nature communications

Article

https://doi.org/10.1038/s41467-025-62463-w

European and African ancestry-specific plasma protein-QTL and metabolite-QTL analyses identify ancestry-specific T2D effector proteins and metabolites

Received: 22 March 2024

Accepted: 19 July 2025

Published online: 11 August 2025

Check for updates

Chengran Yang[1,2], Priyanka Gorijala[1,2], Jigyasha Timsina[1,2], Lihua Wang[1,2], Menghan Liu[1,2], Ciyang Wang[1,2], William Brock[1,2], Yueyao Wang[1,2], Fumihiko Urano[3,4], Yun Ju Sung[1,2,5] & Carlos Cruchaga[1,2,6,7]

In this study, we generated and integrated plasma proteomics and metabolomics with the genotype datasets of over 2300 European (EUR) and 400 African (AFR) ancestries to identify ancestry-specific multi-omics quantitative trait loci (QTLs). In total, we mapped 954 AFR pQTLs, 2848 EUR pQTLs, 65 AFR mQTLs, and 490 EUR mQTLs. We further applied these QTLs to ancestry-stratified type-2 diabetes (T2D) risk to pinpoint key proteins and metabolites underlying the disease-associated genetic loci. Using INTACT that combined trait-imputation and colocalization results, we nominated 270 proteins and 72 metabolites from the EUR set; seven proteins and one metabolite from the AFR set as molecular effectors of T2D risk in an ancestry-stratified manner. Here, we show that the integration of genetic and omic studies of different ancestries can be used to identify distinct effector molecular traits underlying the same disease across diverse ancestral groups.

Until very recently, human genetics studies have mainly focused on participants of European ancestry[1]. However, there has been a recent increase in studies that include multiple ancestral backgrounds[2–4]. With these efforts, human geneticists have now published numerous genome-wide wide association studies (GWAS) on complex diseases that encompass multiple ancestral groups or specifically target non-European genetic ancestries. For example, three of the largest studies on type-2 diabetes (T2D) included participants of five different ancestries[5–7]: Europeans (EUR), Africans (AFR), Hispanics, East Asians, and South Asians, leading to the identification of tens of hundreds of loci. Mahajan et al.[5], Vujkovic et al.[6], and Suzuki et al.[7], reported 338, 568, and 1289 genetic loci associated with multi-ancestry T2D respectively. These studies reported 11 to 145 loci were novel with sample size from 1.3 million to 2.5 million participants (21–49% non-EUR). These T2D genetic studies also found variants were enriched in T2D-relevant tissues including pancreatic islets and adipose tissue. However, T2D GWAS alone are not enough to identify the effector genes (an "effector" is defined as a molecular trait that is associated with a genetic variant and in fact plays a role in the disease) underlying those associations.

Post-GWAS approaches, including colocalization[8,9], or trait-imputation (such as Transcriptome-wide association study (TWAS) for RNA expression[10,11]), can prioritize variants to genes and later identify the pathways implicated in diseases. In fact, studies that

[1]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. [2]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, MO, USA. [3]Division of Endocrinology, Metabolism, and Lipid Research, Washington University School of Medicine, MSC 8127-0021-09, St. Louis, USA. [4]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, USA. [5]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA. [6]Hope Center for Neurological Disorders, Washington University School of Medicine, St. Louis, MO, USA. [7]Charles F. and Joanne Knight Alzheimer Disease Research Center, Washington University School of Medicine, St. Louis, MO, USA.
e-mail: cruchagac@wustl.edu

integrate molecular phenotypes (such as gene expression of the RNA levels) have been instrumental to facilitate the interpretation of the T2D GWAS-associated loci. Within the two T2D risk GWA studies by Mahajan et al.[5], and Vujkovic et al.[6], the authors integrated the GWAS results with expression QTLs (eQTLs), promoter-focused chromatin confirmation capture (pcHi-C) links, via TWAS, variant annotation, and genetic colocalization to nominate the potential effectors for the identified loci. However, few studies have used protein- or metabolite-QTLs to nominate effectors. As proteins and metabolites are closer to the disease endpoint other than RNA, performing such genetic analysis with protein and metabolites levels will be useful to study T2D. In addition, to date, most T2D post-GWAS studies have only used molecular phenotype datasets generated in participants of the EUR ancestry, and thus lacking the integration of ancestry-matched analyses.

A recent preprint[12] tried to fill this gap by integrating four EUR based pQTL datasets and two EUR based mQTL datasets to identify effector proteins and metabolites through genetic colocalization. The authors identified 1728 unique effector proteins and 731 effector metabolites highly colocalized with the T2D risk. This preprint, however, did not use ancestry-matched proteomic and metabolomic datasets on the multi-ancestry T2D risk GWAS and the preprint only used colocalization analysis with a strict assumption of the single variant to identify T2D effectors. Thus, additional genetic studies to study T2D risk with multi-omics in an ancestry-matched manner are needed. Moreover, using additional bioinformatic tools to integrate the QTL and GWAS datasets and considering multiple variants underlying the same genetic loci when performing colocalization will facilitate the discovery of more T2D effectors.

Here, we aim to identify ancestry-specific T2D effector proteins and metabolites by integrating the ancestry-matched deep molecular phenotyping datasets with the ancestry-specific T2D GWAS (Fig. 1). We first identify the ancestry-specific genetic associations in both the plasma proteome and metabolome. Ancestral group stratification was performed using principal component analysis (PCA) with the reference by 1000 Genome project[13] (See "Methods" for details). Next, we apply them to pinpoint key proteins/metabolites underlying the risk loci of ancestry-matched T2D. Our study used plasma omics to study T2D because plasma analytes are more relevant to metabolic disorders, such as T2D, compared to diseases that are enriched in certain tissue types, such as the brain for neurological disorders. In our study, we also consider trans associations in all our molecular trait QTL mapping sets. By incorporating trans associations, we anticipate uncovering more findings compared to those previous cis-centric studies.

Results

Genetic architecture of the plasma proteome in participants of African and European ancestry

To build ancestry-stratified genetic maps of the plasma proteome and metabolome, we performed pQTL and metabolite-QTL (mQTL) analyses on participants of African and European ancestry separately (Fig. 1; and Supplementary Fig. 1, Supplementary Data 1–7). We utilized an aptamer-based assay (SomaScan 7k platform[14]) to measure protein levels and a mass-spectrometry assay (Metabolon HD4 platform[15]) to quantify the metabolite levels. The proteomic and metabolomic datasets were generated from the same cohort. Following quality control procedures for the omics data and integration with array-based post-imputation genotype data, we constructed four maps: i) AFR pQTL (414 participants and 6907 proteins), ii) EUR pQTL (2338 participants and 6907 proteins), iii) AFR mQTL (417 participants and 1413 metabolites), and iv) EUR mQTL (2392 participants and 1483 metabolites). To determine the study-wide significant QTLs, we derived it from genome-wide significance after further accounting for the number of independent features accounting for 95% of the variance of each dataset within each

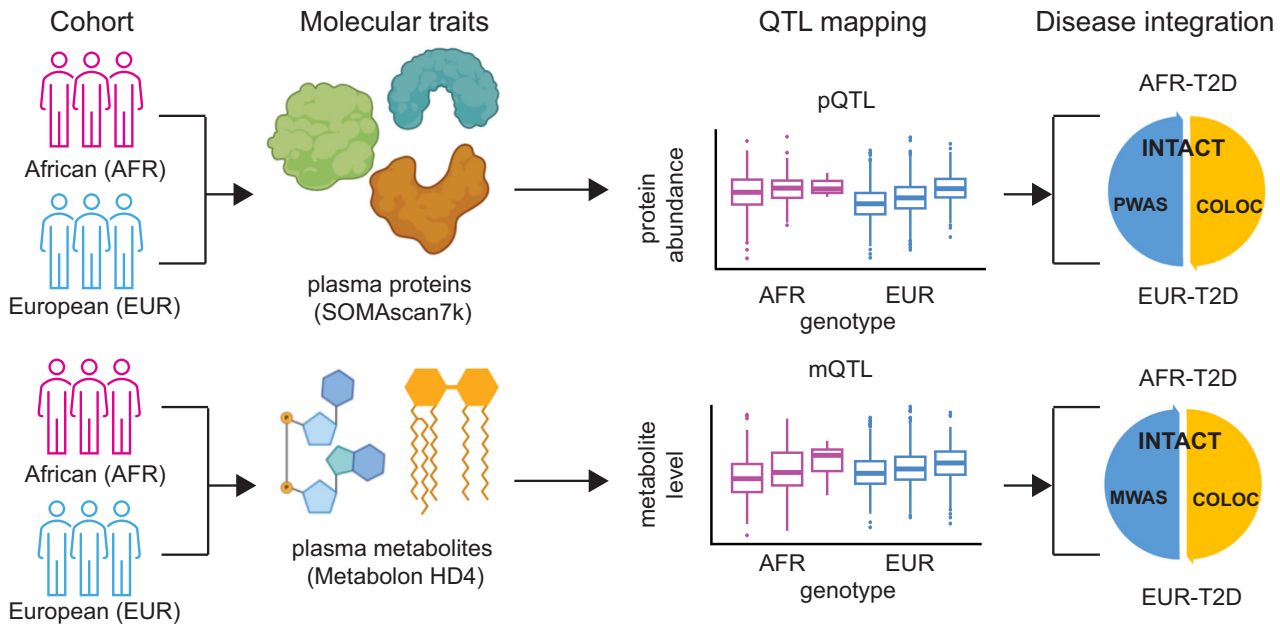


Fig. 1 | Schematics of the study design on genetic architecture of the plasma proteome and metabolome in participants with African and European ancestry. Top: The plasma proteins in participants of African (AFR, in magenta) and European (EUR, in blue) ancestries were profiled together with the SomaScan 7k platform. Integrating the abundance of each protein with the array-based genotype data, we identified protein quantitative trait loci (pQTLs) in both ancestries. We further used these pQTLs to prioritize proteins in the ancestry-matched type-2 diabetes (T2D) risk via INtegration of Transcriptome-wide association study And ColocalizaTion (INTACT), with proteome-wide association study (PWAS) and colocalization (COLOC) as input. Bottom: The plasma metabolites in participants of African and European ancestries were profiled together with the Metabolon HD4 platform. Integrating the level of each metabolite with the array-based genotype data, we identified metabolite quantitative trait loci (mQTLs) in both ancestries. We further used these mQTLs to prioritize metabolites in the ancestry-matched T2D risk via INTACT, with metabolome-wide association study (MWAS) and COLOC as input. Schematics were created with icon library from the Microsoft PowerPoint and BioRender (https://BioRender.com/ephif2e) in Affinity Designer.

