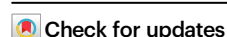


Multi-ancestry sequencing-based genome-wide association study of C-reactive protein in 513,273 genomes

Received: 3 November 2024

Accepted: 14 April 2025

Published online: 24 April 2025

Hongru Li^{1,7}, Jingyi Zhao^{1,7}, Jinglan Dai¹, Dongfang You^{1,2}, Yang Zhao^{1,3}, David C. Christiani^{4,5}, Feng Chen^{1,2,6}✉ & Sipeng Shen^{1,3,6}✉

C-reactive protein (CRP) serves as a pivotal marker of systemic inflammation, yet its genetic architecture has predominantly been explored within European populations. Our multi-ancestry sequencing-based genome-wide association study (seqGWAS) meta-analysis encompasses 447,369 Europeans, 10,389 Africans, 9685 Asians, and 9200 Hispanics in the discovery set, and 23,521 Europeans, 7160 Africans, 771 Asians, and 5178 Hispanics in the replication set. We identify 113 independent association signals ($P_{\text{discovery}} \leq 5 \times 10^{-9}$ and $P_{\text{replication}} \leq 0.05$), including 21 loci that passed the conditional analysis, among which 3 are European-specific. Cross ancestry fine-mapping pinpoints 19 of 113 independent signals within the 95% credible set. Functional annotation reveals significant enrichment in blood tissue, H3K27me3 histone marks, and exonic regions. Leveraging the Polygenic Priority Score (PoPS) and gene-based analyses, we implicate 151 genes as potential regulators of CRP levels, 55 of which have not been previously reported. Among these, 17 genes and four proteins show causal evidence or strong colocalization with CRP-related pathologies.

C-reactive protein (CRP) is a key member of the acute phase reaction protein (APRP) family and has been extensively investigated as a biomarker for systemic inflammation. Accumulating evidence has demonstrated that CRP levels are associated with a diverse range of diseases, including cardiovascular diseases (CVDs)¹, certain malignancies^{2,3}, diabetes mellitus^{4,5}, and immune-mediated diseases (IMDs)⁶.

Genetic factors play a significant role in determining serum CRP levels⁷. A previous genome-wide association study (GWAS) identified 266 independent loci based on SNP-array data from 575,531 individuals of European descent⁸. However, array-based studies face several limitations, particularly in capturing ancestral diversity and the

contribution of rare variants (RVs) with allele frequencies <0.01 ⁹. Specifically, current GWAS struggles to fully characterize the human genetic landscape due to limitations in available reference panels, such as the 1000 Genomes Project, which has a limited sample size, and the Haplotype Reference Consortium, which lacks sufficient ethnic diversity. Furthermore, recent studies suggest that the missing heritability of complex traits and diseases, as well as the identification of causal variants, may be partly attributed to RVs^{10,11}. Array-based imputation, however, is not well-suited to capture RVs, which constitute a significant portion of human genetic variation¹². Whole-exome sequencing (WES) studies have identified significant protein-coding RVs, which account for ~2% of the human genome, in the context of

¹Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China. ²China International Cooperation Center of Environment and Human Health, Nanjing Medical University, 211166 Nanjing, China. ³Key Laboratory of Biomedical Big Data of Nanjing Medical University, Nanjing 211166, China. ⁴Department of Environmental Health, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA. ⁵Pulmonary and Critical Care Division, Massachusetts General Hospital, Department of Medicine, Harvard Medical School, Boston, MA 02114, USA. ⁶Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, 211166 Nanjing, China. ⁷These authors contributed equally: Hongru Li, Jingyi Zhao. ✉e-mail: fengchen@njmu.edu.cn; sshen@njmu.edu.cn

complex traits and diseases^{13–15}. In comparison to SNP arrays and WES, whole-genome sequencing (WGS) offers a more comprehensive view of the genome, allowing for the detection of both coding and non-coding genetic variations and the exploration of complex genomic regions that are challenging to genotype using traditional approaches, showing efficiency for evaluation of RVs¹⁶. Moreover, multi-ancestry studies can enhance the power and fine-mapping resolution of GWAS, thereby improving the prediction, prevention, diagnosis, and treatment of complex traits and diseases across diverse populations^{17,18}. Therefore, our multi-ancestry sequencing-based genome-wide association study can effectively overcome the aforementioned limitations, providing a more comprehensive framework for the in-depth exploration of CRP and its genetic underpinnings.

In the present study, we performed a large-scale multi-ancestry sequencing-based genome-wide association study (seqGWAS) of CRP leveraging data from UK Biobank (UKB) and the Trans-Omics for Precision Medicine (TOPMed) program, including Atherosclerosis Risk in Communities Study (ARIC), Multi-Ethnic Study of Atherosclerosis (MESA), Framingham Heart Study (FHS), Cleveland Family Study (CFS), and Hispanic Community Health Study – Study of Latinos (HCHS-SOL) as the discovery set, with All of Us (AoU) data serving as an independent external replication set. Using the Polygenic Priority Score (PoPS) and gene-based analyses, we identified 151 genes as potential regulators of CRP levels, 55 of which were not previously reported. Among them, 17 genes and four proteins provided causal evidence or strong

colocalization with CRP-related pathologies. As the largest CRP seqGWAS to date, these findings advance our understanding of the genetic basis of CRP and its involvement in complex diseases, while underscoring the need for more diverse genomic research to refine ancestry-specific insights.

Results

Sequencing-based genome-wide association analysis of CRP

We undertook seqGWAS of CRP from seven cohorts (Supplementary Methods, Fig. 1, Supplementary Fig. 1). The UKB, ARIC, FHS, MESA, CFS, and HCHS-SOL were used as the discovery set, and AoU was used as an independent external replication set. Our multi-ancestry study of up to 513,273 participants comprised individuals of Hispanic ($n=14,378$), Asian ($n=10,456$), African ($n=17,549$), and European ancestry ($n=470,890$) (Supplementary Data 1). In the discovery set, we identified 70,437 genetic variants significantly associated with CRP levels in European ancestry group ($P < 5.0 \times 10^{-9}$). Notably, by adopting a more stringent significance threshold (5.0×10^{-9} vs. the previously used 5.0×10^{-8}), we identified an overlap of 238 loci from the 315 loci reported in our prior study^{7,8}. In other ancestry groups, we identified 660 significant associations in Asian populations, 635 in African populations, and 1 in Hispanic population (Supplementary Figs. 2–5, Supplementary Data 2–5).

