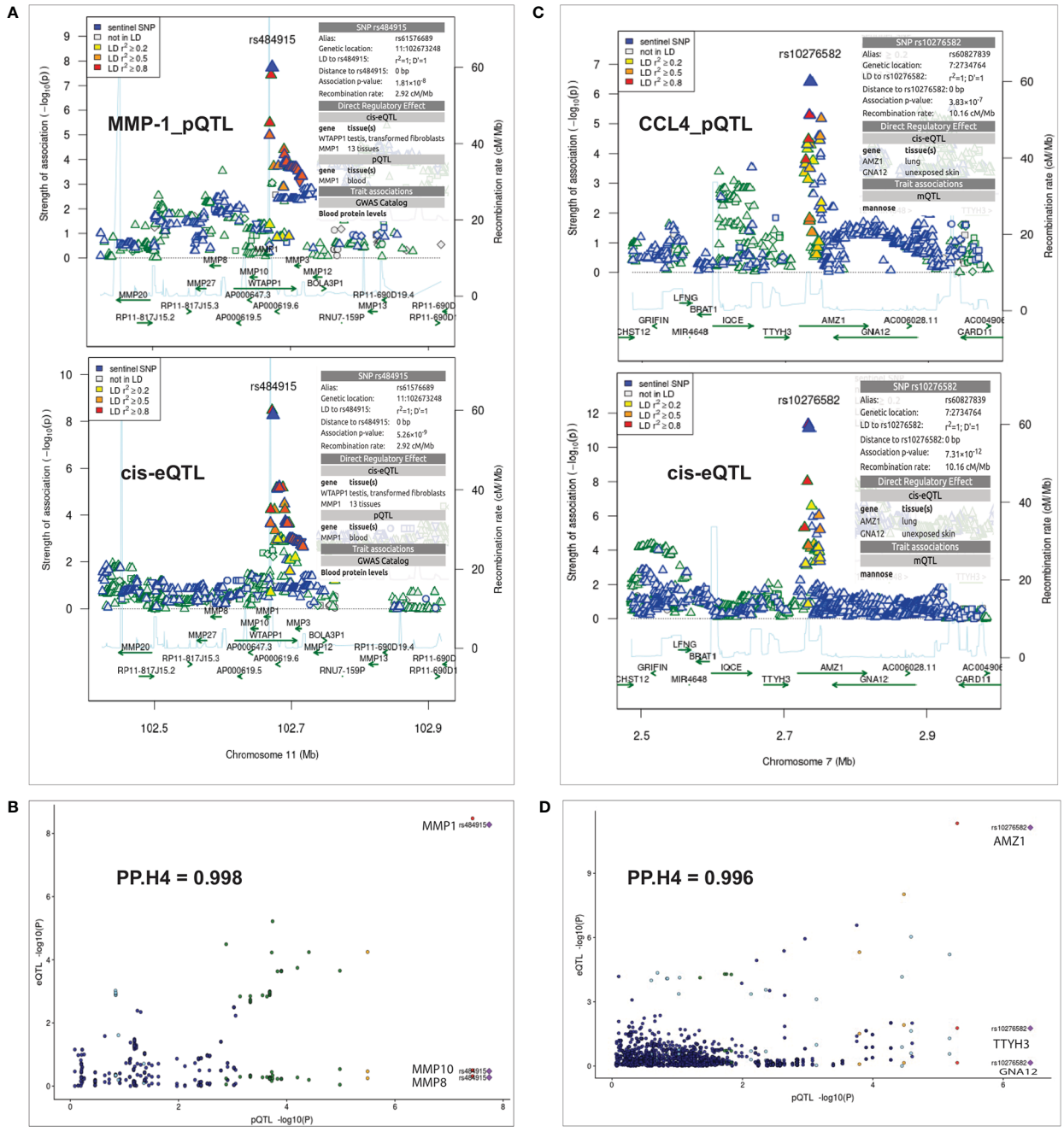

FIGURE 3
Summary of colocalization analyses between pQTL and eQTL. **(A)** Regional association plots at the MMP-1 locus for MMP-1 protein QTL result (top panel) and cis-QTL expressions result (bottom panel) on chromosome 11. **(C)** Regional association plots at the AMZ1 locus with CCL4 protein QTL result (top panel) and cis-QTL expressions result (bottom panel) on chromosome 7. The sentinel SNPs (rs484915 and rs10276582) are indicated with blue diamond shape and other surrounding SNPs are colored with different levels of linkage disequilibrium with the sentinel SNPs. The horizontal axis indicates chromosomal positions (NCBI human genome build 37) and the vertical axes represent negative $\log_{10}$ p-values and recombination rates (cM/Mb) estimated from 1000 Genome Project (European population) version 3.3. **(B, D)** Correlation plots with strong evidence of colocalization between pQTLs and eQTLs as indicated by PP.H4 (posterior probability of shared causal variants) values.