https://doi.org/10.1038/s41467-025-62463-w

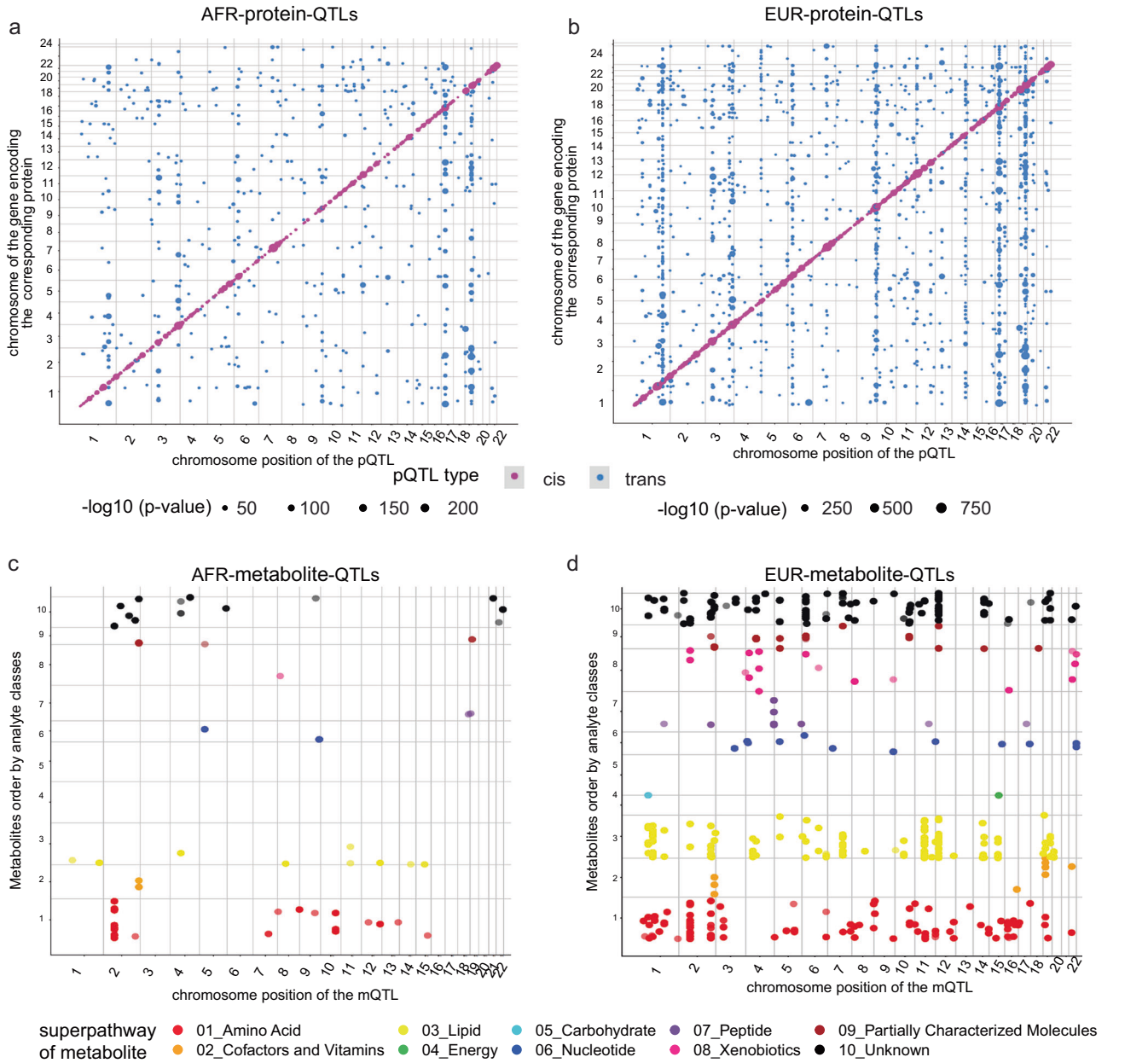


Fig. 2 | Four genetic maps of the plasma proteome and metabolome in participants with African and European ancestry. a Map of 954 AFR pQTLs. The X-axis is the chromosome position of the pQTL; the Y-axis is the chromosome of the gene encoding the corresponding protein. The color code is magenta as cis-pQTLs and blue as trans-pQTLs. The point size represents the negative-log10-*p* value, the scales are from 0 to 200. *P* values are unadjusted, two-sided, and determined via linear regression analysis. The *p* value threshold is after multiple testing correction. The value threshold for cis-pQTL is 5×10^{-8} and for trans-pQTL is 1.49×10^{-10} (5×10^{-8}/336). **b** Map of 2848 EUR pQTLs. Same X, Y, and color code as panel-A. The point size represents the negative-log10-*p* value, the scales are from 0 to 750. The *p* value threshold for cis-pQTL is 5×10^{-8} and for trans-pQTL is 3.40×10^{-11} (5×10^{-8}/1472). **c** Map of 65 AFR mQTLs. The X-axis is the chromosome position of the mQTL; the Y-axis is the metabolite order (based on the super pathway). The color code is red as 01_Amino Acid; orange as 02_Cofactors and Vitamins; yellow as 03_Lipid; limegreen as 04_Energy; cyan as 05_Carbohydrate; blue as 06_Nuleotide; purple as 07_Peptide; deepplink1 as 08_Xenobiotics; brown as 09_Partially Characterized Molecules; black as 10_Unknown categories. *P* values are unadjusted, two-sided, and determined via linear regression analysis. The *p* value threshold is after multiple testing correction, and it is 1.78×10^{-10} (5×10^{-8}/281). **d** Map of 490 EUR mQTLs. Same X, Y, and color code as panel-C. The *p* value threshold is after multiple testing correction and it is 6.53×10^{-11} (5×10^{-8}/766).

separate map (see "Methods" for more details). Briefly, the *p* value threshold for cis-pQTLs was 5×10^{-8}; for trans-pQTLs were 3.40×10^{-11} (EUR) and 1.49×10^{-10} (AFR), respectively. Unlike proteins, metabolites do not correspond to a specific gene, so it is not possible to define *cis* or *trans* mQTL for them, the *P* value thresholds for mQTLs were 6.53×10^{-11} (EUR) and 1.78×10^{-10} (AFR).

To identify genetic variants associated with the plasma proteome in individuals of African ancestry, we conducted pQTL mapping on African ancestry participants (Fig. 2a). Of 6907 proteins that passed QC, we identified 881 proteins with 954 study-wide significant pQTLs (Supplementary Data 8). Among these findings, 420 pQTLs were classified as *cis*, while 534 were *trans*-pQTLs. Consistent with previous studies[16–18], we observed that the absolute effect size was negatively correlated with the minor allele frequency (Supplementary Fig. 2a). After assigning each pQTL to its ancestry-matched linkage disequilibrium (LD) block[19], we identified a total of 548

unique genetic loci, including 154 pleiotropic regions. Notably, we found 33 proteins associated with the *APOE* locus (Supplementary Fig. 3a), which ranked second in terms of AFR proteomic-associated pleiotropic regions and genomic hotspots. The other top-five pleiotropic loci included *VTN* (chr17q11.2) with 35 proteins, *ABO* (chr9q34.2) with 24 proteins, *MHC* region with 21 proteins, and the *CFH* (chr1q31.3) locus with 16 proteins. In the European ancestry-stratified analyses, of the 6907 proteins, 2400 proteins showed 2848 significant pQTLs; 1282 *cis*-pQTLs and 1566 *trans*-pQTLs (Fig. 2b; and Supplementary Data 9; Supplementary Fig. 2b). Of the top five pleiotropic pQTL loci in EUR (totally 746 regions), the *APOE* locus (chr19q13.32) was associated with 126 proteins (Supplementary Fig. 3b). The other top-five pleiotropic loci included the *VTN* (chr17q11.2) with 182 proteins, *CFH* (chr1q31.3) with 151 proteins, *MHC* region with 86 proteins, and the *ABO* locus (chr9q34.2) with 82 proteins. As expected, the strength of associations of cis-pQTLs was indeed stronger in both AFR (Wilcoxon p value = 0.0145 < 0.05) and EUR (Wilcoxon p value = $4.02 \times 10^{-9} < 0.05$) sets.

To determine how many of the pQTLs have been reported before, we conducted a comparison between our study-wide pQTLs and the three largest to-date external studies covering both cis and trans associations while encompassing two genetic ancestries (see "Methods"). Of these four datasets from three external pQTL studies (Supplementary Data 10), Ferkingstad et al.[20], and Sun et al.[17], included participants of EUR ancestry, while Surapaneni et al.[16], and Sun et al.[17], sampled individuals of AFR ancestry. Overall, out of the 954 AFR pQTLs (Supplementary Data 11) identified, we found that 561 had been previously reported with a study-wide significant p value threshold of 5×10^{-11}. Additionally, among the remaining 393 AFR pQTLs, 14 had been reported with a genome-wide threshold, 45 had passed a nominal ($p < 0.05$) threshold, and 242 did not show nominal significance in previous studies. Among the pQTLs that were not tested, 82 were due to missing proxy variants, and 10 were due to missing protein data. Considering the largest number of proteins (~5k) profiled from a large-scale European cohort in the study by Ferkingstad et al.[20], we can replicate the highest number of our findings for the EUR pQTLs, with a p value below 5×10^{-2} in this external study. Of the 2848 EUR pQTLs identified (Supplementary Data 11), 2052 had been reported as study-wide significant ($p < 5 \times 10^{-11}$), 43 were below a genome-wide threshold, 241 were below a nominal threshold ($p < 0.05$), while 395 were above the nominal threshold. This indicates that 86% of the tested pQTLs have supportive evidence from previous studies. Among the untested pQTLs, 81 were due to missing proxy variants and 36 were due to the missing protein data. Either in AFR or in EUR, we found that cis associations tend to be more likely replicated than trans (Supplementary Fig. 4a, AFR: The cis and trans replication rate was 95% and 41%; Supplementary Fig. 4b, EUR: The cis and trans replication rate was 90% and 65%, respectively).

Genetic architecture of the plasma metabolome in participants of African and European ancestry

To detect genetic variants associated with the plasma metabolome in African and European ancestry respectively, we performed mQTL mapping using the same participants from which the proteomic data was generated (Fig. 2c–d). After quality controlling for both genotype and metabolome datasets, we identified 65 significant mQTLs in 34 genetic loci, associated with 60 metabolites (out of a total of 1413 metabolites tested) in the African ancestry cohort (Fig. 2c, and Supplementary Data 12). Notably, a significant number of these hits (27 out of 34 loci) were involved in and enriched for the super pathway of amino acids (enrichment ratio = 2.78, Fisher exact test, p value = 6.63×10^{-5}) compared to other pathways. In the European cohort, we found 490 significant mQTLs in 124 genetic regions associated with 403 metabolites (Fig. 2d, and Supplementary Data 13). The majority of the hits were observed in two super pathways: lipids (198 metabolites, enrichment ratio = 1.26, Fisher exact test, p value = 0.019) and amino acids (106 metabolites, enrichment ratio = 1.52, Fisher exact test, p value = 0.0015). In line with a previous cross-platform mQTL study[21], we observed both sets of mQTL followed a trend where the absolute effect size negatively correlated with the minor allele frequency across all variants (Supplementary Fig. 2c–d). Additionally, we identified top-ranked LD blocks associated with metabolites were loci near *ALMS1P1* (chr2p13.1), *UGT1A6* (chr2q37.1), and *PYROXD2* (chr10q24.2) in the African cohort (Supplementary Fig. 3c), and *FADS1/2* (chr11q12.2), *SLCO1B1* (chr12p12.1), *ALMS1P1* (chr2p13.1) in the European cohort (Supplementary Fig. 3d). These regions represent metabolite-associated pleiotropic regions and genomic hotspots specific to their respective ancestral groups.

To assess the presence of previously reported mQTLs, we examined our study-wide mQTLs using three of the up-to-date external studies with full summary statistics available, which included participants from both genetic ancestries (details see "Methods"). Among these three external mQTL studies (Supplementary Data 14), Yin et al.[22], and Chen et al.[23], focused on individuals of EUR ancestry alone, while Rhee et al.[24], included participants of AFR ancestry (Supplementary Data 15). Of the 65 AFR mQTLs identified (Supplementary Data 16), we found that 48 had already been reported after multiple testing corrections with a study-wide significance threshold of 5×10^{-11}. One passed a nominal threshold, seven were above the nominal threshold, and nine were not examined as the metabolites were missing. On the other hand, of the 490 EUR mQTLs identified (Supplementary Data 16), we found 412 passed the study-wide threshold of p value as 5×10^{-11}, seven passed a genome-wide threshold, ten passed a nominal threshold, while nine did not reach the nominal threshold. Among the mQTLs that were not tested, two were due to the missing proxy variants and 50 were due to the missing metabolite data. These results indicate that approximately 88% of tested mQTL pairs in the African ancestry participants and 98% of tested mQTL pairs in the European ancestry participants have supportive evidence from previous studies.