In the discovery set, we performed seqGWAS cross-ancestry meta-analysis. The genomic inflation factor λ of the multi-ancestry meta-

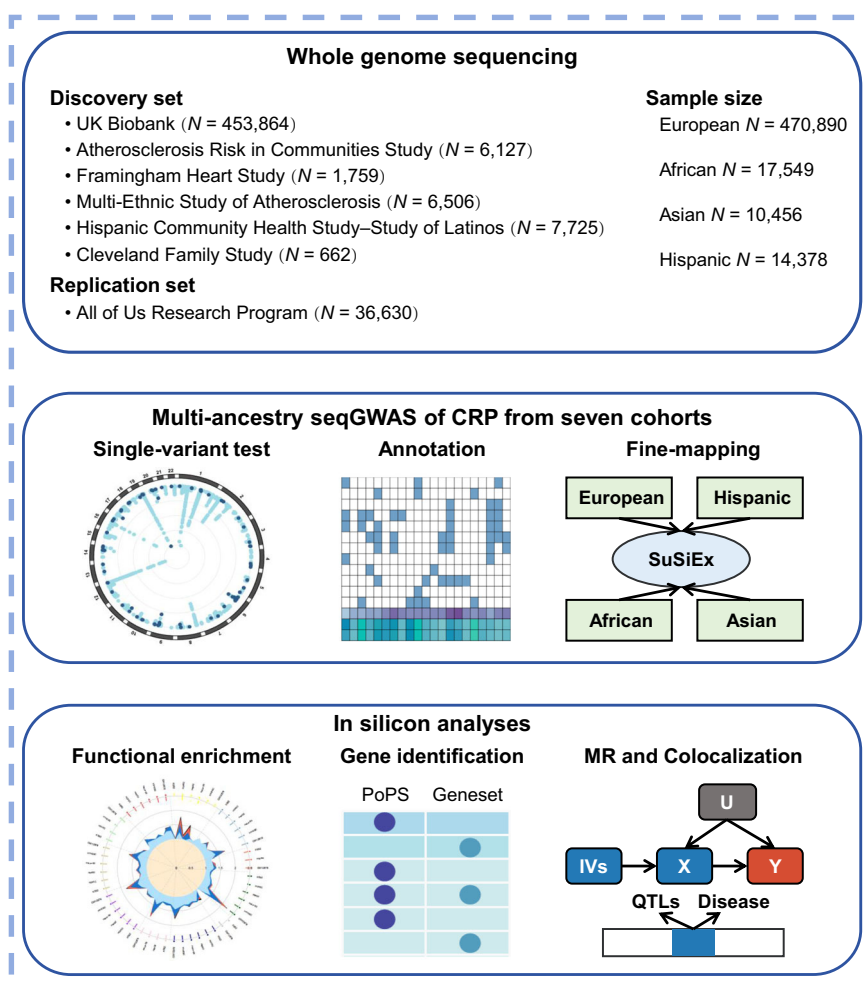


Fig. 1 | Study workflow. Firstly, we acquired multi-ancestry data of WGS and CRP from UKB, TOPMed, and AoU. Secondly, we executed unconditional and conditional seqGWAS of CRP in six cohorts, AoU data serving as an independent external

replication set and further carried out a multi-ancestry fine-mapping analysis. Finally, we conducted functional enrichment analysis, PoPS, gene-based, MR, and colocalization analysis for European-ancestry.

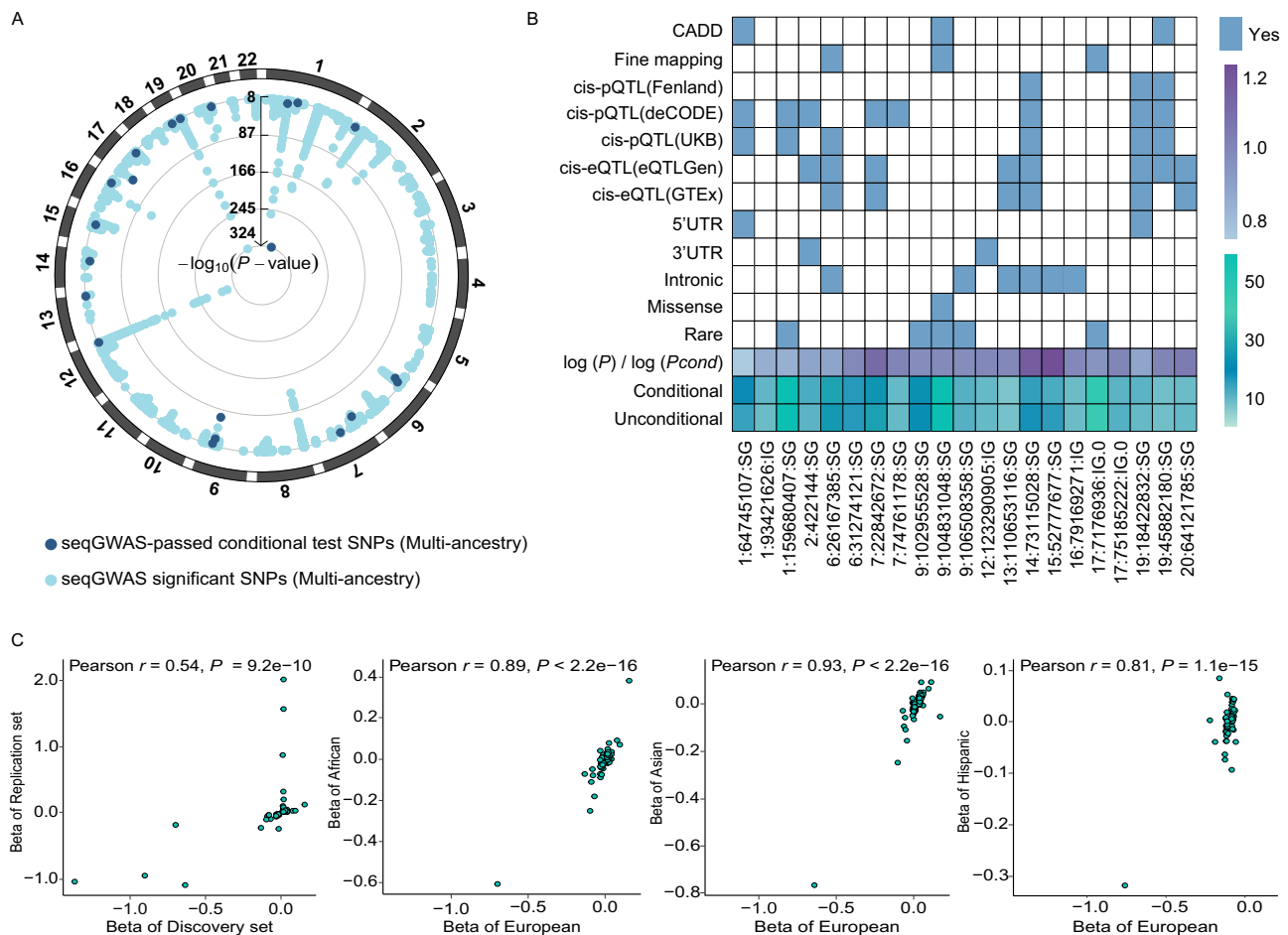


Fig. 2 | Genetic architecture of CRP. A Circle Manhattan plot. Genome-wide significant hits at $P < 5.0 \times 10^{-9}$ are presented for multi-ancestry meta-analysis (light blue dots). Dark blue dots indicate 21 loci that passed the conditional analysis. The Bonferroni-corrected significant two-sided P value of $P \leq 5.0 \times 10^{-9}$. **B** The blue in the first 12 rows indicates that 21 signals have functional annotations. Where cis-eQTLs or cis-pQTLs are defined as within a region of ± 1 MB ($\text{FDR-}q < 0.05$), and