Integrated Protein-Protein Interaction Reference (HIPPIE) (25) and defined trans genes as all genes falling within 1Mb window centered on the top 25 identified trans-pQTL loci. Generally, we did not observe any interacting partners between the trans genes and the trans affected proteins. We have presented some examples to demonstrate the findings. (Figures S4-5).

Co-regulation of Candida albicans-induced protein levels

It is possible that some of the genetic variants may affect multiple proteins, so we wanted to explore whether there is strong correlation between protein concentrations upon stimulation. To explore the

patterns underlying *Candida*-induced protein production, we performed correlation analyses. We observed mostly significant positive pairwise correlation (with the exception of MCP-1 and MCP-2) between the 26 proteins in the 500FG cohort, which contains the largest number of individuals from which samples were collected (Figure 4A). However, in the 1M-scBloodNL cohort, divergent patterns of correlation strengths were observed, including weaker and negative correlations between CSF-1, IL-12B and IFN-gamma, contrasting the positive correlation observed in the 500FG cohort (Figure 4B). To understand what might underlie this observation, we performed the Fligner-Killen test to evaluate whether there is significant donor variation of these proteins between the two cohorts. While we observe significant difference for CSF-1 (Fligner-Killeen:med chi-squared = 19.893, p-value = 8.19 x 10^{-6}) and IL-12B (Fligner-Killeen:med chi-squared = 5.0642, p-value = 0.02442), there was no evidence to suggest that the variance in IFN-gamma concentrations significantly differ between both cohorts (Fligner-Killeen:med chi-squared = 0.3097, p-value = 0.5779).

In the 500FG cohort, the strongly pairwise correlated proteins acting as protein network consisted of 24 out of the 26 proteins common between both cohorts (Figure 4C). For example, the correlation coefficient between uPA and OSM was as high as 0.91, while CD40 exhibited an approximately 0.9 between CD5 and CD6. Next, we sought to more accurately infer the protein network by replicating the analysis in the 1M-scBloodNL cohort. While MCP-1 exhibited strong negative correlation with some proteins (such as OSM and uPA) in the 500FG cohort, this pattern of correlation was hidden in the network obtained from the 1M-scBloodNL cohort (Figure 4D). A rather weaker and positive correlation was manifested between MCP-1 and OSM (0.26) and uPA (0.24) and thus, MCP-1