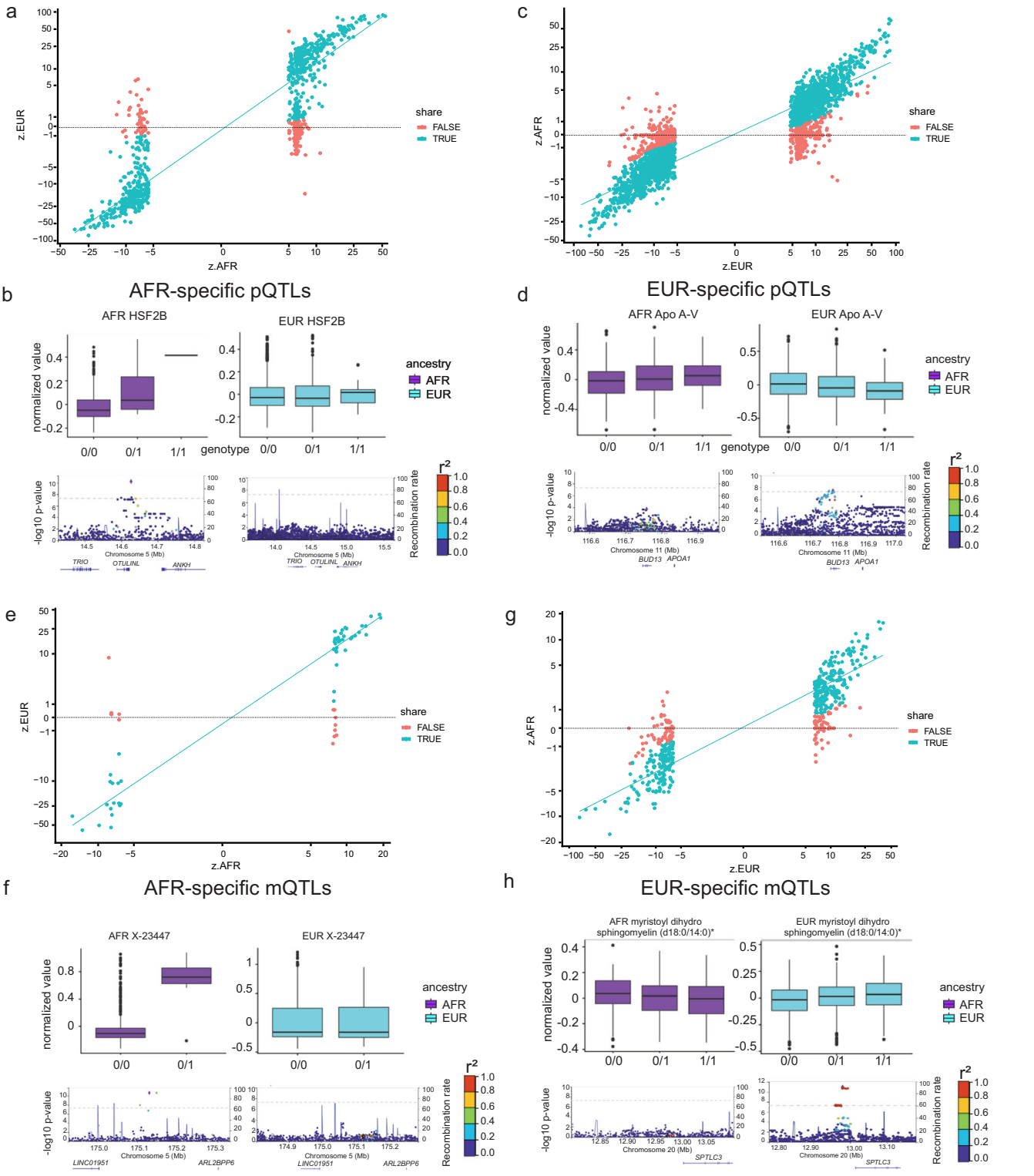
All four QTL datasets can be interactively explored on the PheWeb[25]-based ONTIME browser (https://ontime.wustl.edu/).

Ancestry-specific pQTLs and mQTLs

To identify ancestry-specific xQTLs (i.e., pQTL and mQTL), we compared results between participants of African and European ancestry. Briefly, ancestry-specific hits were determined based on fold-change criteria and considering both the effect size and standard error (for deriving the Z-normalized effect size), following the methodology used in previous condition-dependent genetic studies[26,27]. The fold-change greater than ten-fold or smaller than 10%-fold was used as the threshold, with log10-based fold change boundaries of +/- 1 to determine context-specific QTLs (see "Methods").

In the case of proteomics, of the 954 pQTLs identified in AFR participants, 29.6% were considered AFR-specific pQTLs (Fig. 3a, and Supplementary Data 17). For example, in African ancestry participants, the protein levels of HSF2B (Heat shock factor 2-binding protein) were positively associated with the genetic variant chr5:14626365:T:C ($Z = 6.77$, $MAF.afr = 0.05$), However, in European ancestry participants, the HSF2B protein levels were similar across all genotypes of the same variant ($Z = -0.007$, $MAF.eur = 0.10$), resulting in a large fold-change difference ($Z.afr/Z.eur\ ratio = -949$, Fig. 3b). Similarly, among the 2,848 pQTL identified in EUR participants, 24.3% were considered EUR-specific pQTLs (Fig. 3c, and Supplementary Data 18). In the European-ancestry cohort, the protein levels of Apo A-V (Apolipoprotein A-V) were significantly decreased with the minor allele dosages of the variant chr11:116780399:C:T ($Z = -5.55$, $MAF.eur = 0.20$) but this association was not observed in African ancestry participants ($Z = 2.51$, $MAF.afr = 0.26$), leading to a fold-

https://doi.org/10.1038/s41467-025-62463-w



change ratio ($Z.afr/Z.eur$) of -0.452 (Fig. 3d). Among the ancestry-specific pQTLs, 38 of 282 AFR and 268 of 692 EUR pQTLs were cis-pQTLs.

For the metabolomics analyses, of the 65 mQTLs identified in AFR participants, 20% were classified as AFR-specific mQTLs (Fig. 3e, and Supplementary Data 19). The metabolite abundances of the X-23447 were increased with the minor allele dosage of the genetic variant chr5:175129766:C:T ($Z = 6.89$, $MAF.afr = 0.0096$) in AFR participants, but in EUR participants, this metabolite displayed similar levels across all genotypes of the variant ($Z = -0.0116$, $MAF.eur = 0.014$), resulting in a substantial fold-change difference (*ratio of* $Z.afr/Z.eur = -594$, Fig. 3f). Likewise, among the 490 mQTLs identified with EUR participants, 23.7% were considered EUR-specific mQTLs (Fig. 3g, Supplementary Data 20). In the EUR cohort, the levels of myristoyl dihydrosphingomyelin (d18:0/14:0) were significantly and positively associated with the minor allele dosages of the variant chr20:12978750:T:C ($Z = 6.84$, $MAF.eur = 0.38$), whereas in the AFR group, this association was not observed ($Z = -1.98$, $MAF.afr = 0.41$), leading to a fold-change ratio ($Z.afr/Z.eur$) of -0.289 (Fig. 3h).

Fig. 3 | Ancestry-specific pQTL and mQTL hits. a AFR-specific pQTL vs queried EUR. The correlation coefficients (Pearson's r) of the standardized effect sizes in the pseudo-log10-scale between the two ancestries are 0.923 and -0.147 for the shared (green) and AFR-specific (red) pQTLs. The lines are based on a linear regression model. **b** One AFR-specific pQTL example as HSF2B visualized by boxplots (AFR in purple: 0/0 $n = 325$, 0/1 $n = 36$, 1/1 $n = 1$; EUR in cyan: 0/0 $n = 1727$, 0/1 $n = 384$, 1/1 $n = 19$, all biologically independent samples) and locus zoom plots (P values are unadjusted, two-sided, and determined via linear regression analysis). **c** EUR-specific pQTL vs queried AFR. The correlation coefficients (Pearson's r) of the standardized effect sizes in the pseudo-log10-scale between the two ancestries are 0.935 and 0.234 for the shared (green) and EUR-specific (red) pQTLs. The lines are based on a linear regression model. **d** One EUR-specific pQTL example as Apo A-V visualized by boxplots (AFR in purple: 0/0 $n = 220$, 0/1 $n = 156$, 1/1 $n = 28$; EUR in cyan: 0/0 n = 1441, 0/1 $n = 764$, 1/1 $n = 98$) and locus zoom plots (*P* values are unadjusted, two-sided, and determined via linear regression analysis). **e**) AFR-specific mQTL vs queried EUR. The correlation coefficients (Pearson's r) of the standardized effect sizes in the pseudo-log10-scale between the two ancestries are 0.927 and -0.51 for the shared (green) and AFR-specific (red) mQTLs. The lines are based on a linear regression model. **f** One AFR-specific mQTL example as X-23447 visualized by boxplots (AFR in purple: 0/0 n = 404, 0/1 $n = 8$; EUR in cyan: 0/0 $n = 2243$, 0/1 $n = 63$, all biologically independent samples) and locus zoom plots (*P* values are unadjusted, two-sided, and determined via linear regression analysis). **g** EUR-specific mQTL vs queried AFR. The correlation coefficients (Pearson's r) of the standardized effect sizes in the pseudo-log10-scale between the two ancestries are 0.927 and -0.51 for the shared (green) and AFR-specific (red) mQTLs. The lines are based on a linear regression model. **h** One EUR-specific mQTL example as myristoyl dihydrosphingomyelin (d18:0/14:0) visualized by boxplots (AFR in purple: 0/0 $n = 70$, 0/1 $n = 192$, 1/1 $n = 143$; EUR in cyan: 0/0 $n = 906$, 0/1 $n = 1105$, 1/1 $n = 349$, all biologically independent samples) and locus zoom plots (*P* values are unadjusted, two-sided, and determined via linear regression analysis). For all boxplots, the box represents the interquartile range, 25th percentile (lower quartile) and 75th percentile (upper quartile), the line within the box represents the median, and the whiskers extend to the minimum and maximum. For locus zoom plots, the SNPs for each regional plot are denoted as a purple diamond. Each dot represents individual SNPs, and dot colors in the regional plots represent R-squared value on linkage disequilibrium (LD) with the named SNP at the center. Blue vertical lines in the regional plots show recombination rates as marked on the right of the Y axis.

