Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας στην Ελλάδα

ΠΑΡΑΔΟΤΕΟ Π5.2

«Εκτίμηση κινδύνου με πολυομικά σκορ»

Φορέας	Ερευνητικό Κέντρο Ελληνικό Ινστιτούτο Παστέρ
Τύπος Παραδοτέου	Έκθεση
Ημερομηνία Υποβολής Παραδοτέου	15 Φεβρουαρίου 2026
Ενότητα Εργασίας	Ενότητα Εργασίας 5 Κοινά νοσήματα: Προγνωστικά μοντέλα πρόβλεψης κινδύνου χρόνιων νοσημάτων

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Δημοσίευση 1 «**Metabolic signature of Mediterranean diet adherence and risk of dementia: A prospective analysis in the UK Biobank cohort»**: Έχει υποβληθεί και είναι σε κρίση στο διεθνές επιστημονικό περιοδικό Neurology.

Η παρούσα μελέτη ερεύνησε τη σχέση μεταξύ ενός μεταβολικού «αποτυπώματος» (metabolic signature) στο πλάσμα του αίματος, το οποίο αντικατοπτρίζει την προσκόλληση στη Μεσογειακή Διατροφή, και του κινδύνου εμφάνισης άνοιας καθώς και των υποτύπων της.

Για την εξαγωγή των συμπερασμάτων, αναλύθηκαν δεδομένα από 92.299 συμμετέχοντες της μεγάλης βρετανικής βάσης δεδομένων UK Biobank. Οι ερευνητές χρησιμοποίησαν δεδομένα μεταβολωμικής που προέκυψαν μέσω της πλατφόρμας NMR, καθώς και διατροφικές πληροφορίες από ερωτηματολόγια ανάκλησης 24 ωρών. Η αξιολόγηση της προσκόλλησης στη Μεσογειακή Διατροφή πραγματοποιήθηκε με τη χρήση του δείκτη MEDAS. Στη συνέχεια, το μοντέλο των μεταβολιτών που αναπτύχθηκε επικυρώθηκε εξωτερικά σε ένα ανεξάρτητο δείγμα 1.608 ατόμων από την ελληνική μελέτη Epirus Health Study (EHS).

Κατά τη διάρκεια μιας μέσης περιόδου παρακολούθησης 13,5 ετών, εντοπίστηκαν 975 νέα περιστατικά άνοιας από κάθε αιτία. Η ανάλυση ανέδειξε ένα συγκεκριμένο αποτύπωμα 26 μεταβολιτών, το οποίο αποτελούνταν κυρίως από λιπίδια και σχετιζόμενους με λιπαρά οξέα μεταβολίτες, επιλεγμένα αμινοξέα, δείκτες του μεταβολισμού της ενέργειας, και παράλληλα χαρακτηριζόταν από χαμηλότερα επίπεδα GlycA, που αποτελεί δείκτη συστημικής φλεγμονής. Το συγκεκριμένο μεταβολικό αποτύπωμα βρέθηκε να συσχετίζεται αντίστροφα με τον κίνδυνο άνοιας από κάθε αιτία, υποδεικνύοντας προστατευτική δράση. Αντίθετα, η αρχικά θετική συσχέτιση με τη νόσο του Αλτσχάιμερ αποδυναμώθηκε όταν λήφθηκαν υπόψη πολλαπλές μεταβλητές, ενώ δεν καταγράφηκε στατιστικά σημαντική συσχέτιση με την αγγειακή άνοια.

Συμπερασματικά, τα ευρήματα της έρευνας επιβεβαιώνουν ότι το μεταβολικό αποτύπωμα πλάσματος που συνδέεται με τη Μεσογειακή Διατροφή σχετίζεται πράγματι με μειωμένο κίνδυνο άνοιας σε πληθυσμιακό επίπεδο. Τα αποτελέσματα αυτά υποστηρίζουν τη χρήση της ανάλυσης του μεταβολώματος ως μια αξιόπιστη και αντικειμενική μέθοδο που συμπληρώνει τα

παραδοσιακά εργαλεία διατροφικής αξιολόγησης, ρίχνοντας φως στα βιολογικά μονοπάτια μέσω των οποίων οι διατροφικές συνήθειες μπορούν να επηρεάσουν τον κίνδυνο άνοιας

Δημοσίευση 2 «**The goldmine of GWAS summary statistics: a systematic review of methods and tools»:** Έχει υποβληθεί και είναι σε κρίση στο διεθνές επιστημονικό περιοδικό BioData Mining