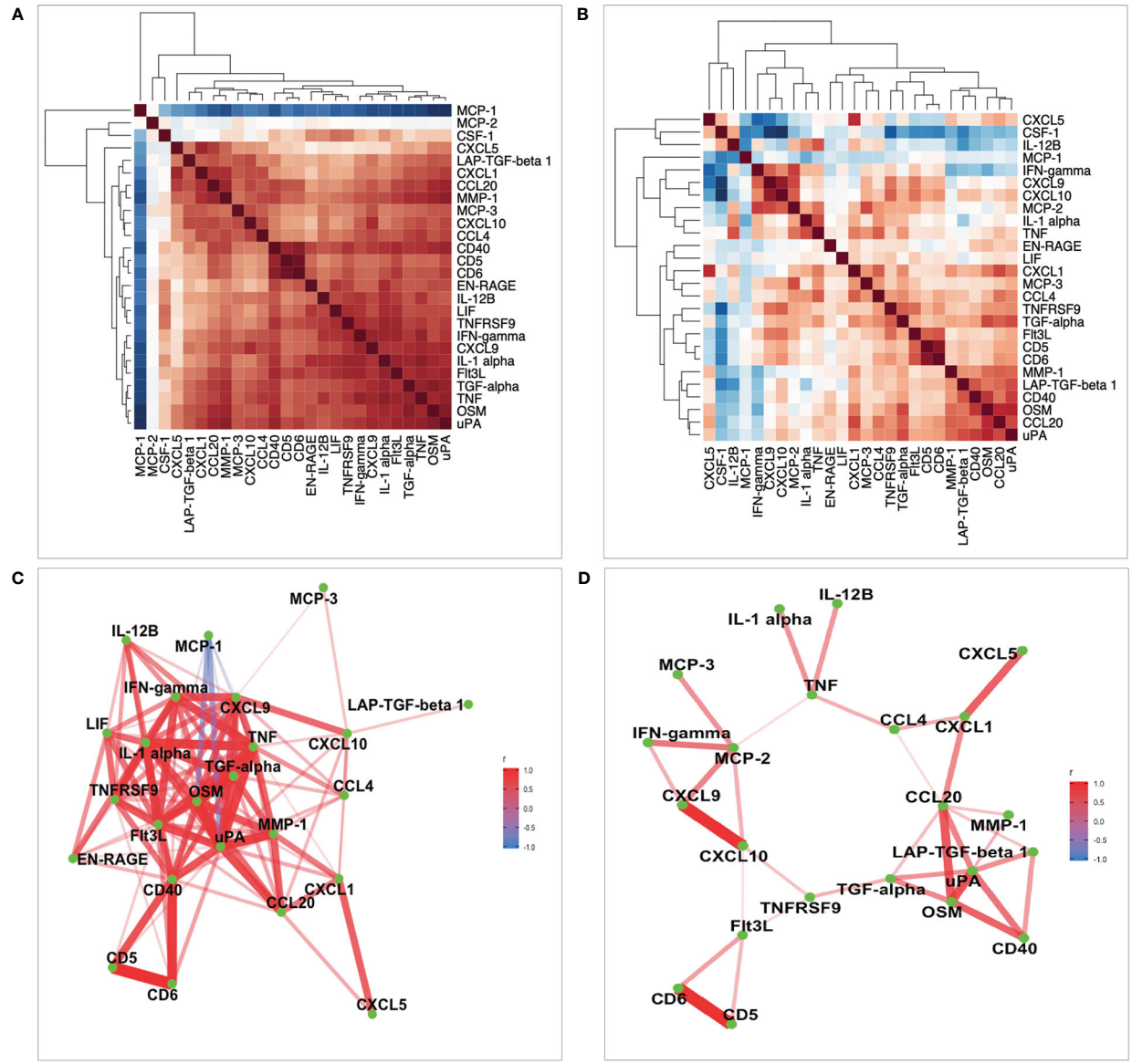

FIGURE 4
Heatmaps of correlation analysis. **(A, B)** Correlation of 26 common proteins for the 500FG and 1M-scBloodNL cohorts respectively. Pairwise correlation between proteins were computed using residual values after adjusting age and sex on protein levels. **(C, D)** Represents network of proteins with strong pairwise correlation coefficients for the 500FG and 1M-scBloodNL cohorts respectively.

was excluded from multivariate genetic association analysis as shared genetic architecture is likely to underlie or trigger the observed strong positive correlations.

Genome-wide identification of pQTLs impacting protein network

Taking advantage of the tightly co-regulated pattern of the proteins, we sought to identify SNPs impacting protein network upon *Candida* stimulations aside the univariate association analysis. After independent analyses in both cohorts, we identified a genome-wide significant hit in the 500FG cohort as well as several suggestive associations. The genome-wide significant SNP was rs938662 (P value = 4.37 x 10^{-8}), residing in the *PRKCE* locus (Figure S6A). This suggest that PRKCE locus exhibit a pleotropic role as the top SNP in this genomic region is associated with multiple highly correlated proteins. There was no evidence of inflation of the test statistics as genomic inflation factor (Λ) was computed to be 1.01 (Figure S6B). In the case of the 1M-scBloodNL dataset, no statistically significant SNP was found to be associated with proteins (Figure S6C), and corresponding genomic inflation factor (Λ) was 0.99, indicating lack of inflation of test statistics (Figure S6D).

To identify true or robust association signals, we synthesized summary statistics from both cohorts through meta-analysis. Even though no loci reached genome-wide significance, we identified strong suggestive associations (Figure 5A). The top signal identified was rs484915 (P value = 1.05x10^{-7}), located at the *MMP1* locus and has previously been shown to alter expression levels of *cis*-genes in multiple tissues and also blood protein concentrations (26). In the univariate analysis, we found the same SNP rs484915 to be associated with MMP-1 with slightly much stronger association based on P-value (1.81x10^{-8}). Interestingly, five proteins, namely MMP-1, CXCL5, CCL20, CXCL1 and CXCL10 among the proteins forming the network (Figure 5B), contributed with relatively stronger weights or correlation coefficients to the observed association result of SNP rs484915 to the protein network. This observation from the multivariate approach demonstrates the pleiotropic effect of SNP rs484915 which cannot be captured directly *via* univariate analysis and also tease apart the main proteins whose expression levels are being regulated.

Multivariate approach improves statistical power

Next, we sought to investigate whether the joint analysis of multiple correlated proteins with genetic variants offers some advantage over univariate analysis. To achieve this aim, we directly compared the P-values or distribution of P-values of genetic variations identified using both approaches and restricted the analysis to only the largest cohort, 500FG. We reasoned that such analysis will be more credible to be conducted in a specific cohort due to the variation seen in proteins' correlation structure between both cohorts. As expected, we observed stronger associations emanating from the multivariate analysis compared to the univariate manner (Figure 5C). The top 6 independent strong suggestive loci (P < 9 x10^{-7}) identified *via* the multivariate approach were in all cases showing a much stronger association when compared to the strength of association (P values) of each protein analyzed separately (Figure 5D). Furthermore, to characterize the identified pQTLs, we targeted the top 6 strongest pQTL variants (rs938662, rs1501565, rs188465730, rs1020993, and rs11902595), given the lack statistically significant SNPs to elucidate their effect on cell-type specific eQTLs. We found 3 out of 6 to be an eQTL as well in at least one cell type albeit one with nominally significant effect. The overlapped SNPs showed weak association with various *cis*-genes. For instance, the strongest effect was observed for intronic SNP rs11902595 on chromosome 2 which was nominally (P= 0.013) correlated with a lincRNA (RP11-191L7.1) in CD8T cells.