To further characterize the ancestry-specific QTLs, we categorized the genetic variants into three bins based on minor allele frequency (MAF). The bins were defined as follows: bin-1, MAF ranging from 0 to 0.01; bin-2, MAF ranging from 0.01 to 0.05; and bin-3, MAF ranging from 0.05 to 0.5. The MAF threshold for each variant was determined by considering the minimum MAF value between the two ancestries. We found that the larger the MAF of a genetic variant, the less likely it was to be specific to a particular ancestry (Supplementary Fig. 5a–g). On average, the proportion of ancestry-specific QTLs decreased from 68.5% to 45% as the MAF bins shifted from bin-1 to bin-2. Moreover, the ancestry-specific QTLs decreased to 10.8% at bin-3. Power analyses (see "Methods") indicated that the sentinel xQTLs used in this ancestry specificity section were well-powered given the current sample size. Even though the variants with lower MAF tend to have a lower power with the same effect size, all the ancestry-specific variants showed >80% power in the other ancestry (Supplementary Data 21–22). After examining the MAF difference between EUR and AFR for those ancestry-specific findings, the ancestry-specific ones had smaller MAF differences than ancestry-shared or non-specific ones (Supplementary Fig. 6a–d). This observation is plausible, as we found that cis pQTLs in the EUR and AFR ancestry-specific findings had significantly different MAFs between the ancestries (Supplementary Fig. 6e–f): for example, in AFR-specific findings, the cis pQTLs from the AFR set had a larger MAF than the EUR set (Supplementary Fig. 6e, 1st box plot); and vice versa. For the EUR-specific findings, the cis pQTL from the EUR set had a larger MAF than the AFR set (Supplementary Fig. 6f, 1st box plot). On the other hand, when checking the shared pQTLs, the cis pQTLs did not always have significantly different MAFs (i.e. the AFR pQTL findings with p = 0.22 in the 3rd box plot, Supplementary Fig. 6e). This principle cannot be generalized to the trans pQTLs (Supplementary Fig. 6e–f) and mQTLs (Supplementary Fig. 6g–h), as the differences in MAF across QTLs were not as simple as the cis pQTLs given the variant and the molecular phenotype may be distally connected.

As a complementary strategy, we employed a more flexible Bayesian approach, multivariate adaptive shrinkage (MASH) framework[26], to calculate the posterior probability and posterior mean for each QTL-trait pair in the two ancestries. The posterior mean fold change also indicated a similar ancestry-sharing proportion ranging from 82.3% to 96.2% (Supplementary Fig. 7a–d), which supports the previous estimations of ancestry-specific QTLs. These results align with previous studies (Zhang et al.[28], for proteomics and Rhee et al.[24], for metabolomics datasets). Zhang et al.[28], reported 10% EUR-specific and 30% AFR-specific cis-pQTLs, while Rhee et al.[24], uncovered 22% ancestry-specific mQTLs. Moreover, our findings extend these previous reports by analyzing different MAF bins, revealing that the ancestry-specific QTLs are more likely to have lower frequencies. This observation could be explained by the fact that functional variants tend to have lower frequencies than non-functional variants, and therefore, these ancestry-specific QTLs may capture those functional variants. Alternatively, participants of African ancestry may have a higher prevalence of rare variants compared to those of European ancestry, which increases the likelihood of finding ancestry-specific associations.

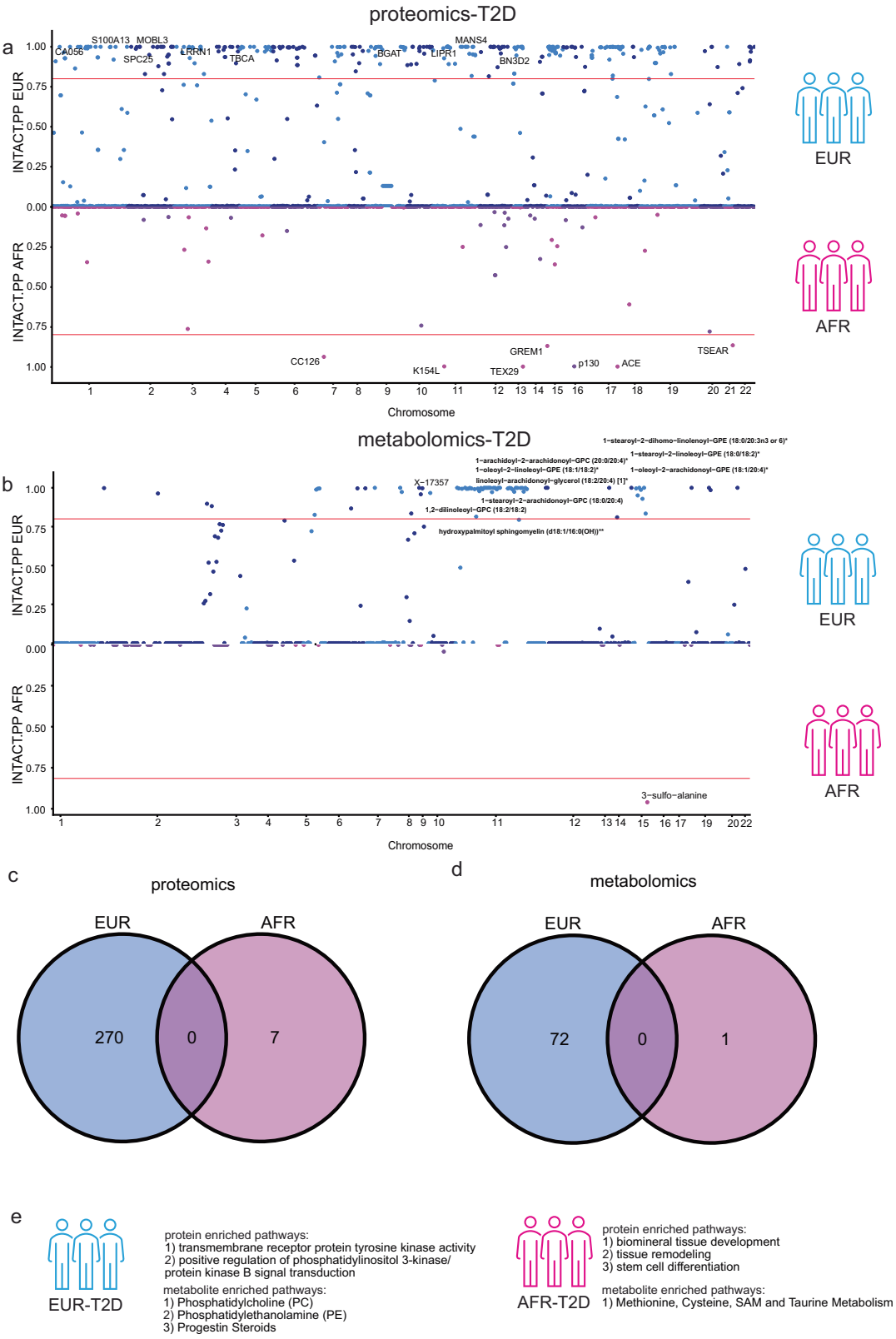
Identification of the T2D effector proteins and metabolites in an ancestry-matched manner via INTACT

Finally, to identify proteins associated with T2D in an ancestry-stratified manner, we turned to the framework of INTACT (INtegration of TWAS And ColocalizaTion)[29], as this framework jointly used the output from the single computational approach and yielded robust and powerful inference results[29]. The framework takes the input of proteome-wide association study (PWAS) z-statistics and locus colocalization posterior probabilities and calculates a posterior probability of the causality for each protein. Specifically, we prioritized proteins that were associated with ancestry-matched T2D risk within each ancestry group, namely EUR- and AFR-stratified analyses (Fig. 4a–b, and Supplementary Fig. 8–9, Supplementary Data 23–24). For the EUR T2D risk analysis, we used the summary statistics from Mahajan et al.[5], that included 80,154 cases and 853,816 controls. On the other hand, to investigate AFR T2D risk, we leveraged the study conducted by Vujkovic et al.[6], which enrolled 24,646 cases and 31,446 controls. These two GWAS reported 225 and 23 genome-wide significant hits in EUR and AFR populations, respectively.

From the ancestry-stratified INTACT analyses, we found 270 proteins in EUR and seven in AFR associated with T2D given a posterior probability greater than 0.8 (Fig. 4a). In the European ancestry group, the top 10 associated proteins included CA056, S100A13, SPC25, MOBL3, LRRN1, TBCA, BGAT, LIPR1, MANS4, and BN3D2 (Fig. 4a, top). For the African ancestry group, the associated seven proteins were CC126, K154L, TEX29, GREM1, p130, ACE, and TSEAR (Fig. 4a, bottom). No proteins were found in common in the EUR and AFR-specific analyses (Fig. 4c).

We applied the same INTACT framework to assess the association between metabolite levels and T2D in an ancestry-specific manner (Fig. 4b, and Supplementary Data 25–26). We identified 72 EUR and one AFR metabolites with a posterior probability greater than 0.8 (Fig. 4b). Of the 72 nominated metabolites from EUR-stratified analysis (Fig. 4b, top), the top-10 metabolites were 1-stearoyl-2-dihomo-linolenoyl-GPE (18:0/20:3n3 or 6), X-17357, 1-oleoyl-2-linoleoyl-GPE (18:1/

https://doi.org/10.1038/s41467-025-62463-w



18:2), 1-stearoyl-2-linoleoyl-GPE (18:0/18:2), 1-arachidoyl-2-arachidonoyl-GPC (20:0/20:4), 1-stearoyl-2-arachidonoyl-GPC (18:0/20:4), linoleoyl-arachidonoyl-glycerol (18:2/20:4), 1-oleoyl-2-arachidonoyl-GPE (18:1/20:4), 1,2-dilinoleoyl-GPC (18:2/18:2), hydroxypalmitoyl sphingomyelin (d18:1/16:0(OH)). Only one metabolite, 3−sulfo−alanine, was identified from the AFR-stratified analysis (Fig. 4b, bottom). No metabolite was in common between EUR and AFR groups (Fig. 4d).

Implications of proteins and metabolites underlying the risk of ancestry-matched T2D

To investigate whether the proteins and metabolites implicated here in T2D were associated with any previously reported T2D effector gene, we cross-referenced the loci associated with molecular features were reported based on variant annotation, genetic colocalization with eQTLs, pcHi-C links, and TWAS significance[5,6]. We found two of seven

Fig. 4 | Integration of proteins and metabolites with the ancestry-matched risk of type-2 diabetes. a Miami plot highlighting the INTACT nominated proteins associated with T2D between two ancestries (AFR in magenta and EUR in blue). Due to the size limit, we only show the top-10 EUR proteins (top). We show all seven AFR proteins (bottom). The Y-axis is the INTACT posterior probability, with a threshold of 0.8 (The horizontal red line represents the threshold). The X-axis is the protein associated genetic variant location across chromosomes. **b** Miami plot highlighting the INTACT nominated metabolites associated with T2D between two ancestries. Due to the size limit, we only show the top-10 EUR metabolites (top). We show the one AFR metabolite (bottom). The Y-axis is the INTACT posterior probability, with a threshold of 0.8. The X-axis is the metabolite associated genetic variant location across chromosomes. The metabolites were labeled with an anchor by their associated genetic variants. This was based on the chromosome of the variant with the strongest association with a given metabolite. **c** Venn diagram of findings per ancestry-matched protein-disease associations after INTACT jointly integrating the PWAS and colocalization results. **d** Venn diagram of findings per ancestry-matched metabolite-disease associations after INTACT jointly integrating the MWAS and colocalization results. **e** Schematic summary of the top enriched pathways using ancestry-stratified proteins or metabolites underlying T2D risk. (*P* values for the pathway enrichment analyses are unadjusted, two-sided, and determined via Fisher's exact test for metabolomic findings. The significance threshold of the pathway being enriched was q-value < 0.1 after multiple testing correction.) Schematics were created with icon library from the Microsoft PowerPoint in Affinity Designer.

AFR proteins (Supplementary Data 27: both proteins had the evidence from TWAS and colocalization) and 26 of 270 EUR proteins (Supplementary Data 28: 3 with colocalization, 5 with missense, 3 with pcHi-C links, 15 with TWAS plus colocalization) were reported as an effector gene already. As for metabolomics, one of one AFR metabolite (Supplementary Data 29: the associated gene *LINGO1* had a TWAS and colocalization evidence) and 18 of 72 EUR metabolites (Supplementary Data 30: 1 with colocalization and 17 with TWAS plus colocalization) with a significant gene-metabolite association were reported as known effector genes.