CADD is defined as a score greater than 10. The last three rows indicated the level of seqGWAS of ratios, conditional, and unconditional. **C** Scatterplot with the effect for 113 distinct signals from the meta-analysis, comparing the Discovery set vs. replication set, and European-ancestry vs. the other three ancestries (African-ancestry, Asian-ancestry, and Hispanic-ancestry). The Bonferroni-corrected significant two-sided P value of $P \leq 7.1 \times 10^{-3}$.

analysis was 0.997 (Supplementary Fig. 6). We identified 27,011 genetic variants for CRP using a stringent threshold of $P < 5.0 \times 10^{-9}$ in the multi-ancestry meta-analysis (Fig. 2A, Supplementary Fig. 7). Of these, 2,320 were nominally associated ($P < 0.05$ and same direction of effect) with CRP in independent replication set. Using a distance-based approach, we identified 113 distinct signals, we compared the effect size, and moderate consistency was found in discovery and replication (Pearson $r = 0.54$, $P = 9.2 \times 10^{-10}$). We found a high consistency between different ancestries (Pearson $r = 0.72 - 0.93$) (Fig. 2C, Supplementary Data 6, Supplementary Fig. 8). The variance of CRP explained by all independent signals was 8.84%. Among the 113 distinct signals, 21 signals passed the conditionality test ($\log_{10}P / \log_{10}P_{\text{conditional}} < 1.5$ and $P_{\text{conditional}} < 5.0 \times 10^{-9}$, Fig. 2B). Of the 21 signals, three signals were identified in Europeans distinctly (Supplementary Data 6). To elucidate comprehensively these signals, we incorporated systematically various forms of relevant evidence. Among the 21 signals identified, five were rare and three were within the fine mapping 95% credible set. The functional annotations of the loci included intronic, one missense mutation variant (9:104831048:SG in *ABCA1* [$\beta = -1.367$, $P = 1.35 \times 10^{-71}$, $\text{MAF} = 0.0002$]), 3'UTR and 5'UTR. Notably, three deleterious variants with Combined Annotation Dependent Depletion [CADD] > 10 were identified, including the previously mentioned missense mutation variant, along with the following two variants: 1:64745107:SG and 19:45882180:SG. In terms of expression quantitative trait Loci (eQTL)

and protein quantitative trait Loci (pQTL) analysis, we identified eight significant cis-eQTLs and nine significant cis-pQTLs (Fig. 2B, Supplementary Data 7).

Cross-ancestry credible set analysis of CRP-associated loci

Fine mapping across diverse ancestries promotes the discovery of putative causal variants underlying CRP-associated genomic loci, identifying 95% credible sets of variants based on the Sum of Single Effects (SuSiE) framework. 19 out of 113 independent signals were included in the 95% credible set. Five (4.4%) signals had a single putative causal variant (posterior probability $> 95\%$), one (0.9%) signal had two variants within the 95% credible set, and 11 (9.7%) signals had ≤ 5 variants within the 95% credible set (Supplementary Data 8).

Functional enrichment analysis

After completing the association analysis, we used GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD) to assess the enrichment of our signals for cell-type-specific regulatory and functional features, providing deeper insight into the biological relevance of the association signals. Using GARFIELD, we assessed the enrichment of our signals for chromatin accessibility peak and DNase I hypersensitivity site, which showed higher enrichment in both blood and fetal intestine (Fig. 3A,

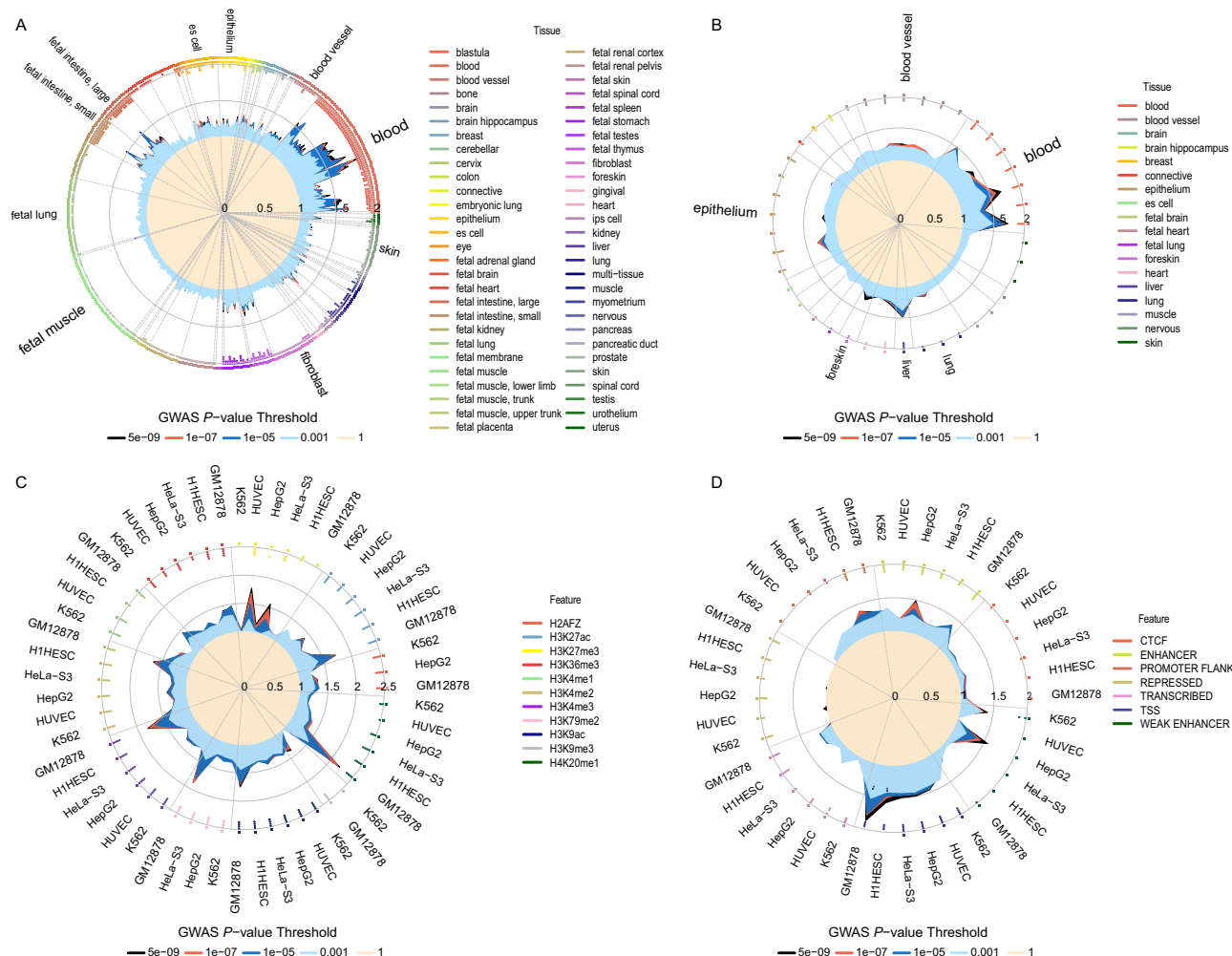


Fig. 3 | Functional enrichment analysis. The wheel plots display functional enrichment for associations with CRP within different types of functional annotation regions in the ENCODE and Roadmap Epigenomics project: **A** Open chromatin peaks **B** Transcription factor footprints **C** Histone modifications **D** Chromatin segmentation states.