Το παρόν επιστημονικό άρθρο αποτελεί μια εκτενή συστηματική ανασκόπηση των διαθέσιμων λογισμικών, εργαλείων και βάσεων δεδομένων που χρησιμοποιούνται για την ανάλυση συγκεντρωτικών στατιστικών δεδομένων (summary statistics) προερχόμενων από Μελέτες Γονιδιωματικής Συσχέτισης (GWAS). Καθώς τα δεδομένα αυτά προσφέρουν το πλεονέκτημα της προστασίας της ιδιωτικότητας των ασθενών και ταυτόχρονα απαιτούν πολύ μικρότερο υπολογιστικό κόστος σε σχέση με τα δεδομένα σε επίπεδο μεμονωμένου ατόμου (IPD), έχουν γίνει πλέον απολύτως απαραίτητα στη σύγχρονη γενετική έρευνα. Ο βασικός στόχος των συγγραφέων ήταν να δημιουργήσουν έναν ολοκληρωμένο οδηγό που θα διευκολύνει τους ερευνητές να περιηγηθούν στο πλήθος των διαθέσιμων επιλογών και να επιλέξουν τα πλέον κατάλληλα εργαλεία για τις δικές τους αναλυτικές ανάγκες.

Ακολουθώντας τις διεθνείς οδηγίες PRISMA για τις συστηματικές ανασκοπήσεις, η βιβλιογραφική έρευνα εντόπισε συνολικά 305 λειτουργικά εργαλεία και βάσεις δεδομένων, τα οποία οι ερευνητές κατηγοριοποίησαν σε τρεις μεγάλες ομάδες με βάση τη λειτουργικότητά τους. Η πρώτη κατηγορία αφορά τη διαχείριση των ίδιων των δεδομένων και περιλαμβάνει εργαλεία για ποιοτικό έλεγχο, εναρμόνιση, καταλογισμό (imputation) καθώς και τα δημόσια διαθέσιμα αποθετήρια αποτελεσμάτων. Η δεύτερη κατηγορία αφορά την ανάλυση ενός μεμονωμένου χαρακτηριστικού (single-trait

analysis), περιλαμβάνοντας μεθόδους για μετα-ανάλυση, εκτίμηση κληρονομησιμότητας, αναλύσεις σε επίπεδο γονιδίου (gene-based tests) ή μονοπατιών (gene set analysis), καθώς και μεθόδους χαρτογράφησης ακριβείας (fine-mapping). Η τρίτη κατηγορία εστιάζει στην ταυτόχρονη ανάλυση πολλαπλών χαρακτηριστικών, ενσωματώνοντας εργαλεία για τη μελέτη της πλειοτροπίας, της γενετικής συσχέτισης, της Μενδελικής τυχαιοποίησης (Mendelian randomization), καθώς και αναλύσεις TWAS και συνεγκατάστασης (colocalization).

Μέσα από την αναλυτική παρουσίαση και σύγκριση αυτών των μεθόδων, διαπιστώθηκε ότι η πλειοψηφία των εργαλείων είναι γραμμένα στις γλώσσες προγραμματισμού R (πάνω από 56%) και Python, ενώ ένα αρκετά μικρό ποσοστό (κάτω από 10%) διατίθεται στο ευρύ κοινό υπό τη μορφή εύχρηστων διαδικτυακών εφαρμογών (webservers). Παράλληλα, οι συγγραφείς υπογραμμίζουν ένα σημαντικό και πάγιο πρόβλημα της βιοπληροφορικής κοινότητας, το οποίο εντοπίζεται στην έλλειψη μακροπρόθεσμης συντήρησης των λογισμικών, γεγονός που συχνά οδηγεί σε μη λειτουργικούς συνδέσμους ή ασύμβατες εξαρτήσεις λόγω παλαιότητας. Συμπερασματικά, η μελέτη αυτή αναδεικνύει την τεράστια ανάπτυξη και τον κεντρικό ρόλο της βιοπληροφορικής στη λεγόμενη «μετα-GWAS» εποχή (post-GWAS era), προσφέροντας έναν πολύτιμο πόρο που αναμένεται να μεγιστοποιήσει την αποτελεσματικότητα της ανάλυσης δεδομένων, με απώτερο σκοπό την καλύτερη κατανόηση της γενετικής αρχιτεκτονικής των πολύπλοκων χαρακτηριστικών και ασθενειών.

Metabolic signature of Mediterranean diet adherence and risk of dementia: A prospective analysis in the UK Biobank cohort

Maria Manou, MSc[1,2,3,*], Christos Papagiannopoulos, MSc[1], Eleftherios Pavlos, MSc[2,3], Georgios Markozannes, PhD[1,5], Jordi Julvez, PhD[6,7], Nikolaos Scarmeas, MD, PhD[8,9], Ioanna Tzoulaki, PhD[2,3,5], Konstantinos K. Tsilidis, PhD[1,5], Christopher Papandreou, PhD[6,10,*]

Affiliations

[1]Department of Hygiene and Epidemiology, School of Medicine, University of Ioannina, Ioannina, Greece.

[2] Biomedical Research Foundation of the Academy of Athens, Athens, Greece.

[3] Hellenic Pasteur Institute, Athens, Greece

[4] Division of Basic Sciences, University of Crete Medical School, Heraklion 71110, Greece

[5] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK.