Moreover, to examine whether the nominated proteins or metabolites themselves were reported as an effector protein or metabolite, we compared our pinpointed molecular features against a recent preprint by Mandla and colleagues[12]. As the study by Mandla et al. represents the largest effort in mapping T2D-associated effectors, additional findings are expected. An effector protein or metabolite for T2D was defined as a protein or metabolite can be used to explain the genetic variants underlying the T2D risk. We found five of seven AFR proteins (Supplementary Data 31) and 136 of 270 EUR proteins (Supplementary Data 32) were reported as an effector protein already. For the effector metabolite comparisons, the only metabolite derived from our AFR-stratified analysis have not reported so far (Supplementary Data 33), while 45 of 72 EUR metabolites (Supplementary Data 34) were reported as effector metabolites. In summary, between 30 to 50% of the effector proteins and metabolites are not reported (except that the only nominated AFR metabolite, 3−sulfo−alanine, was also unreported).

Next, to detect the patterns of the proteins and metabolites implicated here in T2D, we performed pathway enrichment analyses. Specifically, using Gene Ontology, AFR proteins were enriched in several pathways (Fig. 4e, and Supplementary Fig. 10a, Supplementary Data 35), including biomineral tissue development (p value = 0.00164, q value = 0.0548), tissue remodeling (p value = 0.00211, q value = 0.0548), and stem cell differentiation (p value = 0.00252, q value = 0.0548); while EUR proteins were enriched in pathways (Fig. 4e, and Supplementary Fig. 10b, Supplementary Data 36), such as transmembrane receptor protein tyrosine kinase activity (p value = 4.12×10^{-07}, q value = 0.000178), and positive regulation of phosphatidylinositol 3-kinase/protein kinase B signal transduction (p-value = 9.13×10^{-07}, q value = 0.00276). Interestingly, there were four pathways shared between EUR and AFR T2D-associated proteins (Supplementary Fig. 10c), despite no proteins themselves were in common. These pathways were peptidyl-tyrosine modification (EUR proteins included CBL and CTF1, with a q value = 0.00763; AFR protein was GREM1, with a q value = 0.0941), peptidyl-tyrosine phosphorylation (EUR proteins included CBL and CTF1, with a q value = 0.0114; AFR protein was GREM1, with a q value = 0.0936), protein autophosphorylation (EUR proteins included TOM1L1 and TYRO3, with a q value = 0.0214; AFR protein was GREM1, with a q value = 0.0548), and vascular process in the circulatory system (EUR proteins included SVEP1 and ANGPT1, with a q value = 0.0214; AFR protein was ACE, with a q value = 0.0654).

As for metabolomics, we again calculated the pathway enrichment ratio to determine whether certain pathways were over-represented among all annotated pathways from Metabolon's database. AFR metabolites were enriched in the Methionine, Cysteine, SAM and Taurine Metabolism pathway (q value = 0.0207) (Fig. 4e, and Supplementary Fig. 11a, Supplementary Data 37); while EUR metabolites were enriched in sub-pathways (Fig. 4e, and Supplementary Fig. 11b, Supplementary Data 38), including Phosphatidylcholine (PC) (q value = 7.65×10^{-15}), Phosphatidylethanolamine (PE) (q value = 1.38×10^{-7}), and Progestin Steroids (q value = 0.0714). Moreover, we performed a cell-type-specific analysis using stratified-LDSC[30]. Regardless of ancestry, these proteins and metabolites displayed high expression in T cells and myeloid cells (Supplementary Fig. 12).

Finally, to nominate proteins and metabolites that are potential drug targets, we queried the proteins and metabolites against the Drugbank database[31]. As a result, we found one AFR protein, ACE, exhibited targetability by 13 FDA-approved drugs (Supplementary Data 39). On the other hand, 67 EUR proteins could be druggable by at least one FDA-approved drug (Supplementary Data 40). Furthermore, among three EUR metabolites (Supplementary Data 41) which were druggable, 5-oxoproline could be targeted by pidolic acid. Notably, pidolic acid has already been approved by the FDA to treat a family history of diabetes.

Discussion

Our study involved large-scale multi-ancestral multi-omic plasma-based QTL mapping from the same cohort. Using the INTACT approach, we identified key proteins and metabolites implicated in T2D. More importantly, even we found overlapping pathways across those two ancestries, our findings also revealed ancestry-specific results. It is known that different ancestries have different genetic architectures of the same traits, here we extend the genetic findings to the downstream functional analytes underlying T2D. Specifically, our EUR-specific analyses uncovered 270 proteins, and 72 metabolites associated with T2D (Supplementary Data 24, 26). Similarly, the AFR-specific analyses identified seven proteins and one previously unreported metabolite (Supplementary Data 23, 25).

LD differences between EUR and AFR may impact the ancestry specific QTL findings through affecting the detection of certain variants in the two ancestries: Given the same protein or metabolite, for example, variant-A is an ancestry-shared QTL, while variant-B is an ancestry-specific QTL. Different LD means the variant-A can be in high LD with variant-B ($r^2 > 0.8$, which indicates in the same LD block) from the EUR, but in low LD with variant-B ($r^2 < 0.2$, which indicates in different LD blocks) from the AFR.

Even though we identified specific proteins and metabolites when integrating the AFR- GWAS and QTLs, the power of identifying unique signals in this ancestry is still lower than in EUR, as the sample size in the disease GWAS and omic data is much lower in AFR, leading to fewer disease-associated loci and xQTLs. Therefore, future research with larger disease GWAS and QTL maps from diverse ancestral groups is

still necessary. In this study, we only nominated proteins or metabolites that met the criterion defined by the INTACT framework. It jointly considers both XWAS and colocalization. We believe that other analytical strategies may be worth considering for investigating other molecular traits. Our study represents a large-scale research endeavor that encompasses multi-ancestry and multi-omics analyses, allowing for a comprehensive exploration of diverse ancestral groups and molecular trait layers at the same time. Notably, we included trans-QTLs into the conventional cis-QTL framework[28,32], enabling the discovery of more heritable features and expanding our understanding of the genetic underpinnings of the studied traits. In addition, we integrated our multi-omic datasets with the INTACT approach to identify high-confidence proteins and metabolites implicated in T2D.

To our surprise, there were four protein-enriched pathways shared between EUR and AFR sets despite no proteins being shared. These pathways were peptidyl-tyrosine modification, peptidyl-tyrosine phosphorylation, protein autophosphorylation, and vascular process in the circulatory system. These common pathways unveiled ancestry-shared post-transcriptional mechanisms underlying T2D genetic architecture. These findings suggest that various proteins and their enriched pathways converge in the disease mechanisms underlying T2D pathobiology. Targeting these pathways could potentially benefit the general ancestries. On the other hand, some of the EUR-specific pathways identified include phosphatidylinositol 3-kinase/protein kinase B signal transduction and transmembrane receptor protein (tyrosine) kinase activity. In contrast, the AFR-specific pathways highlighted involve the development of biomineral tissue, kidneys, and the renal system. For pathways enriched by metabolites, the EUR set revealed three sub-pathways (PC, PE, and progestin steroids) all belonging to the lipid super-pathway, whereas the AFR set identified one sub-pathway from the amino acid super-pathway. PC and PE are the two most abundant phospholipids of mitochondria of mammalian cells and the mitochondrial dysfunction has been implicated in diabetes already[33]. Progestin steroids belong to sex steroid hormones, and these hormones were a risk factor for obesity and T2D via regulating adipose tissues[34].

There are several limitations to consider in our study. First, our study did not integrate proteomics and metabolomics data due to the lack of colocalization between protein, metabolite, and the ancestry-matched T2D risk loci. Nonetheless, we believe that genetic colocalization of proteomics and metabolomics could exist if we were not solely focused on T2D-associated loci. Second, there were unequal sample sizes between participants of EUR and AFR ancestry. This discrepancy in sample sizes affected the power to detect QTLs, even though we were well-powered to identify the sentinel variants (Supplementary Data 9). To mitigate this bias in identifying ancestry-specific findings, we employed standardized *z*-values that accounted for both effect size and standard error, rather than relying solely on *p* values, when comparing the two ancestral groups. Third, our study utilized plasma bulk-tissue, which may not reflect the cell type of interest when studying T2D, such as pancreatic islets or beta cells. But as plasma circulates throughout the body[35], our study holds value in investigating human metabolic disorders in general. Fourth, our study utilized one certain platform for measuring proteomics and the other for metabolomics, this can lead to platform-biased results. We, however, addressed this concern by querying our findings with external pQTL and mQTL studies that used both the same and different platforms. Fifth, our cohort consisted of participants with various disease statuses, including Alzheimer disease, frontotemporal dementia, and healthy individuals. We and other researchers, however, have reported that few pQTLs[18,36–38] or mQTLs[22,39,40] were status-specific. Thus, it suggests that the disease status is unlikely to have a significant impact, although further studies will be necessary. Sixth, we defined our participants based on genetic ancestry[41], thus we used the term "African". We acknowledge our study may contain admixed participants, which could potentially underestimate the ancestry-specific features observed. Seventh, after our manuscript was drafted, a new large-scale T2D GWAS was published by Suzuki and colleagues[7]. We anticipate more findings would be gained using this latest T2D genetic architecture.

In summary, by performing ancestry-matched omics-disease integration, it enhanced the accuracy of our findings compared to previous T2D studies that included ancestry-mixed data[5,6]. This approach may contribute to more precise identification of T2D effector genes within specific ancestral groups, as we uncovered different proteomic or metabolomic enriched pathways between AFR and EUR ancestral groups. To ensure the robustness of our conclusions, we cross-referenced our findings against the two largest multi-ancestry T2D studies[5,6], as well as one recent preprint[12] on mapping multi-modal effectors of T2D. This stringent evaluation allowed us to determine whether our identified effectors had been reported previously or not (on average, 80% of genes associated with the molecular phenotypes were not reported as effector genes and 40% of the molecular phenotypes were not reported as effector proteins and metabolites), further reinforcing the significance of our results. Our study demonstrates the power to discover effector genes/molecular traits of T2D via integrating the ancestry-matched datasets. We anticipate that our study might be informative to prioritize and/or rank potential targets and therapeutic strategies tailored to diverse genetic ancestries. We, however, acknowledge that a deeper examination still warrants an effective therapy to be developed with the relevant tissue of action, including interventional studies and randomized controlled trials. The findings of our study hold significant implications for advancing the understanding of T2D ethiology. For instance, the distinct pathways uncovered in EUR and AFR ancestry groups highlight the potential for personalized therapeutic interventions.

While our study focused on applying these plasma xQTLs to the study of T2D, it is worth noting that these QTL maps can be expanded to explore other diseases as well. These nominated proteins and metabolites are key intermediate phenotypes that can connect the genotype to the disease endpoint. Therefore, identifying these effectors in diverse ancestral groups may be an initial step toward developing more precise prediction models and therapies.

Methods

Ethics declarations

The Institutional Review Board (IRB) of Washington University School of Medicine in St. Louis approved the study with the IRB number 201109148, and research was performed in accordance with the approved protocols.

Cohort information

All participants were recruited at the Knight Alzheimer Disease Research Center (Knight ADRC). In total, 3170 participants from all genetic ancestries were selected for both proteomics and metabolomics profiling.