Supplementary Fig. 9A). Transcription-factor footprints and Transcription factor binding sites were mainly enriched in blood and liver (Fig. 3B, Supplementary Fig. 9B). Our signal shows enrichment in a wide variety of histone modifications, including higher enrichment in H3K27me3, H4K20me1, and H3K9ac (Fig. 3C). Chromatin segmentation states were mainly enriched in Transcription Start Site (TSS) and ENHANCER (Fig. 3D). Genic annotation was mainly present in exon regions and FAIRE enrichment was mainly located in liver (Supplementary Fig. 9C, Supplementary Fig. 9D).

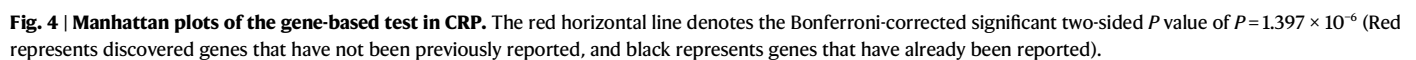
Identification of CRP-related genes

Gene-prioritization strategies can be placed into two broad categories: locus-based and similarity-based methods. A similarity-based method for gene prioritization, the Polygenic Priority Score (PoPS), has been shown to improve gene prioritization¹⁹. In our study, PoPS was leveraged to summary-level data of the European ancestry meta-analysis. PoPS analysis of 113 distinct CRP-associated signals prioritized 107 candidate genes corresponding to 105 genetic signals (Supplementary Data 9). Notably, PoPS focused only on the effect of common variants ($MAF > 0.01$). To capture the effects of rare and ultra-rare variants, gene-based association studies were performed. After Bonferroni correction, 32 unique protein-coding genes (Mask: LoF and missense variants) and 18 unique ncRNAs (Mask: ncRNA exonic, ncRNA exonic-splicing, and ncRNA splicing variants) were considered significant. ($P < 1.397 \times 10^{-6}$, Fig. 4, Supplementary Figs. 10–11, Supplementary Data 10).

Through comprehensive PoPS and gene-based analyses, we identified 151 unique CRP-associated genes. Six genes were identified by both methods, including *ABCA1*, *APOE*, *FDFT1*, *GPR146*, *NEK7*, and *SALL1*. Previous studies have associated these genes with metabolic, inflammatory, and cancers. For instance, *ABCA1* regulates mTORC1 activity by regulating cholesterol metabolism and is involved in aging and inflammation²⁰. *APOE*, the product of *APOE* gene, plays a key role in lipid metabolism, immune regulation, and neurology^{21,22}. *FDFT1* is a key downstream target of the fasting response and may be involved in colorectal cancer (CRC) cell glucose metabolism²³. *GPR146*, encoded by the *GPR146* gene, activates the liver CAMP-PKA-CREB pathway and regulates lipid metabolism²⁴.

Mendelian randomization and Colocalization analyses

To explore the broader phenotypic implications of CRP-related genes, we performed additional Mendelian randomization (MR) and colocalization analysis involving 55 diseases that were associated with CRP. Our multi-faceted analytical framework systematically integrated diverse lines of evidence through seven complementary approaches: (1) identification of genes with the highest polygenic priority score (PoPS); (2) rare variant gene-based association analyses; (3) colocalization analyses between 55 CRP-associated diseases and expression quantitative trait loci (eQTLs); (4) colocalization analyses between 55 CRP-associated diseases and protein quantitative trait loci (pQTLs); (5) Mendelian randomization analyses of 55 CRP-associated diseases with eQTLs; and (6) Mendelian randomization analyses of 55



A total of 116 unique genes were covered in the eQTLs databases (GTEx and eQTLGen), and 27 unique proteins were included in the pQTLs databases (UKB pQTLs, deCODE pQTLs, and Fenland pQTLs) in colocalization analyses. We identified 32 candidate genes and 9 candidate proteins (posterior probability of colocalization ≥ 0.8) through colocalization (Fig. 5C, D, Supplementary Fig. 12A–C, Supplementary Data 18–19). Among these candidate genes (proteins), 13 candidate genes and four candidate proteins were identified as previously unreported. For example, Colocalization evidence for *SNX17* was detected in type 2 diabetes, gout, atopic dermatitis, Crohn's disease, and allergic rhinitis. *BCL3* was associated with ischemic heart disease and myocardial infarction based on colocalization, and MR evidence

In the multi-ancestry meta-analysis, we identified 113 distinct signals. Among these, 21 were classified based on conditional analysis, and three were identified as European-specific. One notable finding is the missense mutation 9:104831048:SG (CADD = 25), located at locus 9q31.1 (ATP Binding Cassette Subfamily A Member 1 [*ABCA1*]), which is annotated in ClinVar with clinical significance of pathogenic. *ABCA1* was similarly reported in the genetic study of CRP by Said et al.⁸. *ABCA1* functions as a cholesterol efflux pump in the cellular lipid removal pathway. Senescence was associated with the upregulation of the cholesterol exporter *ABCA1*, which is rerouted to the lysosome, inducing lysosomal microdomains rich in mTORC1 scaffold complexes. This metabolic shift maintains mTORC1 activity to support the expression of the pro-inflammatory senescence-associated secretory phenotype²⁰. In addition, we also found a significant association between *ABCA1* and *CRP* in gene-based analysis. The prioritized gene for potential pathogenic variant 19:45882180:SG (CADD = 19.27) is Glucose-dependent insulinotropic polypeptide receptor (*GIPR*). This

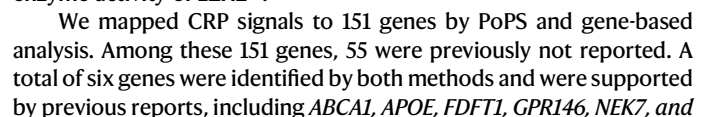


Fig. 5 | Plot of MR and colocalization analysis of CRP-associated diseases.