Overlap of pQTLs with SCALLOP consortium data

We further evaluated the overlap and strength of association between our *Candida* 24h stimulated pQTL associations and previously reported highly powered (21,758 participants) unstimulated pQTL study published under the SCALLOP consortium (5). We found only 8 proteins (MMP-1, CSF-1, CXCL1, EN-RAGE, CCL4, MCP-1, CD40, CCL20) common between the 90 cardiovascular proteins measured in the previous study and the 26 proteins measured in the inflammatory panel used in our study. Using nominal significance P-value < 0.05, the percentage of shared *Candida* 24h stimulated pQTLs vs pQTLs identified using the SCALLOP consortium data, ranged from 4.5% to 5.6%. Of these, the top SNPs among the shared variants show nominal association with 6 proteins after stimulation. However, *cis*-acting genome-wide significant pQTL variant rs484915 (P = 1.81 x 10^{-8}) correlating with MMP-1 upon stimulation showed very strong association (P = 1.87 x 10^{-220}) in the SCALLOP consortium data (Figure 6A). We further observed one *trans*-acting pQTL variant rs3014874 exhibiting suggestive association with EN-RAGE protein (5.87 x 10^{-4}), but exceeded the genome-wide significant threshold with P = 1.64 x 10^{-29} in the SCALLOP consortium data (Figure 6B). Evidence from previous studies indicate that this downstream variant (rs3014874) located on chromosome 1 affects the expression levels of *cis*-genes in blood as well as different tissues (Figure 6C). Given that eQTL effects are dependent on context and tissue being studied, we investigated the effect on this SNP on nearby genes in different cell types after candida stimulation. Among all the cis-genes, SNP rs3014874 showed association with only *S100A9* gene (Figure 6D), making it the likely causal gene. This observation further highlights the role of trans-regulatory network in determining the abundance of circulating plasma proteins in blood.

Discussion

In this study we applied integrative analysis of genomics, proteomics and single cell transcriptomics in two independent cohorts to better understand the genetic mechanisms that link mRNA expression and proteins abundance following *Candida albicans* stimulation of immune cells. First, we compared univariate