[6] Clinical and Epidemiological Neuroscience (NeuroÈpia), Institut d'Investigació Sanitària Pere Virgili (IISPV), Reus, Spain

[7] ISGlobal, Institut de Salut Global de Barcelona-Campus MAR, PRBB, Barcelona, Spain

[8] 1st Department of Neurology, Aiginition Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece

[9] Department of Neurology, Columbia University, New York, NY, USA

[10]Department of Nutrition and Dietetics Sciences, School of Health Sciences, Hellenic Mediterranean University (HMU), Siteia, Greece

***Corresponding authors:** Maria Manou, M.Sc, PhD candidate, Department of Hygiene and Epidemiology, School of Medicine, University of Ioannina, 45110, Ioannina, Greece. E-mail: m.manou@uoi.gr, Christopher Papandreou, PhD, Department of Nutrition and Dietetic Sciences, School of Health Sciences, Hellenic Mediterranean University, 72300 Sitia, Greece. E-mail: papchris@hmu.gr

ORCID

Maria Manou: https://orcid.org/0000-0002-9584-0373

Christos Papagiannopoulos: https://orcid.org/0000-0003-4030-882X

Georgios Markozannes: https://orcid.org/0000-0001-8481-579X

Jordi Julvez: https://orcid.org/0000-0003-0818-4003

Nikolaos Scarmeas: https://orcid.org/0000-0001-6453-8908

Christopher Papandreou: https://orcid.org/0000-0002-6803-507X

Ioanna Tzoulaki: https://orcid.org/0000-0002-4275-9328

Konstantinos K Tsilidis: https://orcid.org/0000-0002-8452-8472

Abstract word count: 321

Manuscript word count:

References: 75

Figures: 3

Tables: 1

ABSTRACT

Background and Objectives: Adherence to the Mediterranean diet (MD) has been associated with a lower dementia risk, yet the biological mechanisms underlying this association remain poorly understood, and objective biomarkers of MD adherence are limited. This study aimed to identify a plasma metabolic signature of MD adherence and evaluate its association with the risk of dementia and its subtypes.

Methods: We analysed 92,299 UK Biobank participants with available NMR-based metabolomics and dietary data derived from at least one 24-hour dietary recall. MD adherence was assessed using the Mediterranean Diet Adherence Screener (MEDAS). A metabolic signature of MEDAS was identified using a multi-step adaptive elastic-net model and externally validated in the Greek Epirus Health Study (EHS; N = 1,608). Cox proportional hazards regression models were used to examine associations of metabolic signature with incident all-cause dementia, Alzheimer's disease (AD) and vascular dementia (VD), adjusting for potential confounders.

Results: Over a median follow-up of 13.5 years, 975 incident cases of all-cause dementia, 346 AD and 151 VD cases were identified. A 26-metabolite signature was identified (r = 0.27, 95% CI: 0.26 - 0.28) and externally validated in EHS (r = 0.22, 95% CI: 0.17 - 0.26). The signature was predominantly composed of lipid and fatty acid–related metabolites, along with selected amino acids and energy metabolism markers, and was characterized by lower levels of glycoprotein acetylation (GlycA), a marker of systemic inflammation. The metabolic signature was inversely associated with all-cause dementia (HR per SD = 0.91, 95% CI: 0.84 - 0.97), while its association with AD was attenuated after multivariable adjustment. Metabolic signature was not significantly associated with VD.

Discussion: The plasma metabolic signature of MD was associated with a reduced risk of dementia in a large-scale population study. These findings support the use of metabolomic profiling as a complementary approach to dietary assessment, providing objective biological insight into diet-related metabolic pathways through which dietary patterns may influence dementia risk.

Keywords: Mediterranean diet, metabolomics, NMR, metabolic signature, UK Biobank

INTRODUCTION

Dementia is a leading cause of disability and dependency among older adults worldwide, affecting over 55 million individuals in 2023, with numbers projected to triple by 2050 [1]. Despite extensive research, effective strategies for dementia prevention remain limited, and the need for modifiable interventions is critical [2]. The Mediterranean diet (MD), characterized by high consumption of fruits, vegetables, whole grains, legumes, nuts, olive oil, and moderate intake of fish and poultry, has emerged as a promising dietary pattern for dementia [3, 4]. Epidemiological studies have demonstrated that greater adherence to the MD is associated with reduced risk of dementia [5, 6]. Notably, adherence to the MD has been shown to lower dementia risk independently of genetic predisposition, as evidenced by findings from the UK Biobank prospective cohort study [7]. A recent meta-analysis of 23 studies confirms that adherence to the MD is associated with an 11-30% reduction in the risk of age-related cognitive disorders, including cognitive impairment, dementia, and Alzheimer's disease (AD), highlighting the diet's potential as a key neuroprotective strategy in public health [8].