A Heatmap displaying functional evidence for CRP-related genes (The amount of evidence ≥ 3). Columns 2 and 3 represent PoPs and geneset tests, included based on significant threshold criteria. Columns 4 to 6 represent colocalization and MR evidence, indicating the association of genes with at least one CRP-related disease. Column 7 specifies whether CRP-related genes have been previously reported in the literature and genes that have not been reported are included. **B** Forest plots of the candidate genes and proteins that have not been previously reported in Mendelian randomization (MR) analyses incorporating expression quantitative trait loci (eQTLs) and protein quantitative trait loci (pQTLs). Samples from GTEx ($N = 800$); eQTLGene ($N = 31,684$); Fenland ($N = 10,708$) were included in the analyses. The x-axis represents OR (95% CI). The points are the odd ratio (OR) estimates from the MR analyses, and the error bars are the 95% confidence intervals. **C** eQTLGene eQTLs. **D** GTEx eQTLs. Statistical tests were two-sided. The red horizontal line denotes the FDR- q of MR < 0.05 or posterior probability of colocalization ≥ 0.8 . The upper figure represents MR, and the lower figure represents colocalization (black represents genes that are both significant in MR and colocalization, light gray represents genes that are only significant in MR or colocalization, where red

represents genes that have not been previously reported and are both significant in MR and colocalization, and light gray represents genes that have not been previously reported and are only significant in MR or colocalization). AA Alopecia areata, ACD Allergic contact dermatitis, AD Atopic dermatitis, AP Allergic purpura, AR Allergic rhinitis, AS Ankylosing spondylitis, AU Allergic urticaria, Bullous Bullous disorders, GB Guillain-Barre syndrome, Hepatitis Autoimmune hepatitis, HPT Hypothyroidism, ID Immunodeficiency with predominantly antibody defects, ITP Idiopathic thrombocytopenic purpura, Lichen Lichen planus, MG Myasthenia gravis, MS Multiple sclerosis, NV Necrotizing vasculopathies, PA Pernicious anemia, PBC Primary biliary cirrhosis, PR Polymyalgia rheumatica, PST Psoriatic and enteropathy arthropathies, RF Rheumatic fever, SD Sarcoidosis, Sicca Sicca syndrome, SLE Systemic Lupus erythematosus, UC Ulcerative colitis, MI Myocardial infarction, T1D Type 1 diabetes, T2D Type 2 diabetes, HF Heart failure, IHD Ischemic heart disease STR stroke, Lung ca Lung cancer, BRCA Breast invasive carcinoma, CRC Colon adenocarcinoma, HNSC Head and neck squamous cell carcinoma, HCC Hepatocellular carcinoma, KIPAN Pan-kidney cohort, NHL Non-Hodgkin's lymphomas, ESCA Esophageal carcinoma, STAD Stomach adenocarcinoma, COPD chronic obstructive pulmonary disease.

SALL1. The aforementioned genes are closely associated with metabolism, inflammation, and cancer. For example, *ABCA1* was identified as the disease-causing gene for Tangier disease (TD) – a rare genetic disorder characterized by severe reductions in HDL and a high incidence of premature CVD^{34–36}. A growing body of research has linked apolipoprotein E (*APOE*) to various neurodegenerative proteinopathies and broader age-related brain changes, including neuroinflammation, energy metabolism failure, loss of myelin integrity and increased blood-brain barrier permeability. These findings have potential implications for longevity and resilience to pathological protein aggregates. Several therapeutic strategies targeting *APOE* have demonstrated efficacy in preclinical studies and hold promise for translation into clinical trials³⁷. *FDFT1* exerts its tumor inhibitory function by downregulating the AKT/mTOR/HIF1 α signaling pathway²³. MR and colocalization analysis revealed extensive associations between CRP candidate genes and related phenotypes. For instance, we observed a causal association between Phosphatidylinositol Transfer Protein Membrane Associated 2 (*PITPNM2*) and four diseases (asthma, type 2 diabetes, atopic dermatitis, and allergic rhinitis). *PITPNM2* is most highly expressed in late double-negative and double-positive thymocytes and is downregulated following positive selection. It plays a key role in both T cell antigen receptor β selection and positive selection in thymocytes. Additionally, *PITPNM2* supports the survival of mature T cells in the periphery²⁶. B-cell lymphomas 3 (*BCL3*) has been linked to ischemic heart disease and myocardial infarction through colocalization, and Mendelian randomization (MR) evidence further supports its association with myocardial infarction. Recent findings from the past decade suggest that, in addition to its well-established role as a co-factor in Nuclear Factor Kappa B signaling, the *BCL3* protein also influences cancer progression and therapy resistance through distinct mechanisms. These mechanisms involve protein modifications and interactions with alternative oncogenic players, including myelocytomatosis, Wingless-related integration site (WNT)/ β -catenin, and Signal Transducer and Activator of Transcription 3²⁹. *BCL3* was identified as a key transcription factor in a study of inflammatory transcriptome features associated with cardiovascular events in psoriasis³⁸.

Our work has several strengths. First, through pooling multiple population-scale and multi-ethnic WGS datasets, this is the largest seqGWAS for CRP so far. We identified genetic associations for CRP and conducted independent external replication, including rare variants and protein-coding variants that were ignored by previous GWASs. Second, we considered functional enrichment involving various tissues and histone modifications. Third, potential causal relationships were investigated through multiple approaches, such as SuSiEx, PoPS, gene-based, MR, and colocalization.

Our multi-ancestry sequencing-based meta-analysis maximized the ability to detect CRP-associated signals across ancestral populations. However, the frequency of ancestry-specific variants is generally low³⁹, the power to detect ancestry-specific variants is still limited by the sample size available within each ancestral group. Therefore, we focused on multi-ancestry rather than ancestry-specific signals. The statistical methods we employed, including the additive random effects model and inverse variance fixed effects model, effectively account for genetic heterogeneity between populations. Heterogeneity is more likely to limit discovery rather than introduce false positives⁴⁰. Furthermore, incomplete sampling of non-European ancestries have led to underrepresentation of these populations in genomic studies⁴¹. Non-European ancestries are underrepresented in genomic studies, which constrains the effectiveness of GARFIELD functional enrichment, PoPS, gene-based, MR, and colocalization analyses in these populations. Therefore, future research should place greater emphasis on increasing the representativeness of non-European ancestry samples to enable a more comprehensive assessment of genetic variation across diverse populations. With the continuous advancement of global genomic research, expanding sample sizes from all ancestries will enhance the capacity and precision of ancestry-specific studies and improve the fine mapping of signals in multi-ancestry meta-analyses, ultimately increasing the generalizability and accuracy of the research findings.

In conclusion, our multi-ancestry seqGWAS study contributes insights into the genetic determinants of CRP, which enhance our understanding of the mechanisms underlying CRP thus laying a foundation for future research by informing functional genomics experiments. Ultimately, this knowledge could guide the development of interventions for a range of complex diseases such as cardiovascular diseases, cancers, and immune-mediated disorders, broadening the scope of potential therapeutic targets.

Methods

Study population of the C-reactive protein WGS

In this multicenter study, we performed seqGWAS across several cohorts from the UK Biobank (UKB) WGS-500k project and TOPMed as a discovery set. We collected CRP and WGS data from the Atherosclerosis Risk in Communities Study (ARIC), the Multi-Ethnic Study of Atherosclerosis (MESA), the Framingham Heart Study (FHS), the Cleveland Family Study (CFS), and the Hispanic Community Health Study – Study of Latinos (HCHS-SOL) in TOPMed. The All of Us (AoU) WGS data was used as an independent external replication. The study population comprised 470,890 individuals of European ancestry, 17,549 individuals of African ancestry, 10,456 individuals of Asian ancestry, and 14,378 individuals of Hispanic ancestry. Details regarding

the population characteristics and genotype data are provided in Supplementary Methods. All the data analyses in UKB were performed on DNAnexus Research Analysis Platform (RAP) under application 92675. All studies were approved by the corresponding institutional review boards (Supplementary Methods).