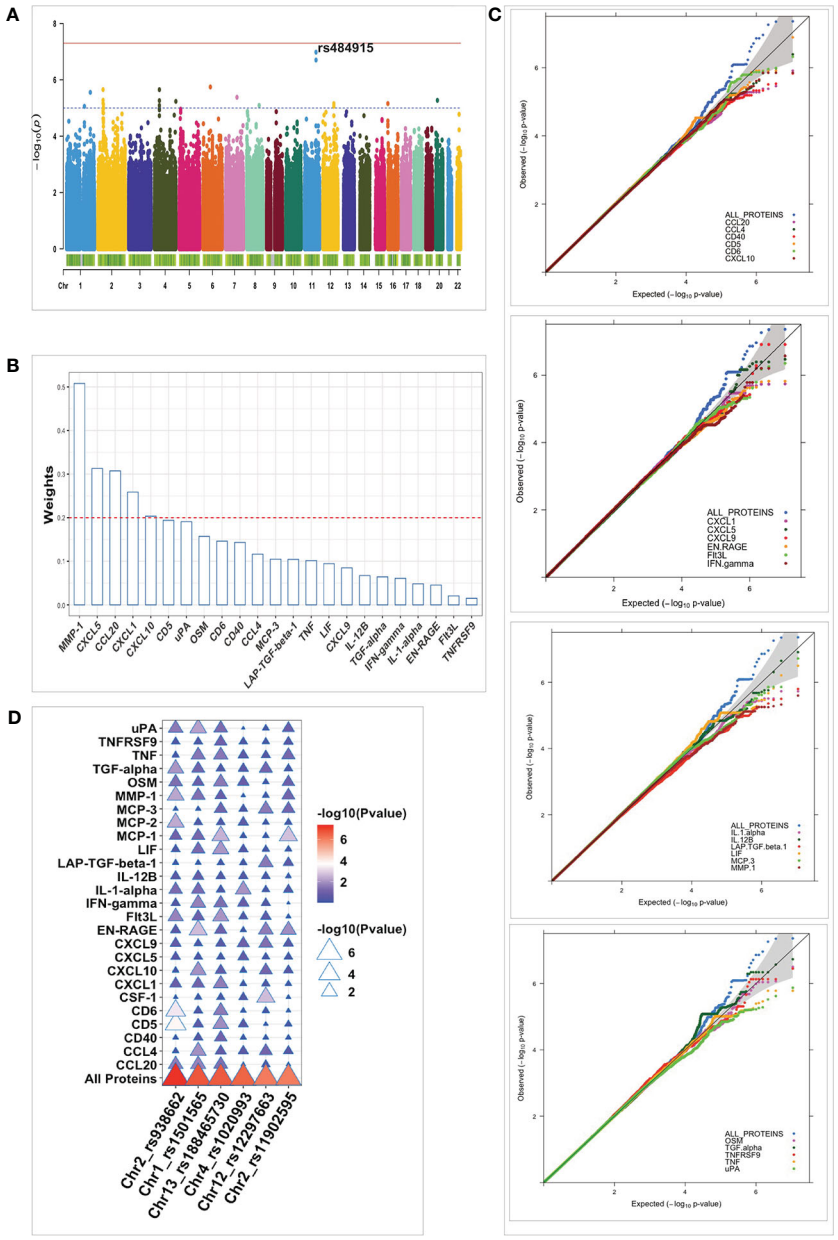

FIGURE 5
Summary of multivariate QTL mapping results and comparison with pQTLs identified using univariate approach. **(A)** Manhattan plot for meta-analyzed pQTL multivariate analysis results of protein network. Strength of association (-log10 meta-p values) is shown on the vertical axis and the horizontal axis depicts the chromosomal position of plotted SNPs. The blue horizontal dashed line and the red bold line represent suggestive (p-value = 1 x10^{-5}) and genome-wide (p-value < 5 x10^{-8}) significant thresholds, respectively. **(B)** Barplot of the weights (contribution to the protein network association) on the vertical axis plotted against all proteins forming the network (horizontal axis) using "ggbarplot" function in R. The red dashed line represents the threshold of significant contributions. **(C)** Quantile-quantile (Q-Q) plots for the pQTL mapping results in the 500GF cohort. The p-values distribution of the multivariate analysis (ALL PROTEINS) results are shown in blue and the remaining colors correspond to p-values of proteins analyzed separately. The gray shaded area represents 95% confidence interval of the null hypothesis. **(D)** Plot of association results of all individual proteins and protein network (All Proteins) plotted against the top six independent SNPs (horizontal axis) identified after multivariate analysis in the 500FG cohort. The color legend represents strength of association of SNPs with protein levels (Pvalue), ranging from blue (weaker associations) to red (stronger associations).

versus multivariate pQTL analyses to test how the cross-trait covariance information which is mostly unutilized in the univariate analysis influence pQTL findings. We then overlaid the identified pQTL SNPs with cell-type-specific eQTL results from PBMCs to help disentangle the underlying mechanism of pQTL results and specify the cell type that could be involved. The strength of this study is the meta-analytic approach of combining two independent population-based cohorts which makes it possible to identify true or consistent genetic associations. Also, for the genes involved in many phenotypes or complex diseases' progression, it is unclear in which cell type gene regulation takes place. Thus, our approach emphasizes the application of cell-type-specific and context-dependent *cis*-eQTL and pQTL data in addressing this challenge.

One of the main observations from our study is that only 35% (7/20) of the pQTLs significantly correlated with mRNA expressions at single cell level, suggesting that eQTLs cannot be used as a proxy for

pQTLs when investigating molecular mechanisms underlying trait-associated variants. The observation of limited overlap is consistent with previously reported findings as the discrepancy between pQTL and eQTL results were also detected in a larger cohort (GTEx Consortium) utilizing over 900 individuals (27). Another recent proteomic study also demonstrated that more than 2000 protein associated variants had no eQTLs (28). Of note, in each of those studies the protein data was a bulk measurement from circulating proteins in the blood (potentially being secreted by any cell type in the body), whereas the matched mRNA data was a bulk measurement from the immune cells themselves. Similarly, in our own study the protein data was a bulk measurement from PBMCs, whereas scRNA-seq data was used to obtain cell-type-specific eQTL data. This discrepancy can potentially result in different significant/top SNPs being identified in the bulk pQTL versus the cell-type-specific eQTL analysis. Given the recent emergence of high-throughput technologies (29, 30) to measure both mRNA and protein levels simultaneously at single cell resolution, future studies could directly compare cell-type-specific pQTL and eQTL results from exactly the same samples. This will provide the most definitive answer regarding the eluded low overlap of both data modalities. Even so, this finding suggests that protein regulation is much more complex than direct mRNA-protein relationship: this is not necessarily surprising, as many post-transcriptional processes are known to influence protein production such as translation, processing and secretion. Therefore, on average low correlation between mRNA and protein QTLs was not unexpected. For instance, 6 different regulatory patterns have been previously described as the mechanisms by which genetic variants can influence the process of transcription to translation (31), such as in scenarios whereby SNPs independently affect transcript levels and protein abundance or SNPs responsible for both transcriptional and translational alterations. Also, it is possible for mRNA decay or extended half-life of the secretory proteins to explain why many pQTL-depenednt effects were observed. Furthermore, proteolytic activities can lead to the limited concordance between pQTLs and eQTLs, which requires data on isoform-specific expressions levels for