Previous studies have identified specific metabolites linked to MD adherence and dementia, such as lipids, polyphenol-derived compounds, and markers of inflammation and oxidative stress [9, 10]. However, many of these studies are limited by small sample sizes, and lack of external validation, which hinder the generalizability of their findings [11]. Additionally, traditional dietary assessment tools, such as food frequency questionnaires (FFQs) and dietary recalls, rely on self-reported intake data, which are prone to measurement errors [12, 13, 14], further complicating dietary research.

Several mechanisms have been proposed to explain the neuroprotective effects of the MD, including its anti-inflammatory and antioxidant properties, cardiovascular benefits, and modulation of gut-brain axis interactions [15, 16]. Preclinical studies in nonhuman primates have further shown that MD consumption reduces pro-inflammatory gene expression in both peripheral monocytes and brain

tissue, which is associated with preserved brain structure, such as cortical thickness and brain volume, and improved socioemotional behaviours. These findings provide mechanistic insights linking diet, peripheral and central inflammation, and neuroprotection [17]. However, the pathways linking MD adherence to dementia risk are not well characterized. Recent advances in metabolomics technology have facilitated the identification of metabolic signatures, comprehensive profiles of small molecules in biological systems, that reflect dietary patterns and their associated physiological effects [18, 19]. These metabolomic analyses have begun to uncover the biological underpinnings of the MD's benefits and offer new opportunities for understanding its role in dementia prevention.

In this study, we leveraged plasma metabolites from the large-scale UK Biobank cohort to identify a metabolic signature associated with MD adherence. We validated our findings in an independent cohort to enhance their robustness. We then assessed whether this signature was associated with risk of dementia, AD-and vascular dementia (VD) in UK Biobank. Because dietary recalls in the UK Biobank were collected several years after baseline blood sampling, diet cannot be used as a strict baseline exposure. Metabolomic profiling, however, offers a temporally valid approach to capture biological pathways related to diet at baseline, thereby enabling prospective investigation of diet-related metabolic signatures in relation to incident dementia.

METHODS

Study population

Discovery cohort: In the UK Biobank cohort (**Supplementary Text 1**), individuals with available metabolomic and dietary data (N = 97,516) were included in the analysis. Participants with baseline diagnoses of dementia, migraine, epilepsy, schizophrenia, bipolar disorder, or Parkinson’s disease were excluded (N = 5,207). These conditions were identified using self-reported data and linked hospital records coded with the following international classification of diseases (ICD): ICD-10 codes G20, G40, G41, G43, F20, and F31; and ICD-9 codes 332.1, 345.0–345.9, 346.9, and 295.0–295.9. There were no mismatches between genetic sex and self-reported sex among individuals.

Additionally, prevalent cases of dementia (N = 10) were excluded from the analysis (**Figure 1**). After applying these criteria, a total of 92,299 participants were included in the final analyses.

Validation cohort: We used baseline data (N = 1,612) from the Epirus Health Study (EHS) and employed the same exclusion criteria as the discovery population in the participant selection process (**Supplementary Text 1**, **Supplementary Figure 1**). Finally, 1,608 participants were included in the validation analysis.

Metabolomic profiling

A high-throughput Nuclear Magnetic Resonance (NMR) platform (Nightingale Health Plc; biomarker quantification version 2022) has been used to assess a random subset of non-fasting baseline plasma samples from 274,353 UK Biobank people, recruited between 2006 and 2010. An array of 168 metabolites including lipids, lipoprotein particle subclass, cholesterol subtypes, amino acids and inflammation markers, were quantitatively profiled (molar concentration units), and 81 ratios derived from their combinations were further included [20, 21]. Additional information about metabolomic biomarker measurements can be found in **Supplementary Text 2**. The same metabolomic profiling methodology used in the UK Biobank was also used to assess fasting plasma samples from 1,608 persons in the EHS. **Supplementary Table 1** includes the concentrations of metabolic markers for the participants in the two distinct studies, together with the median and interquartile ranges.

Dietary assessment and calculation of MEDAS

Validated for use in large-scale observational studies, the Oxford WebQ is a web-based, self-administered 24-hour dietary assessment tool [22, 23]. By having users pick the number of standard portions they ingested for each meal and drink item, this tool gathers data on the intake of 206 different food types and 32 different drink types over a 24-hour period. During their visit to the baseline assessment centre, individuals who were recruited between April 2009 and September 2010 completed the Oxford WebQ. Furthermore, via their home computer, participants were asked to complete the Oxford WebQ assessment every three to four months between February 2011 and June