Quality control for the whole-genome sequencing data

The whole-genomes of 490,640 UKB participants were sequenced to an average coverage of 32.5x (>23.5x per individual) using Illumina NovaSeq 6000 sequencing machines. Initial quality control was performed by deCODE and Wellcome Sanger, which included sex discordance, contamination, unresolved duplicate sequences, and discordance with microarray genotyping data checks. 1,037,556,156 SNPs and 101,188,713 indels were called using GraphTyper⁴². A large majority of variants, 1,025,188,151 (98.80%) SNPs and 97,190,353 (96.05%) indels were reliable (AAscore > 0.5 and <5 duplicate inconsistencies). We further processed the jointly called genotype data in Hail v0.2, where multiallelic sites were first split and normalized. Variants were then filtered based on low allelic balance (ABHet <0.175, ABHom <0.9), low quality-by-depth normalized score (QD <6), low phred-scaled quality score (QUAL <10) and high missingness (call rate <90%).

TOPMed's Freeze 9 includes more than 80 different studies with about 158,000 samples with whole-genome sequencing (WGS), which targeted a mean depth of at least 30x (paired-end, 150-bp reads) using Illumina HiSeq X Ten instruments was carried out over several years at six sequencing centers. All sequences were remapped to the hs38DH 1000 Genomes build 38 human genome reference, incorporating decoy sequences, using BWA-MEM algorithm⁴³. Variant discovery and genotype calling were conducted in a unified manner across TOPMed studies for all samples within each freeze, employing the GotCloud pipeline to ensure comprehensive analysis⁴⁴.

The WGS library in the All of Us dataset is constructed using the PCR Free Kapa HyperPrep kit, followed by sequencing on the Illumina NovaSeq 6000. Primary QC includes library quantification via qPCR, pooling for sequencing, and removal of non-conforming samples. Data are processed on the Illumina DRAGEN platform with GRCh38dh reference alignment, generating high-resolution QC metrics at each analysis stage. Predefined QC measures ensure rigorous control, including coverage thresholds ($\geq 30\times$), genome coverage ($\geq 90\%$ at $20\times$), and contamination limits ($\leq 1\%$)⁴⁵. The Allele Count/Alele Frequency (ACAF) threshold dataset was used in the analysis, which included variants with population-specific allele frequency AF > 1% or population-specific allele count AC > 100, in any computed ancestry subpopulations. The remaining rare variants were extracted from the Hail VariantDataset (VDS).

Sequencing-based genome-wide association analysis of CRP

REGENIE (v3.2.6) was used for association analysis of CRP levels across all autosomes⁴⁶. REGENIE is a toolkit developed for genome-wide association tests using machine-learning method for fitting a whole-genome regression model in biobank-level datasets. CRP was rank-based inverse-normal transformed after adjustment for age, sex, and the top 10 genetic principal components. For the replication set, additional correction was applied to the CRP measurement method. Referring to the approach described by Li et al.⁴⁷, we defined a list of associated loci ("top hits") for CRP based on a distance-based criterion: within each chromosome, variants $P < 5.0 \times 10^{-948}$ were identified, the variant with the smallest P -value was selected within a ± 500 KB window. This process was repeated recursively until no additional variants with $P < 5.0 \times 10^{-9}$ remained.

Multi-ancestry genome-wide Meta-analysis

We utilized the SNP effect estimates and standard errors for each SNP-CRP site pair from the candidate list in the meta-analysis.

Inverse-variance fixed effect (FE) meta-analysis based on the data sets with the same ancestry was performed on the cleaned files using METAL⁴⁹.

The trans-ethnic meta-analysis was performed using the additive random effects (RE) model of DerSimonian-Laird⁵⁰, implemented the¹² modified METAL of Min et. al. (<https://github.com/explodecomputer/random-metal>). The variance explained was calculated for the variants within the top hits of the multi-ancestry seqGWASs results using the formula⁷: $\sum_{i=1}^n \left[\frac{2 \times \text{MAF}_i (1 - \text{MAF}_i) \times \beta_i^2}{\text{var}(\ln \text{CRP})} \right]$, where $\sum$ is the sum, MAF_i is the MAF of associated variant i , β_i is the absolute effect estimate of the corresponding variant i and var is the variance of natural log CRP levels obtained from Said S et al. study⁸.

Conditional analysis

We selected 113 independent signals based on the distance criterion. For conditional analysis using REGENIE, we included 315 independent CRP-associated signals reported in previous study^{7,8}. To determine whether the above-reported sentinels correspond to the same nearby signal from our results, we conditioned our CRP signals on the previously reported nearby sentinels. Following the definition of a signal proposed by Shrine et al.⁵¹, we compared the original P -value with the conditional P -value of our signals to determine whether the reported sentinel predominantly explains our CRP signal. Specifically, we classified our signal as being explained by a previously reported signal if there was substantial attenuation, defined by a $\log_{10} P / \log_{10} P_{\text{conditional}}$ ratio > 1.5 and $P_{\text{conditional}} > 5.0 \times 10^{-9}$. The Combined Annotation Dependent Depletion (CADD) of signals was derived from Schubach et al.⁵².

Cross-population fine-mapping analysis

SuSiEx⁵³ extends the single-population fine-mapping model, Sum of Single Effects (SuSiE)⁵⁴, by integrating population-specific GWAS summary statistics and linkage disequilibrium (LD) reference panels from multiple populations. Using the SuSiE framework, each single effect is coupled with the assumption that the causal variants are shared across populations. SuSiEx was conducted on the CRP-associated genomic loci to identify likely causal variants. The resulting variants of each locus to the identification of likely causal variants. The resulting variants of each locus constitute the smallest list of variants that cumulatively have a $\geq 95\%$ probability of encompassing causal variants. For the LD reference panel, 10,000 Europeans were randomly sampled from the UK Biobank (UKB), whereas for African and Asian populations, the complete datasets from the UKB were utilized, while for the Hispanic population, the HCHS-SOL dataset was used to maximize the robustness of ancestry-specific analyses.

Genomic regulatory and functional enrichment analysis

We used GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD)⁵⁵ with default parameters to perform functional annotations, covering regions including DNase I hypersensitivity hotspots, open chromatin peaks, transcription-factor footprints and formaldehyde-assisted isolation of regulatory elements, histone modifications, chromatin segmentation states, genic annotations and transcription-factor binding sites. Based on European meta-analysis (to make the analysis more computationally feasible, only variants with $P < 0.1$ were included), we applied GARFIELD to DNase I hypersensitivity hotspots annotation across 424 cell lines and primary cell types from ENCODE and Roadmap Epigenomics, deriving enrichment estimates at CRP-genotype association P -value thresholds of $P < 1.0 \times 10^{-3}$, $P < 1.0 \times 10^{-5}$, $P < 1.0 \times 10^{-7}$ and $P < 5.0 \times 10^{-9}$.

Polygenic Priority Score for gene prioritization

Polygenic Priority Score (PoPS) is a similarity-based gene prioritization method that uses extensive set of bulk and single-cell expression

datasets, curated biological pathways and predicted protein-protein interactions¹⁹. For each locus, the gene with the numerically highest PoPS score was determined to be the PoPS-prioritized gene. In our study, we used Polygenic Priority Score v.0.2 (<https://github.com/FinucaneLab/pops>), incorporating summary statistics from European-specific meta-analysis and a reference panel of 10,000 randomly selected Europeans from the UKB. We gave priority to genes for all autosomal CRP signals within a 500-kb (± 250 kb) window centered on the sentinel and identified the top-prioritized genes in the region. For the signals lacking prioritized genes within the 500-kb window, we extended the search to a 1-MB (± 500 kb) window.