We used the TOPMed recommendations[42] when defining ancestries based on genetic information (See the following section "Genotype QC, imputation, and ancestral group stratification"). Therefore, we used the terms of "European (EUR)" and "African (AFR)" when referring to participants recruited at the Knight-ADRC (USA) with European and African genetic backgrounds, respectively and regardless of the country of origin. We used genetic principal component analysis to define the European or African ancestry for our downstream analyses (see the header "Genotype QC, imputation, and ancestral group stratification" for details). This has an important repercussion on the study design, as the goal of this study is to leverage ancestry-specific QTL datasets to perform the post-GWAS analyses. The existence of ancestry-specific QTL is because different ancestral populations have different LD structures.

The plasma proteomic and metabolomic datasets were generated from participants, which included 1254 AD patients, 1720 healthy controls, 34 frontotemporal dementia patients, and 162 individuals with an unclassified neurodegenerative disease. The cohort was a subset of the participants recruited from the Knight ADRC, which includes community-dwelling adults older than 27 years old via prospective studies of memory and aging since 1979. All participants recruited from the Knight ADRC are required to participate in core study procedures, including annual longitudinal clinical assessments, neuropsychological testing, neuroimaging, and biofluid biomarker studies. The corresponding genotype was a priori to choosing participants for profiling proteomics and metabolomics specifically in this study. If such a participant was genotyped, we prioritized the one for multi-omics. Plasma samples were collected in the morning after an overnight fast, immediately centrifuged, and stored at −80 °C.

Proteomics data QC

In brief, 3132 participants and 6907 aptamers passed proteomics QC. 7584 aptamers were measured before proteomics QC using the SomaScan 7k platform. Plasma proteomics data from all genetic ancestries were QCed with seven steps (Supplementary Fig. 13a, details see Supplementary Notes): Step 1) Limit of detection, scale factor difference, and coefficient of variation: we kept proteins/aptamers with ≥85% limit of detection, ≥0.5 scale factor difference, ≥0.15 coefficient of variation; Step 2) IQR-based outlier expression level detections: we log10-transformed the values and detection the outlier per the 1st and 3rd quantile for larger than 1.5 fold of IQR; Step 3) Remove analytes and samples with < 65% call rate (Supplementary Fig. 14a–b); Step 4) Re-calculate call rate for analytes and remove analytes with call rate < 85%; Step 5) Re-calculate missing rate for subjects and remove subjects with < 85% call rate cut-off (Supplementary Fig. 14c–d); Step 6) Back transformation into raw values from log10-scale; Step 7) Removal of Non-Human and analytes without protein targets and output the final matrix.

Metabolomics data QC

Briefly, 3169 participants and 1508 metabolites passed metabolomics QC. 1718 metabolites were measured before QC with the Metabolon HD4 platform. Plasma metabolomics data from all genetic ancestries were QCed with 11 steps (Supplementary Fig. 13b, details see Supplementary Notes): Step 1) Volume Normalization; Step 2) Sample Missingness: we kept samples < 50% missingness; Step 3) Metabolite Missingness: we kept metabolites from non-xenobiotics groups < 80% missingness; Step 4) Fischer's Exact Test and Differential Expression Check to recover metabolites if their missingness was associated with disease status; Step 5) Minimum Value Imputation on non-xenobiotic group of metabolites; Step 6) Remove non-informative metabolites after log10-transformation: non-informative metabolites were defined as IQR = 0 and variance < 0.001; Step 7) IQR-based outlier detection: remove outlier values based on if exceed the 1st and 3rd quantile for larger than 1.5*IQR; Step 8) Metabolites with < 50 data points; Step 9) Sample Outlier removal based on metabolite-PCA: > 5 standard deviation of the PC1 and PC2; Step 10) Batch effects of metabolomics data; Step 11) output the final matrix and back transformation of metabolite levels. Notably, for step 5, missing values were imputed. However, as mentioned earlier Xenobiotics are indeed expected to be missing, and imputing their values could skew the results. So, imputation is performed only for the non-xenobiotic group of metabolites. The metabolomic imputation was later tested and was determined not to affect the effect size of the mQTL identified in both ancestries (see "Identification of mQTLs" for the modeling), compared to the values without imputation (Supplementary Fig. 15a–d).

Genotype QC, imputation, and ancestral group stratification

At the pre-imputation stage (see **Supplementary Notes** for the details), the directly genotyped variants were kept agreeing to three criteria: (1) genotyping successful rate ≥98% per variant or per individual; (2) MAF ≥ 0.01; and (3) Hardy–Weinberg equilibrium (HWE) ($P \geq 1 \times 10^{-6}$). Imputation was performed on the TOPMed[3] imputation server using the hg38 Version R2 reference panel. The TOPMed Imputation Reference panel contains information from 97,256 deeply sequenced human genomes. Imputed genotypes with imputation quality of $R^2 \geq 0.3$ were kept. At the post-imputation stage, the genotyped and imputed variants remained on the two criteria: (1) genotyping missing rate ≤90% per variant; (2) MAF≥ 0.0005. Multiple genotyping arrays were included for this cohort. Including CoreEx, GSA_v1, GSA_v2, GSA_v3, Human1M.Duov3, NeuroX2, OmniEx, quad660 (Supplementary Data 1). For each genotype array, we performed pre-imputation, imputation, and post-imputation separately. We merged all into one dataset before performing the QTL analyses (see **Supplementary Notes** for the details). Thus, the final number of genetic variants located on autosomal chromosomes was 10,448,203. In total, 3081 out of 3170 participants had corresponding genotype data.

Ancestral group stratification was performed using Plink1.9[43] pca function, 2598 participants were classified as EUR and 433 as AFR per principal component analysis (PCA) with the reference by 1000 Genome project (Supplementary Fig. 16a–d, and Supplementary Data 2). We defined the genetic ancestry per genotype PCA anchored with participants from 1000 Genome Project within the boundary within the mean +/- 3 times of the standard deviation of each PC. Moreover, we noted that over 400 AFR participants we defined as AFR were within 3 standard deviations of the AFR participants in 1000 Genome project, we cannot omit the admixed ancestral groups within these participants. Relatedness was performed using plink1.9[43] genome function on the IBD. Unrelated participants were defined as PI_HAT < 0.25. 2395 EUR and 418 AFR participants were kept as unrelated participants.

Identification of pQTLs

An additive linear regression model was used from plink2[43] glm function for each protein. Protein-abundances were log-10 transformed first and z-scale normalized next. Covariates were age, sex, genotyping array types, genotype PC 1-10, and proteomics PC 1-2 (to correct such batch effects from the proteomics data alone: we identified two different batches when visualizing the scatterplots of proteomic PC1 and PC2 (Supplementary Fig. 14e). But after adjusting for the proteomic PC1 and PC2, the batch effect was corrected (Supplementary Fig. 14f)). Genotype array types included Quad660, CoreEx, GSA_v1, GSA_v2, GSA_v3, NeuroX2, Human1M.Duov3, when using as covariates, dummy variable included n-1 rather n. The final sample size for EUR and AFR pQTL analyses were 2338 and 414 (Supplementary Fig. 1a, and Supplementary Data 3, average age EUR = 75 and AFR = 75; female percentage EUR = 54% and AFR = 71%). The final numbers of proteins for EUR and AFR pQTL analyses were both 6907 (Supplementary Fig. 1b, and Supplementary Data 4–5).

For cis and trans definitions, we used a window of the variants within 1 Mb upstream and downstream of the gene start site by which each protein was coded. The cis threshold was 5×10^{-8}, as cis pQTLs were only corresponding to the variants near the protein encoded by its gene, thus no further multiple testing correction on the independent proteins needs to be considered across the proteome. For the trans-pQTL analysis, we used the study-wide significance considering the number of independent proteins, as for non-cis variants given the same corresponding protein, the additional protein factor needs to be considered. As many proteins were correlated, we used the proteomic PCA to estimate the effective number of independent proteins within each ancestry. The estimated number of independent proteins was determined as the minimum number of protein PCs that cumulatively explain 95% of the variance given the proteomics expression matrix per ancestry after QC. We calculated that the estimated number of

independent proteins for EUR and AFR were 1472 and 336, respectively. They were next used to adjust the p-value for multiple testing of trans-pQTL using a Bonferroni correction for EUR was 3.40×10^{-11} (5×10^{-8}/1472) and for AFR was 1.49×10^{-10} (5×10^{-8}/336).

Identification of mQTLs

An additive linear regression model was used from plink2[43] glm function for each metabolite. Metabolite levels are first normalized by the median value given the same metabolite and log-10 is transformed next. Covariates are age, sex, genotyping array types, genotype PC1-10, and metabolomics PC 1-2. Genotype array types included Quad660, CoreEx, GSA_v1, GSA_v2, GSA_v3, NeuroX2, Human1M.Duov3, when using as covariates, dummy variable included n-1 rather n. The final sample size for EUR and AFR mQTL analyses was 2392 and 417 (Supplementary Fig. 1a, and Supplementary Data 3, average age EUR = 75 and AFR = 75; female percentage EUR = 54% and AFR = 71%). The final numbers of metabolites for EUR and AFR mQTL analyses were 1483 and 1413 (Supplementary Fig. 1b, and Supplementary Data 6–7).

For the mQTL analysis, the number of independent metabolites of EUR and AFR metabolomics used as denominators were 766 and 281, respectively. (The number of metabolites was derived as the minimum metabolite PC number that cumulatively explains 95% of the variance for the metabolomics expression matrix of each ancestry after QC.) Thus, the p-value threshold for EUR was 6.53×10^{-11} (5×10^{-8}/766) and for AFR was 1.78×10^{-10} (5×10^{-8}/281).

Filtering the inflation features

For the inflated features (i.e., associated with variants over 5/3/7/3 different chromosomes corresponding to EUR pQTL/AFR pQTL/EUR mQTL/AFR mQTL [the thresholds are collected empirically]), we first removed the variants given this feature with MAF < 0.05 and genotyping call rate < 97%. If we found the features were still inflated, we removed the features eventually. The unique features of removal were listed below: EUR proteomics: 142 aptamers; AFR proteomics: 132 aptamers; EUR metabolomics: six metabolites; AFR metabolomics: five metabolites.

Annotation of the xQTLs

To annotate our QTL findings, we used the command line tool Variant Effect Predictor (VEP[44]) from the Ensembl-version107. We used the default options for all four QTL maps.

Replication of xQTLs with external studies

To replicate our QTL findings, we queried all study-wide feature-variant pairs from our study against several largest external studies. These studies all released their full summary statistics and set the genetic coordinates in the hg38. The proxy variant was defined as LD $r^2 \geq 0.8$ using the reference at TOPMed[3] WGS data curated by the tool TOP-LD[45]. The LD used from TOP-LD for LD was matched separately for EUR and AFR to match this study.

To replicate our proteomics findings, we used four datasets from three studies. Ferkingstad et al.[20], used 35k participants of European ancestry and the SomaScan 5k platform to measure plasma proteome. Surapaneni et al.[16], used 466 participants of African ancestry and the SomaScan 7k platform to measure serum proteome. Sun et al.[17], used 34k participants of European ancestry as well as 931 participants of African ancestry and the OLINK 3k platform to measure plasma proteome. We set six categories when comparing the study-wide significant findings from this study and its corresponding external studies: 1) validated with a p value below the Bonferroni-corrected study-wide threshold (5×10^{-11}) account for 1000 independent features; 2) known with a p value below the genome-wide threshold (5×10^{-8}); 3) replicated with a p value below the nominal threshold (5×10^{-2}); 4) not replicated with a p value greater or equal to the nominal threshold; 5) not reported with a matching protein but a missing proxy variant; 6) not reported with a non-matching protein.