2012, for a maximum of five assessments (including the baseline assessment). We excluded participants with 24-h recalls with extreme energy intakes (defined as < 800 or > 4,200 kcal/d for males and < 600 or > 3,500 kcal/d for females). We used the residuals method to energy-adjust the dietary data (2,000 kcal/d) for each time point in accordance with earlier studies [7, 24]. This allowed for the evaluation of food quality to be done independently of diet amount [25]. Before determining each participant's diet score, data were then averaged over all possible time points (minimum 1, maximum 5). We measured the MEDAS score, which is a 14-point score that was created as part of the Prevención con Dieta Mediterránea (PREDIMED) trial [26] and has been used extensively in observational studies and trials [9, 27, 28]. The MEDAS is typically calculated using a binary evaluation for each of the 14 food components, awarding one point if the participant's consumption meets a pre-established cut-off (e.g., intake of a certain amount of vegetables), and zero points if they do not. The total possible score can range from 0 to 14 points (**Supplementary Table 2**). Points are awarded by MEDAS for using olive oil as the primary cooking fat and, separately, for consuming a certain quantity of olive oil (four tablespoons or more per day). The maximum possible score for the MEDAS in this study was 13 points because it was not possible to determine the amount of olive oil consumed from the available dietary data, even though we were able to determine the use of olive oil as a culinary fat and award points for consumption (1 point) or non-consumption (0 points) accordingly. Participants with at least one 24-hour recall were included in the main analyses. Because dietary data were obtained after baseline assessment (median = 1.9 years, Interquartile range (IQR) = 0 - 2.8 years), dietary variables were not treated as baseline exposures but were used for the derivation of the diet-related metabolomic signature.

Dementia Ascertainment

In the UK Biobank sample, all-cause dementia status was identified based on hospital inpatient records and death registry records. The ICD was utilized to identify cases with all-cause dementia, AD and vascular dementia (VD), with the list of ICD-9 and ICD-10 codes demonstrated in the **Supplementary Table 3**. ICD-9 codes were included to improve our sensitivity in identifying

dementia cases that were confirmed before enrolment. An individual was considered to have dementia if they had either received a primary or secondary diagnosis of dementia (primary care/ hospital records) or if their primary or secondary cause of death was dementia related. In cases where there were several dates available for the variables, the date of diagnosis was established using the earliest of two dates: the date of death obtained from primary care records, or the date of a primary care diagnosis. In the UK Biobank, the censoring date was 2022-09-27 which was the date until which diagnoses were available. Death records were used to obtain the date of death. For individuals with dementia, time at risk was calculated as the time in days between their baseline assessment and the earliest dementia diagnosis. For individuals with a death record and no diagnosis of dementia, time at risk was computed as the time between their baseline assessment and the date of their death record. The time between a person's baseline evaluation and the censoring date was used to calculate time for those who were still alive and free of dementia.

Covariates assessment

In the UK Biobank, the Townsend deprivation index (TDI), a composite measure of deprivation based on unemployment, non-ownership of a car, non-ownership of a home, and household overcrowding, was estimated using data from the previous national census [29]. Body Mass Index (BMI) was calculated using weight and height recorded during the recruitment. Information about smoking status (never/previous/current smoker) and education status (higher/lower) was collected through touch-screen questionnaires. Chronic conditions were assessed using a combination of self-reported data and linked hospital records coded with ICD-10 and ICD-9 classifications. Participants were classified as having diabetes if they had glycated haemoglobin (HbA1c) levels exceeding 48 mmol/mol, were on antidiabetic medications (**Supplementary Table 4**), or had self-reported or previously diagnosed diabetes. ICD-10 and ICD-9 codes (E10, E11, 250.0) from linked hospital records were also used to identify diabetes cases. Cancer status and depression status at baseline were determined using both self-reported data and linked hospital records coded with ICD-10 and ICD-9 codes (C00-C97, F32, F33, 311.9). Cardiovascular disease (CVD) status included any history of ischemic heart disease,

myocardial infarction, stroke, heart failure, or other vascular conditions. Both ICD-10 codes (I516, I519, I20-I25, I60-I69, I110-I139, I500-I509, I420-I429, I700-I739, G450-G459) and ICD-9 codes (4100-4109, 4110-4119, 4120-4129, 4140-4149, 4340-4349, 4350-4359, 4300-4309, 4280-4289) were used to identify these conditions. APOE4 carrier status was determined using genotypic data for the APOE rs429358 and rs7412 variants. Participants were classified as APOE4 carriers if they had at least one copy of the ε4 allele, defined as having one or two C alleles at rs429358 and no T alleles at rs7412. Participants who answered 'do not know' or 'prefer not to answer' to any of the self-reported questions were considered missing values. Data field identifiers for all utilized features are described in **Supplementary Table 5**.