Gene-set association analysis

To evaluate the impact of rare (minor allele frequency [MAF] <0.01) and ultra-rare variants (minor allele count [MAC] <10), we performed gene-level collapsing analysis. In constructing the mask, functional annotations and alternative allele frequency (AAF) cutoff values below 1% were used.

Our primary focus was on rare and ultra-rare protein-coding variants (loss-of-function (LoF) and missense variants) and non-coding variants (ncRNA exonic, ncRNA exonic-splicing, and ncRNA splicing variants). For each gene-level collapse, we report the association *P*-value derived using the ACAT-O method, implemented in REGENIE (v3.2.6).

All association analyses were adjusted for covariates, including age, sex, and the top 10 principal components (PCs), to account for potential confounding factors and mitigate population stratification. *P*-values from different datasets were aggregated using the Aggregated Cauchy Association Test (ACAT)⁵⁶. Bonferroni correction was used to determine genome-wide significance thresholds for gene-level associations ($0.05 / 35,790 = 1.397 \times 10^{-6}$, 18,949 protein-coding genes and 16,841 ncRNAs).

Identification of eQTLs and pQTLs Mendelian randomization and colocalization of priority genes and CRP-related diseases

Data sources. To investigate the broader phenotypic implications of the prioritized genes associated with CRP, we examined their links to 55 diseases^{1–6,57,58}, including 37 immune-mediated diseases (IMDs), 9 types of cancers, 4 cardiovascular diseases (CVDs), as well as non-insulin-dependent diabetes mellitus, chronic obstructive pulmonary disease (COPD), obesity, depression, and schizophrenia. Summary statistics for these diseases were available from the FinnGen Consortium's summary results (R11 release)⁵⁹. For colocalization analysis, two blood eQTL and three pQTL resources were used: GTEx V10 (*N* = 800 and 13,903 genes)⁶⁰, eQTLGen (*N* = 31,684 and 19,942 genes)⁶¹, UK Biobank pQTLs (*N* = 33,995 and 2911 proteins)⁶², deCODE pQTLs (*N* = 35,449 and 4719 proteins)⁶³, and Fenland pQTLs (*N* = 10,708 and 4979 proteins)⁶⁴.

Mendelian randomization and colocalization. For the selection of instrument variables (IVs), we retained SNPs identified based on the presence of significant cis-eQTLs or cis-pQTLs within a region of ± 1 MB. For eQTLs databases, significant eQTLs were defined as specified by the respective data source. In the case of pQTLs databases, the Benjamini-Hochberg procedure was applied to control the false discovery rate (FDR) at 5%. To minimize correlated horizontal pleiotropy, we retained SNPs independent of each other (LD window: 1MB, $R^2 < 0.01$ in 1000 G). To quantify the statistical power of the eQTLs (pQTLs), the strength of SNPs was evaluated by *F*-statistics, where an *F*-statistic thresholds ≥ 10 of IV indicates sufficient statistical strength⁶⁵. If any IVs had *F*-statistics <10, we considered those to have limited power (potentially causing weak instrument bias⁶⁶ and removed these from the MR). The presence of pleiotropy was further investigated using MR-PRESSO⁶⁷ and MR-Egger method⁶⁸ to estimate the potential effect of pleiotropy⁶⁹. For MR-PRESSO *P* value < 0.05 or MR-Egger intercept *P* value < 0.05, we

considered these gene (protein)-disease signals as influenced by horizontal pleiotropy. We also applied Cochran's *Q* test to estimate the potential heterogeneity of MR estimates (*P* value < 0.05). Genes (proteins) with horizontal pleiotropy and heterogeneity were excluded from any of the follow-up analyses. MR Steiger test of causality directionality was performed using 'directionality_test' function in TwoSampleMR package⁷⁰ to filter reverse causation. The Wald ratio (only one IV was available) and the inverse variance weighted (two or more IVs were available) method were used to estimate MR effects^{71,72}. A 5% FDR correction threshold was applied to correct for multiple testing. The coloc package⁷³ was employed for colocalization analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data are available in a public, open access repository. deCODE Genetics is available from: <https://www.decode.com>. Fenland cohort is available from: <https://www.omicscience.org/apps/pgwas/>. UK Biobank data is available from: <https://www.ukbiobank.ac.uk/>. All of Us data is available from: <https://allofus.nih.gov/>. FinnGen R11 release data is available from: <https://www.finnngen.fi/>. GTEx data is available from: <https://www.gtexportal.org/home/>. eQTLGen data is available from: <https://www.eqtngen.org/phase1.html>. The dbGaP accessions for TOPMed cohorts are as follows: Atherosclerosis Risk in Communities study (ARIC), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001211.v5.p4 and phs000280.v8.p2. Multi-Ethnic Study of Atherosclerosis (MESA), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001416.v4.p1 and phs000209.v13.p3. Framingham Heart Study (FHS), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000974.v6.p5 and phs000007.v34.p15. Cleveland Family Study (CFS), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000954.v4.p2 and phs000284.v2.p1. Hispanic Community Health Study – Study of Latinos (HCHS-SOL), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001395.v2.p1 and phs000810.v2.p2. The Source Data file is provided at <https://zenodo.org/records/15099550>.

Code availability

All software programs used in the study are publicly available and described in the Methods and Reporting Summary.

References

1. Bisioendial, R. J., Boekholdt, S. M., Vergeer, M., Stroes, E. S. & Kastelein, J. J. C-reactive protein is a mediator of cardiovascular disease. *Eur. Heart J.* **31**, 2087–2091 (2010).
2. Muller, D. C. et al. Circulating high sensitivity C reactive protein concentrations and risk of lung cancer: nested case-control study within Lung Cancer Cohort Consortium. *BMJ (Clin. Res. ed.)* **364**, k4981 (2019).
3. Zhu, M. et al. C-reactive protein and cancer risk: a pan-cancer study of prospective cohort and Mendelian randomization analysis. *BMC Med.* **20**, 301 (2022).
4. Pradhan, A. D., Manson, J. E., Rifai, N., Buring, J. E. & Ridker, P. M. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *Jama* **286**, 327–334 (2001).
5. Hu, G. et al. Association of serum C-reactive protein level with sex-specific type 2 diabetes risk: a prospective finnish study. *J. Clin. Endocrinol. Metab.* **94**, 2099–2105 (2009).
6. Yang, L. et al. Large-scale whole-exome sequencing analyses identified protein-coding variants associated with immune-mediated diseases in 350,770 adults. *Nat. Commun.* **15**, 5924 (2024).