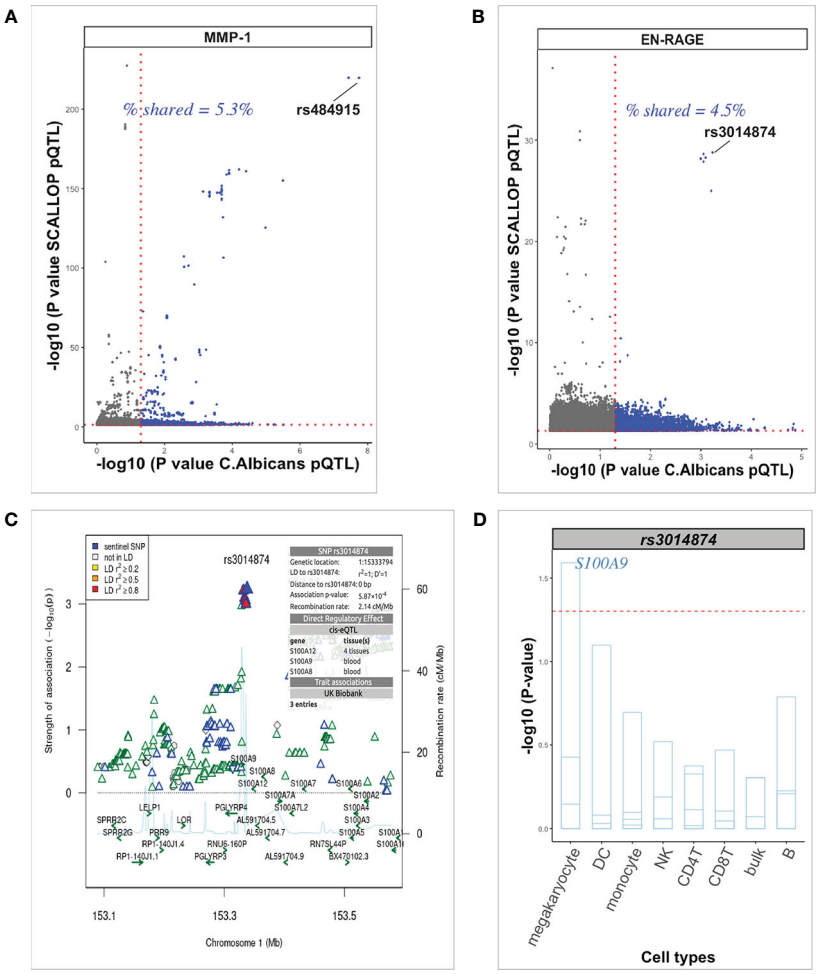

FIGURE 6
Exploration of pQTL variants in plasma from unstimulated samples. **(A, B)** Scatter plots generated with the "ggbarplot" function of pQTLs from unstimulated protein levels (SCALLOP) against candida-induced protein levels showing the proportion of shared pQTLs. Top cis-acting pQTL and trans-acting pQTL are labeled respectively. Dot gray color represents unique pQTLs from the SCALLOP data analysis without stimulation and dot blue color represents the number of overlapping pQTLs irrespective of stimulation status. **(C)** Regional association plot at the S100A9 locus using EN-RAGE pQTL association results after candida stimulation. The top SNP rs3014874 on chromosome 1, with direct regulatory effect on multiple cis-genes is indicated in blue diamond shape. **(D)** Barplot of the top trans-acting pQTL (rs3014874) association on multiple cis-genes (strips in the barplot) in the S100A9 locus in various cell types (horizontal axis). At nominal level (red dashed horizontal line), SNP rs3014874 significantly affect only S100A9 gene in megakaryocyte. .

further investigation. Thus, we advocate for future studies targeting genome-wide *Candida albicans*-induced mRNA (*cis* and *trans* inclusive) and cell-type-specific protein expressions analysis at different time-windows aside 24h, which has the potential to refine this observation and makes it feasible for genome-wide comparison of the proportion of shared or unique pQTL and eQTL variants.

To help with the interpretation of how cell-type-specific eQTLs regulate trans-acting pQTLs, colocalization analysis of lead pQTLs and eQTL signals implicated the *AMZ1* gene. This finding implies that *AMZ1* gene might be involved in the molecular pathways underlying complex diseases, most probably the pathogenesis of candidemia. Also, our analysis therefore predicts the sentinel SNP rs1027658 associated with the trans-acting protein (CCL4) as a probable causal variant and further shows that regulation of CCL4 protein levels is mediated by gene transcription. Apart from the trans-genomic region, similar analysis also showed strong colocalization at the *MMP1* locus with the leading SNP (rs484915) located in the *cis* region, suggesting a direct regulatory effect on MMP-1 protein concentrations.