Statistical analysis

Baseline characteristics of participants were described using percentages for categorical categories and mean and standard deviation (SD) for continuous variables. In the discovery dataset, no metabolite had more than 20% missing values (**Supplementary Figure 2A**). Participants with up to 5% missing values in their metabolite data were excluded from the analysis, as a result 26,065 participants were removed because they had at least one missing metabolite value. Due to the high percentage of missing values (>20%), two of the 250 plasma metabolites, glycerol and beta hydroxybutyrate, were eliminated from the external validation set. The remaining metabolites' missing values (**Supplementary Figure 2B**) were imputed, by using MissForest algorithm [30]. In the discovery population, metabolites were inverse-normal transformed and analysed using multi-step adaptive elastic-net (MSA-Enet) with stability selection. Model tuning was performed using cross-validation. Full details are provided in **Supplementary Text 3.**

Associations between the metabolic signature with dementia

In the UK Biobank, the metabolic signature was converted to a z-score (mean = 0; SD = 1) before Cox regression analyses. The time-to-event variable for our analyses of the incidence of all-cause dementia and AD was the difference between the date of enrolment and the event, death, or end of

follow-up, whichever occurred first. We evaluated the proportional hazards assumption by examining the relationship between time and standardized Schoenfeld residuals, and we did not find any evidence suggesting it was violated. To handle missing data in covariates, we performed multiple imputation using multivariate imputation by chained equations (MICE) with the random forest method, generating five imputed datasets [31]. To investigate the relationships between the standardized metabolic signature and the risk of all-cause dementia, AD and VD, we fitted two multivariable Cox proportional hazards models. The first model was subjected to age and sex adjustments. The second model (multivariable-adjusted model) was further adjusted for potential confounders, identified via Directed Acyclic Graphs (DAGs) (**Supplementary Figure 3**), such as BMI, TDI, diabetes, smoking status, education level, history of depression, cancer and cardiovascular disease and APOE4 status. Effect modification by age, sex, BMI, TDI, educational level, smoking status, and APOE4 status was assessed by adding a multiplicative interaction term between each modifier and the exposure (metabolic signature) in multivariable Cox models and comparing models with and without the interaction using likelihood ratio tests. Resulting p-values were adjusted for multiple testing with the Benjamini–Hochberg false discovery rate (FDR) procedure [32].

Several sensitivity analyses were conducted to assess the robustness of the findings. First, to better capture habitual dietary intake, analyses were restricted to participants who completed at least two 24-hour diet recalls [22]. Second, we repeated the original analyses after removing cases that occurred during the first one and five years of follow-up, respectively, to account for the possibility of reverse causality.

R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria) was used for all analyses. Statistical significance was defined as a two-tailed P value <0.05. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The Strengthening the Reporting of Observational studies in Epidemiology reporting guidelines were followed.

Standard Protocol Approvals, Registrations, and Patient Consents

The UK Biobank research received ethical approval from the Northwest Multicenter Research Ethics Committee in the United Kingdom, as well as authorization from the National Information Governance Board for Health and Social Care in England and Wales, and the Community Health Index Advisory Group in Scotland. Every participant gave written, informed consent, and the study was carried out in compliance with the Helsinki Declaration.

Data availability

Data in the manuscript, code book, and analytic code will not be publicly available due to sensitivity concerns but may be obtained from the corresponding author upon reasonable request. The UK Biobank dataset used to conduct the research in this paper is available via application directly to the UK Biobank. Applications are assessed to meet the required access criteria, including legal and ethical standards. More information regarding data access can be found here: access@ukbiobank.ac.uk

RESULTS

Baseline characteristics

The mean age was 56·3 years (SD = 7·9) in the UK Biobank and 46·8 years (SD = 11·1) in the EHS (**Table 1**, **Supplementary Table 6**). In UK Biobank, during a median follow-up period of 13·5 years (IQR = 12·8 - 14·2), a total of 975 incident all-cause dementia events, 346 AD incident events, and 151 incident vascular dementia occurred. Participants who developed dementia were older at baseline (mean age 64.3 vs. 56.2 years), more likely to be male (56.4% vs. 47.4%), and had slightly higher BMI (27.4 vs. 27.1 kg/m^2). They also had higher prevalence of cardiovascular disease (17.5% vs. 0.9%), diabetes (11.7% vs. 0.9%), depression (7.1% vs. 1.0%), and cancer (12.2% vs. 0.9%). Dementia cases were more likely to carry the APOE4 allele (50.9% vs. 25.4%) and had lower levels of education (52.6% vs. 39.9%). The mean MEDAS score was lower among those who developed

dementia compared to those who did not (3.8 vs. 4.2). Differences in socioeconomic deprivation and smoking status were modest across groups.