7. Ligthart, S. et al. Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders. *Am. J. Hum. Genet.* **103**, 91–706 (2018).
8. Said, S. et al. Genetic analysis of over half a million people characterises C-reactive protein loci. *Nat. Commun.* **13**, 2198 (2022).
9. Dai, J. et al. Boosting the power of rare variant association studies by imputation using large-scale sequencing population. medRxiv:2023.10.28.23297722. (2023).
10. Wainschein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
11. Hernandez, R. D. et al. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* **51**, 1349–1355 (2019).
12. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
13. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
14. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
15. Shen, S. et al. A large-scale exome-wide association study identifies novel germline mutations in lung cancer. *Am. J. respiratory Crit. care Med.* **208**, 280–289 (2023).
16. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
17. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
18. Tobin M. D., Izquierdo A. G. Improving ethnic diversity in respiratory genomics research. *Eur. Respir. J.* **58**, 2101615 (2021).
19. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
20. Roh, K. et al. Lysosomal control of senescence and inflammation through cholesterol partitioning. *Nat. Metab.* **5**, 398–413 (2023).
21. Colton, C. A., Brown, C. M. & Vitek, M. P. Sex steroids, APOE genotype and the innate immune system. *Neurobiol. aging* **26**, 363–372 (2005).
22. Huang, Y. & Mahley, R. W. Apolipoprotein E: structure and function in lipid metabolism, neurobiology, and Alzheimer's diseases. *Neurobiol. Dis.* **72**, 3–12 (2014).
23. Weng, M. L. et al. Fasting inhibits aerobic glycolysis and proliferation in colorectal cancer via the Fdft1-mediated AKT/mTOR/HIF1 α pathway suppression. *Nat. Commun.* **11**, 1869 (2020).
24. Han, F. et al. Hypercholesterolemia risk-associated GPR146 is an orphan G-protein coupled receptor that regulates blood cholesterol levels in humans and mice. *Cell Res.* **30**, 363–365 (2020).
25. Lev, S. et al. Identification of a novel family of targets of PYK2 related to Drosophila retinal degeneration B (rdgB) protein. *Mol. Cell. Biol.* **19**, 2278–2288 (1999).
26. Lu, W. et al. The phosphatidylinositol-transfer protein Nir3 promotes PI(4,5)P₂ replenishment in response to TCR signaling during T cell development and survival. *Nat. Immunol.* **24**, 136–147 (2023).
27. Zhang, J. et al. Reactive oxygen species regulation by NCF1 governs ferroptosis susceptibility of Kupffer cells to MASH. *Cell Metab.* **36**, 1745–1763.e6 (2024).
28. McNally, K. E. et al. Retriever is a multiprotein complex for retromer-independent endosomal cargo recycling. *Nat. cell Biol.* **19**, 1214–1225 (2017).
29. Seaton, G., Smith, H., Brancale, A., Westwell, A. D. & Clarkson, R. Multifaceted roles for BCL3 in cancer: a proto-oncogene comes of age. *Mol. cancer* **23**, 7 (2024).
30. Zhang, Q. et al. The glucose-dependent insulinotropic polypeptide (GIP) regulates body weight and food intake via CNS-GIPR signaling. *Cell Metab.* **33**, 833–44.e5 (2021).
31. Marsolier, J. et al. H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nat. Genet.* **54**, 459–468 (2022).
32. Maeda, M. et al. Cancer cell niche factors secreted from cancer-associated fibroblast by loss of H3K27me3. *Gut* **69**, 243–251 (2020).
33. Gbate, N. B. et al. Phosphorylation and stabilization of EZH2 by DCAF1/VprBP trigger aberrant gene silencing in colon cancer. *Nat. Commun.* **14**, 2140 (2023).
34. Bodzioch, M. et al. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat. Genet.* **22**, 347–351 (1999).
35. Brooks-Wilson, A. et al. Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. *Nat. Genet.* **22**, 336–345 (1999).
36. Rust, S. et al. Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. *Nat. Genet.* **22**, 352–355 (1999).
37. Jackson, R. J., Hyman, B. T. & Serrano-Pozo, A. Multifaceted roles of APOE in Alzheimer disease. *Nat. Rev. Neurol.* **20**, 457–474 (2024).
38. Garshick, M. S. et al. An inflammatory transcriptomic signature in psoriasis associates with future cardiovascular events. *J. Eur. Acad. Dermatol. Venereology: JEADV* **37**, 1361–1365 (2023).
39. Keightley, P. D. & Jackson, B. C. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics* **209**, 897–906 (2018).
40. Zhou, H. et al. Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals. *Nat. Med.* **29**, 3184–3192 (2023).
41. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
42. Eggertsson, H. P. et al. GraphTyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv.1303. (2013).
44. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
45. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
46. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
47. Li S., Carss K. J., Halldorsson B. V., Cortes A. Whole-genome sequencing of half-a-million UK Biobank participants. medRxiv:2023.12.06.23299426. (2023).
48. Pulit, S. L., de With, S. A. & de Bakker, P. I. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol.* **41**, 145–151 (2017).
49. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
50. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7**, 177–188 (1986).
51. Shrine et al. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat. Genet* **55**, 410–422 (2023).
52. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic acids Res.* **52**, D1143–d54 (2024).

53. Yuan, K. et al. Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nat. Genet.* (2024).
54. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
55. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* **51**, 343–353 (2019).
56. Zheng, J. et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).
57. Pletnikoff, P. P. et al. Cardiorespiratory fitness, C-reactive protein and lung cancer risk: A prospective population-based cohort study. *Eur. J. Cancer (Oxf., Engl.: 1990)* **51**, 1365–1370 (2015).
58. Gunter, M. J. et al. Circulating Adipokines and Inflammatory Markers and Postmenopausal Breast Cancer Risk. *J. Natl. Cancer Inst.* **107**, 10.1093/jnci/djv169 (2015).
59. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
60. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369:1318–1330. (2020).
61. Vösa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
62. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
63. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).
64. Pietzner, M. et al. Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
65. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
66. Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat. Med.* **30**, 1312–1323 (2011).
67. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of wide-spread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
68. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
69. Bowden, J. et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I² statistic. *Int. J. Epidemiol.* **45**, 1961–1974 (2016).
70. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, 7. (2018).
71. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
72. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Stat. methods Med. Res.* **26**, 2333–2355 (2017).
73. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

Acknowledgements

We want to acknowledge the participants and investigators of deCODE Genetics, Fenland cohort, UK Biobank, All of Us, ARIC, MESA, FHS, CFS, HCHS-SOL and FinnGen. This work was supported by the National Natural Science Foundation of China (82220108002 to F.C., 82373685 and 82103946 to S.S., 82373690 and 82173620 to Y.Z.), and US NIH (NCI) grant #U01CA209414 to DCC, Open Research Fund Program of Changzhou Institute for Advanced Study of Public Health, Nanjing Medical University (No. CPHM202301 to S.S.).

Author contributions

S.S. and F.C. contributed to the study design. S.S. and H.L. contributed to data collection. H.L. and J.Z. performed statistical analyses and interpretations. H.L. and J.Z. drafted the manuscript. J.D., D.Y., Y.Z., and D.C. revised the final manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59155-w>.

Correspondence and requests for materials should be addressed to Feng Chen or Sipeng Shen.

Peer review information *Nature Communications* thanks Marina Ciullo, who co-reviewed with Daniela Ruggiero, Symen Ligthart and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025