Another interesting observation made in this study is the strong correlation among proteins concentrations released by human PBMCs upon *Candida albicans* stimulation, suggesting their concerted role in immune regulation. In genetic studies, joint analysis of correlated phenotypes in a single model, a so-called multivariate approach, has been demonstrated to increase statistical power relative to a univariate approach (32–34). Indeed, this was clearly demonstrated in this study as well in the context of Candida-stimulation using the larger cohort (500FG). For example, as the intergenic SNP rs938662 was statistically significant when the multivariate method was adopted, the lowest P value of the same SNP in the univariate approach showed suggestive association (1.45 $x10^{-4}$), correlating with CD5 proteins levels and strikingly, did not show any association with as many as 16/26 proteins analyzed separately.

However, the added value of coupling proteins for joint genetic analysis was not substantially detected after meta-analysis of both cohorts as we expected since multivariate analysis is known to perform relatively better especially, in the case of presence of pleiotropy (35). The dissimilarity in the correlation structure between the two datasets and the relatively weaker correlation strength in the 1M-scBloodNL cohort is mostly capable of causing the multivariate test from not out performing or boosting power than the univariate test after the combined analysis. Even though smaller sample sizes can lead to instability in estimating correlation coefficient (36), technical and experimental variations can partially explain the observed differences in correlation pattern seen in both datasets. Yet, findings from the largest cohort demonstrates that these two approaches are entirely orthogonal to detecting genotype-phenotype relationship.

Identifying such context-dependent pQTLs may have implications in understanding human complex diseases. Supporting this argument, a recent powered study evaluating the relationship between pQTLs and GWAS loci of 81 diseases and other clinical traits found 69 out of 76 (number of phenotypes associated with the genome-wide significant loci investigated) representing 90.8% of the tested genetic associations with phenotypes were also associated with at least one protein with strong evidence of colocalization (37). We therefore need larger studies with context-specific data to show implications of context-specific pQTLs in explaining GWAS findings.

Several limitations of the present study are worth highlighting. First, the use of a specific Olink panel with overrepresentation of inflammatory proteins hinders broad analysis of proteins as the current high-throughput Olink Explore panel is capable of profiling thousands of plasma proteins. Second, sample size limitation made it impossible to comprehensively characterize pQTLs. We therefore acknowledge that upscaling the sample size might help identify more significant genetic loci especially distal QTLs with relatively smaller effective sizes and thus requiring large sample sizes to be detected.

In conclusion, our study has pinpointed several possible mechanisms through which protein levels in circulation are regulated and delineate the specific cell type involved. In addition, we have prioritized candidate genes at pQTL loci, providing great insight into the genetic architecture of secretory protein levels following *Candida albicans* stimulations. We believe that our functional genomic approach can be extended to larger cohorts to obtain mechanistic insights into pathogen-dependent protein regulations.

Data availability statement

The summary-level association statistics of the meta-analyses results have been deposited at the EBI GWAS Catalogue under the accession numbers GCST90244822-GCST90244847.

Ethics statement

The 500FG cohort was approved by the Arnhem-Nijmegen Medical Ethical Committee (500FG: NL42561.091.12) and performed in accordance with the Declaration of Helsinki. All individuals gave written informed consent to donate venous blood for research. The Lifelines DEEP study was approved by the ethics committee of the University Medical Center, Groningen, document number METC UMCG LLDEEP: M12.113965. All volunteers signed an informed consent from prior study of enrollment. The patients/participants provided their written informed consent to participate in this study.

Author contributions

VK and MW designed and supervised the study. MN, MW, and LF contributed with data generation and curation. CB, RO, and KL performed data analysis. CB performed pQTL mapping and other statistical analyses with critical input from VK and CB drafted the manuscript & prepared the figures. All authors contributed to the article and approved the submitted version.

Funding

VK was supported by a Hypatia tenure track grant RadboudUMC. MN was supported by an ERC Advanced grant

(#833247). MW was supported by a NWO Veni grant (#192.029). LF was supported by a NWO-VICI (#917.14.374) and an Oncode Investigator grant.