Metabolic signature of MEDAS

A total of 26 metabolites and metabolites' ratios after 10-CV (RMSE = 0.963), were selected from MSA-Enet to be associated with the MEDAS score. Of these, 13 metabolites were positively associated with MEDAS, while 13 were inversely associated (**Figure 2, Supplementary Table 7**). Positively associated metabolites included linoleic acid to total fatty acids, docosahexaenoic acid (DHA) to total fatty acids, omega-3 fatty acids, free cholesterol to total lipids ratio in very large VLDL, triglycerides in LDL, valine, phospholipids in medium HDL, albumin, phospholipids in very large VLDL, beta-hydroxybutyrate, acetate, acetone, and saturated fatty acids to total fatty acids. Inversely associated metabolites included omega-6 fatty acids to total fatty acids, cholesteryl esters to total lipids ratio in very large VLDL, omega-6 fatty acids, Glycoprotein acetylation (GlycA), degree of unsaturation, omega-6 to omega-3 fatty acid ratio, creatinine, phospholipids to total lipids ratio in very small VLDL, triglycerides to total lipids ratio in large HDL, lactate, glutamine, isoleucine, and polyunsaturated to monounsaturated fatty acid ratio. These metabolites and metabolites' ratios explained 7.3 % of the total variance of MEDAS. The derived metabolic signature was significantly correlated with MEDAS [Pearson correlation coefficient [r = 0.27 (95% CI, 0.26 - 0.28]). In the external validation set (EHS), the metabolites included in the metabolic signature accounted for 4.8 % (RMSE = 0.976). The external validation set showed similar magnitudes of correlation for MEDAS (r = 0.22, 95% CI, 0.17 - 0.26).

Associations of metabolic signature with all-cause dementia, AD and VD

Higher metabolic signature was associated with lower risk of all-cause dementia and AD (**Figure 3**, **Supplementary Table 8**). In age- and sex-adjusted models, the metabolic signature was associated with a 13% lower risk (HR per SD increment = 0.87, 95% CI: 0.81 – 0.93). These associations remained significant after multivariable adjustment for BMI, TDI, diabetes, smoking status,

education, depression, cancer, cardiovascular disease, and APOE4 status (adjusted HR per SD increment = 0.91, 95% CI: 0.84 – 0.97). For AD, a similar pattern of associations was observed. In age- and sex-adjusted models, the metabolic signature was associated with an 11% lower risk (HR per SD increment = 0.89, 95% CI: 0.80 – 0.99). After multivariable adjustment, the association for the metabolic signature was attenuated and no longer significant (adjusted HR per SD increment = 0.89, 95% CI: 0.79 – 1.01). For VD, inverse associations were observed in age- and sex-adjusted models, with each SD increment in the metabolic signature associated with a 19% lower risk (HR = 0.81, 95% CI: 0.68 – 0.95). However, after multivariable adjustment, these associations were attenuated and no longer statistically significant (HR = 0.94, 95% CI: 0.79 – 1.13).

Stratified and sensitivity analysis

After FDR correction, the interaction test indicated heterogeneity by TDI for the metabolic signature. Associations were more protective, for the metabolic signature, among more deprived participants (TDI ≥ −2.43: HR = 0.84, 95% CI: 0.77 − 0.93 vs TDI < −2.43: HR = 0.98, 95% CI: 0.88 − 1.08). Interactions by age, sex, BMI, education, smoking, and APOE4 status were not significant after FDR (**Supplementary Table 9**). In sensitivity analyses (**Supplementary Table 10**), results were consistent across alternative model specifications. When restricting to participants with at least two dietary assessments (N = 54,693; Sensitivity Analysis 1), the inverse associations of metabolic signature with dementia risk were directionally consistent, though slightly attenuated. Excluding dementia cases diagnosed within the first year of follow-up (Sensitivity Analysis 2; N = 92,297) yielded estimates nearly identical to the main analyses. Similarly, excluding cases within the first five years of follow-up (Sensitivity Analysis 3; N = 92,249) produced comparable results.

DISCUSSION

By applying machine learning approaches, we developed and externally validated a plasma metabolic signature of MEDAS. This signature, identified in the UK Biobank and validated in the independent EHS, was associated with a lower risk of all-cause dementia. Importantly, metabolic score was associated with dementia risk, supporting the relevance of MD adherence as assessed through 24-hour dietary recalls and objective metabolomic measures. These findings support the potential value of metabolomic profiling in capturing diet-related metabolic pathways and in providing objective measures of dietary adherence relevant to dementia prevention.

Our findings are strongly supported by numerous previous studies demonstrating the protective effects of the MD on cognitive aging and dementia prevention. Higher adherence to the MD is linked to a significant reduction in the risk of cognitive impairment, dementia, and AD, according to a recent meta-analysis [8] of 23 studies, underscoring its potential as a population-level neuroprotection strategy. Also, systematic reviews confirm that greater MD adherence correlates with lower incidence of cognitive decline and dementia, supported by robust epidemiological data [33, 34]. The abundance of anti-inflammatory, antioxidant, and neurotrophic nutrients found in the MD, such as vitamins, polyphenols, and omega-3 fatty acids, has been linked to its neuroprotective effects [35]. These advantages are further supported by neuroimaging research, which shows that following a MD is linked to a decrease in white matter lesions and hippocampal atrophy, two important indicators of neurodegeneration [36]. Although related dietary patterns such as Dietary Approaches to Stop Hypertension (DASH) and Mediterranean-DASH Intervention for Neurodegenerative Delay (MIND) also confer cognitive benefits, the MD remains the most consistently associated with favorable trajectories in cognitive performance and brain structural preservation [37, 38]. A recent study highlighted a metabolic signature of MD adherence related to reduced neuroinflammation, improved lipid metabolism, and enhanced energy regulation [39], which was associated with improved cognitive outcomes. Overall, this growing body of evidence establishes the MD as a biologically plausible approach to mitigating dementia risk.