To replicate our metabolomics findings, we used three studies. Yin et al.[22], used 6136 participants of European ancestry from Finland and the Metabolon HD4 platform to measure plasma metabolome. Chen et al.[23], used 8299 participants of European ancestry and the Metabolon HD4 platform to measure plasma metabolome. Rhee et al.[24], used 687 participants of African ancestry and the Broad Institute platform to measure plasma metabolome. For mQTLs of African ancestry by Rhee et al. 2022, the number of metabolites overlapping between their platform (Broad Institute) with our platform (Metabolon HD4) was 207. This overlap was performed via HMDB-ID matching, rather than chemical name (Supplementary Data 15). We set four categories when comparing the study-wide significant findings from this study and its corresponding external studies: 1) validated with a p value below the Bonferroni-corrected study-wide threshold (5×10^{-11}) account for 1000 independent features; 2) known with a p value below the genome-wide threshold (5×10^{-8}); 3) replicated with a p value below the nominal threshold (5×10^{-2}); 4) not replicated with a p value greater or equal to the nominal threshold; 5) not reported with a matching metabolite but a missing proxy variant; 6) not reported with a non-matching metabolite.

Pathway enrichment of metabolites with mQTLs

To test the pathway enrichment of metabolites with mQTLs in AFR or EUR set, the analysis was performed using the Metabolon's officially annotated super pathway information[46]. The outcome was tested based on the enrichment ratio of the metabolite with mQTL (pathway-1 alone over all pathways) over the metabolite passed QC (pathway-1 alone over all pathways) with Fisher exact test. For example, in AFR metabolomics dataset, we found 27 metabolites with one mQTL belong to amino acid pathway, 65 metabolites with one mQTL belong to eight different pathways; 211 metabolites passed QC belong to amino acid pathways, 1413 metabolites passed QC belong to ten different pathways. Thus, the enrichment ratio is derived as 2.78 per (27/65)/(211/1413).

Definition of LD block and pleiotropy

To define LD blocks for each ancestry, we used the 1000 Genome project EUR (1703 blocks) and AFR (2583 blocks) as the reference ancestral group per ldetect by Berisa and Pickrell[19]. We performed liftover to map the hg19 coordinates into hg38. We next used the index to group the variants and obtained the pleiotropic region, which was the index associated with multiple molecular traits. For proteomics, we used karyoploteR[47] package to visualize the top findings as an ideogram. For metabolomics, we used circlize[48] package to visualize the top findings as a chord diagram.

Identification of ancestry-specific QTLs

Ancestry-specificity was defined as fold-change over 10-fold or below 0.1-fold between the Z-normalized effect sizes (beta divided by standard error) of the protein-variant pairs or metabolite-variant pairs given the same variants. The fold-changes of the same feature-variant pairs were also calculated after setting 3 bins of MAF as 0 to 0.01, 0.01 to 0.05, and 0.05 to 0.5.

MASH method[26] (implemented as mashR package) was also used. Briefly, after fitting the model into the mash function with the beta and standard error of the same QTLs from both EUR and AFR datasets as the input plus the covariance matrices set up given the same input. The fold-change of posterior means of the protein-variant pairs or metabolite-variant pairs given the same variants were calculated. The same thresholds of 10-fold and 0.1-fold were used to determine QTL sharing or not.

Boxplots were drawn with the ggplot2 package; Locus-zoom plots were drawn with the LocusZoom.js tool[49].

Power analysis of ancestry-specific QTLs

We performed two separate power analyses using the powerEQTL.ANOVA function from the R package powerEQTL[50] is listed below:

a) We calculated the power values for all our current sentinel variants from each of the four QTL sets after splitting them into ancestry-shared and ancestry-specific subtypes.

Given the input of MAF, the average standardized effect size per ancestry-specificity, the sample size, and using the genome-wide significance thresholds (FWER = 0.05, nTests=1e6), we found all the sentinel variants from each of the four QTL sets with a power of 0.8 or more (Supplementary Data 21).

b) We next fixed the effect size and number of tests, while varying MAF per each of the four QTL sets. We split the MAF into minimum, average, and maximum by ancestry-shared and specific xQTLs. We empirically learned the MAF and the average standardized effect size and used the genome-wide significance thresholds for consistency between ancestries and specificity of xQTLs.

We found that min-MAF led to underpowered findings, especially in the ancestry-specific xQTLs. For example, in the EUR mQTL set, the power was 0.007 for identifying EUR-specific findings in AFR given minMAF from ancestry-specific variants (Supplementary Data 22). On the other hand, the power turned to 1 for identifying the EUR-shared mQTL in AFR given minMAF from shared variants.

Cross-reference of the ancestry-specific xQTLs with external studies

We cross-referred to Zhang et al.[28], for proteomics to examine the proportion of the ancestry-specific pQTLs. We calculated the percentage of ancestry-specific pQTLs using the variable "EA-specific" as TRUE over all EUR-pQTLs in their Supplementary Table 3.1 and the variable "AA-specific" as TRUE over all AFR-pQTLs in their Supplementary Table 3.2.

We cross-referred Rhee et al.[24], for metabolomics datasets to check the proportion of the ancestry-specific mQTLs. We derived the percentage of ancestry-specific mQTLs with ones outside the 10-fold-change per the Z-normalized effect sizes (beta divided by standard error) of EUR and AFR over the total 45 mQTLs in their Table 2.

T2D risk GWAS

We used the two population-scale T2D risk GWAS from EUR and AFR, separately.

For EUR T2D risk GWAS, we used the summary statistics from the study by Mahajan et al.[5], covering 80,154 cases and 853,816 controls. The full summary statistics for EUR and multi-ancestry GWAS were available at (http://diagram-consortium.org/downloads.html), while AFR GWAS was not available as of September 2023.

Thus, for AFR T2D risk, we turned to another study from Vujkovic et al.[6], containing 24,646 cases and 31,446 controls. The full summary statistics for AFR GWAS were downloaded after approval for the request authorize access via dbGAP at (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/analysis.cgi?study_id=phs001672.v11.p1&pha=4943).

We plotted the two summary statistics along with the ancestry-matched molecular modalities using CMplot[51].

INTACT procedure of proteins/metabolites on T2D

We used the INTACT framework[29] to compute a posterior probability (PP) for each protein/metabolite underlying T2D risk with PWAS or metabolome-wide association study (MWAS) and colocalization as input: In total, we tested 2184 EUR and 732 AFR proteins; 388 EUR and 53 AFR metabolites given a non-missing value for the input. We defined the putative causal protein or metabolite given the INTACT PP > 0.8. We detailed PWAS/MWAS/colocalization methods separately in the next headings.

PWAS weight calculation

A modified version of FUSION[11] was used as for the proteins associated with more than one genetic region, all variants from each region were included in the weight calculation. The same covariates used for pQTL identification were used for PWAS weight calculation. The SNP-based heritability of each protein SOMAmer was estimated using the GCTA GREML[52] tool. The proteins with negative h^2 values were removed before performing the weight calculation. The window size of the sentinel QTL region was ± 1 Mb. Using the FUSION R package, we constructed imputation models for 881 AFR and 2400 EUR SOMAmers, as we only focused on the proteins with at least one study-wide significant pQTLs, not all proteins in the study. The imputation model for a SOMAmer was trained by the best models (out of Elastic Net, TOP1, and BLUP) using all variants in ± Mb upstream and downstream of the sentinel pQTL sites of the target protein. The Elastic Net model was refitted using all data and the tuning parameters per 5-fold cross-validation. The metric of best model was based on the cross-validation performance with the minimum p value out of the three models including Elastic Net, TOP1, and BLUP.

PWAS association test with T2D

A modified version of FUSION[11] was used as we incorporated multiple regions into account by breaking the single-chromosome requirement per protein-disease associations. We used the 881 (AFR) and 2400 (EUR) proteins with imputation models to perform the PWAS on the ancestry-matched T2D risk. The tool Functionally-informed Z-score Imputation (FIZI[53]) was used first to impute the summary statistics of AFR T2D risk (Vujkovic et al., [6]) and EUR T2D risk (Mahajan et al., [5]) with the in-sample reference linkage-disequilibrium (LD) information.

The multiple testing corrections for the PWAS results were selected per the total number of imputation models for all weight-non-missing plasma proteins (p value < 0.05/797 in AFR and 0.05/2285 in EUR). The Z value from PWAS was used to determine the effect size of protein-T2D associations within each ancestry.

MWAS weight calculation

Similar to the above PWAS weight calculation section, a modified version of FUSION[11] was used, as for the same metabolites associated with more than one genetic region, all variants from each region were included in the weight calculation. The same covariates used for mQTL identification were used for MWAS weight calculation. The SNP-based heritability of each metabolite was estimated using the GCTA GREML[52] tool. The metabolites with negative h^2 values were removed before performing the weight calculation. The window size of the sentinel QTL region was ± 1 Mb. Using the FUSION R package, we constructed imputation models for 60 AFR and 403 EUR metabolites, as we only focused on the metabolites with at least one study-wide significant mQTLs, not all metabolites in the study. The imputation model for a metabolite was trained by the best models (out of Elastic Net, TOP1, and BLUP) using all variants in ± 1 Mb upstream and downstream of the sentinel mQTL sites of the corresponding metabolite. The Elastic Net model was refitted using all data and the tuning parameters per 5-fold cross-validation. The metric of best model was based on the cross-validation performance with the minimum p value out of the three models including Elastic Net, TOP1, and BLUP.

MWAS association test with T2D

Similar to the above PWAS association test section, a modified version of FUSION[11] was used as we incorporated multiple regions into account by breaking the single-chromosome requirement per metabolite-disease associations. We used the 60 (AFR) and 403 (EUR) metabolites

with imputation models to perform the MWAS on the ancestry-matched T2D risk (AFR T2D risk (Vujkovic et al.[6]) and EUR T2D risk (Mahajan et al.[5])) after FIZI imputation.

The multiple testing corrections for the MWAS results were selected per the total number of imputation models for all weight-non-missing plasma proteins (p value < 0.05/401 in AFR and 0.05/58 in EUR). The Z value from MWAS was used to determine the effect size of metabolite-T2D associations within each ancestry.

Colocalization of molecular traits (proteins/metabolites) and T2D

We performed colocalization analysis using both coloc.abf function from R package coloc v3.1[8] and coloc.susie function from R package coloc v5.1[9] with a wrapper for susie_rss function from susieR[54,55] package. We next set the window size to +/- 1 Mb centering on IV per trait-T2D pair. We used default priors, with p1 as 1×10^{-4}, p2 as 1×10^{-4}, and p12 as 1×10^{-5}. Evidence for colocalization was assessed using the posterior probability (PP) for hypothesis 4 (indicating there is an association for both protein and disease and they are driven by the same causal variant(s)). We used PP.H4_final > 80% as a threshold to suggest that associations were highly colocalized. Under the assumption of only a single causal variant, we used the PP.H4 from coloc.abf output of the trait-disease pair. Under the assumption that multiple causal variants exist[54], we used the maximum PP.H4 of multiple credible sets from coloc.susie output.

Identification of ancestry-specific T2D effector findings

Comparison of the posterior probability (PP) with the same analyte-T2D associations given the same trait from the two ancestries. If the PP of analyte-T2D association from both ancestries were higher than 0.8, the analyte was ancestry-shared. If the PP of analyte-T2D association from only one ancestry was higher than 0.8, the analyte was ancestry-specific. The Miami plots for proteomics-T2D and metabolomics-T2D comparing EUR and AFR findings were plotted using the R package hudson[56].

Cross-reference on the effector genes of proteins/metabolites on T2D

We first assembled the effector genes of T2D risk GWAS from two multi-ancestry studies[5,6]. For the study by Mahajan et al., we combined their missense annotations, colocalization with seven-tissue eQTL and plasma pQTL, pcHi-C annotations. In total, 834 unique genes were nominated by the authors. For the study by Vujkovic et al., we combined their missense annotations, TWAS, and colocalization results from 52 tissue eQTLs. In total, 754 unique genes were nominated by the authors. We next queried our protein and metabolite findings using the nearest gene to the genetic locus. If the gene can be found in the effector list, we define it as the finding that was reported.