Acknowledgments

The authors are most grateful to all volunteers from the 500FG and Lifelines DEEP cohorts for participation in the studies.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1069379/full#supplementary-material

References

1. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* (2013) 45(10). doi: 10.1038/ng.2756

2. Mu Z, Wei W, Fair B, Miao J, Zhu P, Li YI. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol* (2021) 22(1). doi: 10.1186/s13059-021-02334-x

3. Zhernakova DV, Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* (2017) 49(1). doi: 10.1038/ng.3737

4. Molendijk J, Parker BL. Proteome-wide systems genetics to identify functional regulators of complex traits. *Cell Syst* (2021) 12. doi: 10.1016/j.cels.2020.10.005

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

6. Li Y, Oosting M, Smeekens SP, Jaeger M, Aguirre-Gamboa R, Le KTT, et al. A functional genomics approach to understand variation in cytokine production in humans. *Cell* (2016). doi: 10.1016/j.cell.2016.10.017

7. Vitrinel B, Koh HWL, Kar FM, Maity S, Rendleman J, Choi H, et al. Exploiting interdata relationships in next-generation proteomics analysis. *Mol Cell Proteomics.* (2019) 18(8). doi: 10.1074/mcp.MR118.001246

8. Assum I, Krause J, Scheinhardt MO, Müller C, Hammer E, Börschel CS, et al. Tissue-specific multi-omics analysis of atrial fibrillation. *Nat Commun* (2022) 13(1). doi: 10.1038/s41467-022-27953-1

9. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: Perspectives for Large population-based studies. *Nat Rev Genet* (2021) 22. doi: 10.1038/s41576-020-0268-2

10. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-Scale integration of the plasma proteome with genetics and disease. *Nat Genet* (2021) 53(12). doi: 10.1038/s41588-021-00978-w

11. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open* (2015) 5(8).

12. Matzaraki V, Le KTT, Jaeger M, Aguirre-Gamboa R, Johnson MD, Sanna S, et al. Inflammatory protein profiles in plasma of candidaemia patients and the contribution of host genetics to their variability. *Front Immunol* (2021) 12.

13. Oelen R, de VDH, Brugge H, Gordon G, Vochteloo M, Consortium B, et al. Single-cell RNA-sequencing reveals widespread personalized, context-specific gene expression regulation in immune cells. *bioRxiv* (2021).

14. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS One* (2014) 9(4).

15. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: An r library for genome-wide association analysis. *Bioinformatics* (2007).

16. Francioli LC, Menelaou A, Pulit SL, Van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* (2014) 46(8).

17. Ricaño-Ponce I, Zhernakova DV, Deelen P, Luo O, Li X, Isaacs A, et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding rnas. *J Autoimmun* (2016).

18. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* (2012) 9(2).

19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. Plink: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* (2007).

20. Ferreira MAR, Purcell SM. A multivariate test of association. *Bioinformatics.* (2009) 25(1).

21. Willer CJ, Li Y, Abecasis GR. Metal: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* (2010). doi: 10.1093/bioinformatics/btq340

22. Whitlock MC. Combining probability from independent tests: The weighted z-method is superior to fisher's approach. *J Evol Biol* (2005) 18(5). doi: 10.1111/j.1420-9101.2005.00917.x

23. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for colocalisation between pairs of genetic association studies using summary statistics. *PloS Genet* (2014) 10(5). doi: 10.1371/journal.pgen.1004383

24. Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* (2019) 51. doi: 10.1038/s41588-019-0404-0

25. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE V2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* (2017) 45(D1). doi: 10.1093/nar/gkw985

26. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun* (2017) 8.

27. Carithers LJ, Moore HM. The genotype-tissue expression (Gtex) project. *Biopreservation Biobanking* (2015) 13.

28. He B, Shi J, Wang X, Jiang H, Zhu HJ. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol* (2020) 18(1).

29. Chung H, Parkhurst CN, Magee EM, Phillips D, Habibi E, Chen F, et al. Simultaneous single cell measurements of intranuclear proteins and gene expression. *bioRxiv* (2021).

30. Reimegård J, Tarbier M, Danielsson M, Schuster J, Baskaran S, Panagiotou S, et al. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol* (2021) 4(1).

31. Wang Y, He B, Zhao Y, Reiter JL, Chen SX, Simpson E, et al. Comprehensive cis-regulation analysis of genetic variants in human lymphoblastoid cell lines. *Front Genet* (2019) 10. doi: 10.3389/fgene.2019.00806

32. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, et al. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PloS Genet* (2012) 8(8). doi: 10.1371/journal.pgen.1002907

33. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. *J Probability Stat* (2012). doi: 10.1155/2012/652569

34. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* (2014) 11(4). doi: 10.1038/nmeth.2848

35. Zhang L, Pei YF, Li J, Papasian CJ, Deng HW. Univariate/Multivariate genome-wide association scans using data from families and unrelated samples. *PloS One* (2009) 4(8). doi: 10.1371/journal.pone.0006502

36. Sari BG, Lúcio AD, Santana CS, Krysczun DK, Tischler AL, Drebes L. Sample size for estimation of the Pearson correlation coefficient in cherry tomato tests. *Ciec Rural* (2017) 47(10). doi: 10.1590/0103-8478cr20170116

37. Gudjonsson A, Gudmundsdottir V, Axelsson GT, Gudmundsson EF, Jonsson BG, Launer LJ, et al. A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat Commun* (2022) 13(1). doi: 10.1038/s41467-021-27850-z