Cross-reference on the effector proteins or metabolites themselves on T2D

We first assembled the lists of effector proteins and metabolites of T2D risk from a recent preprint[12]. For the effector protein list by Mandla et al., we queried the protein name to map against our protein list after INTACT. In total, 1572 unique effector proteins were found in the SomaScan7k platform. For the effector metabolite list by Mandla et al., we queried the metabolite ID to map against our metabolite list after INTACT. In total, 390 unique effector metabolites were in overlap with the Metabolon platform. We next compared our protein and metabolite lists with each data separately to determine whether the proteins/metabolites were reported as an effector itself.

Pathway enrichment of proteins and metabolites implicated in T2D

To test the pathway enrichment of proteins pinpointed via INTACT in EUR and AFR stratified analyses, we used enrichGO function from clusterProfiler package[57]. The background list contained all SomaScan 7k proteins that passed QC. P values are unadjusted, two-sided, and determined via over-representation test. The significance threshold of the pathway being enriched was q value < 0.1.

To test the pathway enrichment of metabolites pinpointed via INTACT in EUR and AFR stratified analyses, we used the Fisher exact test testing against the Metabolon's officially annotated sub-pathway information. The outcome was tested based on the enrichment ratio of the metabolite with INTACT evidence (sub-pathway-A alone over all other pathways) over the metabolite passed QC (sub-pathway-A alone over all other pathways). P values are unadjusted, two-sided, and determined via Fisher's exact test for metabolomic findings. The significance threshold of the pathway being enriched was q value < 0.1.

Cell-type specificity analysis of the proteins/metabolites on T2D

Using the nomination of INTACT results, we performed cell-type specificity analysis of the proteins/metabolites on T2D using S-LDSC[30] with the cell-type annotation from ImmGen. Overall, 295 cell subtypes were used and grouped into five major cell types: B, Myeloid, Natural Killer (NK), T, and other cells. The input for S-LDSC was each protein or metabolite genome-wide summary statistics. The output of the S-LDSC was the regression coefficient of the cell type-specific annotation from ImmGen as well as the p value of the coefficient for each feature. We then ranked the feature by the p value and used the top cell type as the enriched cell type for this feature. We used two input files: one from the INTACT integration results; and the other from all features with at least one QTL. We calculated the fold-change of the two sets given the same feature-ancestry combination (i.e., EUR proteins, AFR proteins, and EUR metabolites, for AFR metabolite, no feature remained).

Druggable target query of the proteins/metabolites implicated in T2D

To nominate druggable targets for repositioning, we queried the proteins and metabolites nominated from the INTACT results against the drugbank database[31] (drugbank_5.1.10.db) downloaded locally. For proteins, we first downloaded the csv file for all Drug Target Identifiers (https://go.drugbank.com/releases/latest#protein-identifiers). The uniprotID and drugbankID were linked in the csv file. We next used the proteins with an overlapping uniprotID to query the corresponding drugbankID via the drugbankR R package (https://github.com/girke-lab/drugbankR). For metabolites, we first used the hmdbID to query via the hmdbQuery R package (https://github.com/vjcitn/hmdbQuery) and extracted the "drugbank_id" for the final query via the drugbankR.

PheWeb Browser for interactively visualizing QTL datasets

To assist users in navigating our QTL results, we implemented the PheWeb[25] to visualize all 16,710 traits (EUR-proteins: 6907; AFR-proteins: 6907; EUR-metabolites: 1483; AFR-metabolites: 1413). The URL is (https://ontime.wustl.edu/), all aligned to the hg38 genomic coordinates.

Reporting summary

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

Data availability

The summary statistics of all four EUR and AFR protein and metabolite QTLs generated in this study have been deposited in the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) with accession NG00180 (https://dss.niagads.org/datasets/) and GWAS catalog (https://www.ebi.ac.uk/gwas/). The NIAGADS team is processing the uploaded files as of 30 MAY 2025. The accession numbers for the GWAS catalog are from GCST90607080 to GCST90623789. The four datasets also are available in the BOX folders listed below: EUR-pQTL: (https://wustl.box.com/s/

a009fncyeykdsyslnci0npa1q2z4vk4d), AFR-pQTL: (https://wustl.box.com/s/sn43u2254qx77h4ekwge073ipmi949fy); EUR-mQTL: (https://wustl.box.com/s/k0pdz3akrwc2nl9fo05i0ng9qs84yspr); AFR-mQTL: (https://wustl.box.com/s/vyxnyduanon1xf8arevk5y4hp8jazl53) The PheWeb browser for visualizing and downloading all plasma omics QTL results is at (https://ontime.wustl.edu/). Plasma proteomics and metabolomics data for the Knight ADRC participants is available at the Knight ADRC at: (https://live-knightadrc-washu.pantheonsite.io/professionals-clinicians/request-center-resources/). Requests for clinical or proteomic/metabolomic data from individual investigators will be reviewed to ensure compliance with patient confidentiality. Please see knightadrc.wustl.edu for details on accessing available data and study protocols. The Genotype data for all the Knight ADRC participants is available at the NIAGADS at (https://dss.niagads.org/datasets/ng00127/).

Code availability

The scripts for performing TWAS with both cis and trans regions are available at https://github.com/cyang-2014/Plasma2omic2pops/tree/master[58].

References

1. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
2. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
3. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
4. Zhou, W. et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).
5. Mahajan, A. et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet* **54**, 560–572 (2022).
6. Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet* **52**, 680–691 (2020).
7. Suzuki, K. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* 1–11 https://doi.org/10.1038/s41586-024-07019-6 (2024).
8. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet.* **10**, e1004383 (2014).
9. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
10. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet* **47**, 1091–1098 (2015).
11. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
12. Mandla, R. et al. Multi-omics characterization of type 2 diabetes associated genetic variation. 2024.07.15.24310282 Preprint at https://doi.org/10.1101/2024.07.15.24310282 (2024).
13. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLOS ONE* **5**, e15004 (2010).
15. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal. Chem.* **81**, 6656–6667 (2009).
16. Surapaneni, A. et al. Identification of 969 protein quantitative trait loci in an African American population with kidney disease attributed to hypertension. *Kidney Int* **102**, 1167–1177 (2022).
17. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* 1–10 https://doi.org/10.1038/s41586-023-06592-6 (2023).
18. Yang, C. et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat. Neurosci.* **24**, 1302–1312 (2021).
19. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
20. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00978-w (2021).
21. Lotta, L. A. et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).
22. Yin, X. et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. *Nat. Commun.* **13**, 1644 (2022).
23. Chen, Y. et al. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat. Genet.* 1–10 https://doi.org/10.1038/s41588-022-01270-1 (2023).
24. Rhee, E. P. et al. Trans-ethnic genome-wide association study of blood metabolites in the Chronic Renal Insufficiency Cohort (CRIC) study. *Kidney Int.* **101**, 814–823 (2022).
25. Gagliano Taliun, S. A. et al. Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).
26. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
27. Kelly, D. E. et al. The genetic and evolutionary basis of gene expression variation in East Africans. *Genome Biol.* **24**, 35 (2023).
28. Zhang, J. et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* https://doi.org/10.1038/s41588-022-01051-w (2022).
29. Okamoto, J. et al. Probabilistic integration of transcriptome-wide association studies and colocalization analysis identifies key molecular pathways of complex traits. *Am. J. Hum. Genet.* **110**, 44–57 (2023).
30. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet* **50**, 621–629 (2018).
31. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074–D1082 (2018).
32. Schubert, R. et al. Protein prediction for trait mapping in diverse populations. *PLOS ONE* **17**, e0264341 (2022).
33. van der Veen, J. N. et al. The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. *Biochimica et. Biophysica Acta (BBA) - Biomembranes* **1859**, 1558–1572 (2017).
34. Mayes, J. S. & Watson, G. H. Direct effects of sex steroid hormones on adipose tissues and obesity. *Obes. Rev.* **5**, 197–216 (2004).
35. Oh, H. S.-H. et al. Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
36. Robins, C. et al. Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **48**, 400–410 (2021).
37. Hansson, O. et al. The genetic regulation of protein expression in cerebrospinal fluid. *EMBO Molecular Medicine* **n/a**, e16359 (2022).
38. Western, D. et al. Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and informs

causal proteins for Alzheimer's disease. *Research Square* https://doi.org/10.21203/rs.3.rs-2814616/v1 (2023).
39. Schlosser, P. et al. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0567-8 (2020).
40. Wang, C. et al. Unique genetic architecture of CSF and brain metabolites pinpoints the novel targets for the traits of human wellness. (2023).
41. Borrell, L. N. et al. Race and genetic ancestry in medicine - a time for reckoning with racism. *N. Engl. J. Med* **384**, 474–480 (2021).
42. Khan, A. T. et al. Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genomics* **2**, 100155 (2022).
43. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
44. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
45. Huang, L. et al. TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. *Am. J. Hum. Genet.* **109**, 1175–1181 (2022).
46. Pietzner, M. et al. Plasma metabolites to profile pathways in noncommunicable disease multimorbidity. *Nat. Med* **27**, 471–479 (2021).
47. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
48. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in. *R. Bioinforma.* **30**, 2811–2812 (2014).
49. Boughton, A. P. et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
50. Dong, X. et al. powerEQTL: an R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab385. (2021).
51. Yin, L. et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteom. Bioinforma.* **19**, 619–628 (2021).
52. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proc. Natl. Acad. Sci.* **113**, E4579–E4580 (2016).
53. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
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. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genet.* **18**, e1010299 (2022).
56. Lucas, A. hudson: an R package for creating mirrored Manhattan plots. (2020).
57. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov. (Camb.)* **2**, 100141 (2021).
58. Yang, C. European and African ancestry-specific plasma protein-QTL and metabolite-QTL analyses identify ancestry-specific T2D effector proteins and metabolites; cyang-2014/Plasma2omic2pops: v1.0. Zenodo https://doi.org/10.5281/ZENODO.15442277 (2025).

Acknowledgements

We thank all the participants for giving consent included in this study. We thank John Budde and Pat Kohlfeld for sample preparation and shipment.

Funding: Research reported in this publication was supported by National Institute on Aging (NIA) of the National Institutes of Health under grant number [R01AG044546 (C.C.), P01AG003991 (C.C.), RF1AG053303 (C.C.), RF1AG058501 (C.C.), U01AG058922 (C.C.), RF1AG074007 (Y.J.S.)]. This work was also supported by grants from the Chan Zuckerberg Initiative, the Michael J. Fox Foundation (C.C.), and the Alzheimer's Association Zenith Fellows Award (ZEN-22-848604, C.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

C.C. and C.Y. conceived the study. C.Y. performed most of the genetic analyses. P.G. conducted the genotype imputation, and Q.C., J.T., and L.W. performed the proteome Q.C. M.L. managed the clinical data. J.T. and C.W. performed the metabolome Q.C. W.B. carried out the sample preparation and shipment. YW implemented the PheWeb. YJS and CC acquired the funding. C.C. supervised the study. C.Y. and C.C. wrote the original draft. F.U. edited the revised version and provided intellectual insights on T2D findings. C.Y., J.T., C.W., and C.C. prepared the peer-review documents. All authors edited and approved the manuscript.

Competing interests

CC has received research support from GSK and EISAI. The funders of the study had no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. CC is a member of the advisory board of Circular Genomics and owns stocks. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-62463-w.

Correspondence and requests for materials should be addressed to Carlos Cruchaga.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.