The validity of MEDAS in capturing important nutritional exposures was reinforced by its strong association with a metabolic profile typical of MD adherence in our large cohort study. The close alignment between MEDAS and several lipid- and fatty-acid related metabolites supports its ability to capture habitual dietary patterns with known biological relevance. The positive correlations of omega-3 fatty acids and, especially linoleic acid and DHA, and negative correlations of the ratio omega-6 to omega-3 with MEDAS score were expected given the dietary origins of these fatty acids and the high polyunsaturated fatty acid content of the MD [10]. Regular intake of omega-3 fatty acids may help maintain adequate and sustained circulating levels, potentially contributing to a lower risk of dementia [40], likely through their roles in neuronal membrane fluidity, anti-inflammatory signalling, and synaptic function [41]. Elevated acetate in those participants more adhered to the MD reflects enhanced microbial fermentation of dietary fiber, which is abundant in fruits, vegetables, legumes, and nuts, all components of the MEDAS score [42]. This finding is in line with research showing that Mediterranean-style diets encourage beneficial changes in the composition of the gut microbiota and the production of short-chain fatty acids, both of which improve the integrity of the gut–brain barrier and reduce neuroinflammation [43]. Higher MEDAS scores are also associated with higher levels of acetone and β-hydroxybutyrate, suggesting a metabolic milieu characterized by greater metabolic flexibility and availability of alternative brain energy substrates, which may promote cognitive resilience [44].

We also observed several lipoprotein-related metabolites associated with MEDAS. Positive associations with phospholipids in HDL, triglycerides in LDL, and free cholesterol in very large VLDL indicate improved lipid transport dynamics and favourable lipoprotein remodelling associated with cardiometabolic health. [45, 46]. The inverse associations with cholesteryl esters in very large VLDL and phospholipids in very small VLDL suggest reduced atherogenic lipoprotein burden and enhanced HDL functionality, mechanisms that support neurovascular health and amyloid-β clearance [47-49]. Together, these findings are consistent with extensive evidence that adherence to the MD

promotes a shift toward less atherogenic, more functional lipid particles, contributing to both cardiovascular and cognitive benefits.

Besides fatty acids and lipoproteins, several amino acids and energy-related metabolites were differentially associated with MEDAS, providing further insight into diet-related metabolic adaptations. Lactate, glutamine, and isoleucine were inversely correlated with MEDAS, while valine was positively associated, potentially reflecting protein sources characteristic of the MD, such as legumes and fish, and supporting nutrient-sensing and energy-regulatory pathways. Reduced lactate aligns with improved mitochondrial oxidative function, while lower glutamine reveals reduced systemic inflammation. The inverse association with isoleucine is consistent with studies linking lower circulating branched-chain amino acids to improved metabolic health [50, 51]. GlycA, a composite marker of systemic inflammation, was inversely associated with MEDAS. Although inflammatory biomarkers were not the dominant component of the signature, this finding aligns with the well-documented anti-inflammatory properties of the MD [52]. Lower creatinine levels associated with higher MEDAS may indicate improved renal function or lower muscle catabolism, consistent with evidence that the MD supports kidney health through reductions in oxidative stress and improved metabolic control [53, 54]. Additionally, higher circulating albumin, suggests better nutritional and inflammatory status, reinforcing the role of the MD in supporting metabolic homeostasis [55].

Our findings for fatty acid profiles were generally in agreement with previous literature, though with some notable distinctions. The inverse relationships with omega-6 fatty acids, omega-6 to omega-3 ratio, polyunsaturated to monounsaturated fatty acids ratio, and degree of unsaturation are consistent with the well-characterized features of the MD, especially its focus on lowering pro-inflammatory omega-6 fatty acids while raising foods high in omega-3 and MUFA. While we found a positive association between saturated fatty acids (SFA) and MEDAS, this could be due to dietary or cultural variations within the UK Biobank population, where moderate SFA intake may coexist with Mediterranean-style foods or reflect specific NMR-derived SFA measures [56, 57]. This variation

highlights the complexity of interpreting fatty acid profiles across populations with different dietary contexts.

Furthermore, diet-related metabolic resilience may partially offset structural disadvantages, potentially by mitigating cardiometabolic stress and inflammation in high-risk populations [34, 58]. This mirrors Foster et al.'s UK Biobank observations that socioeconomic deprivation amplifies the harmful effects of unhealthy lifestyles, while demonstrating a symmetric benefit for protective exposures such as healthy dietary patterns [59]. These results highlight the potential for dietary interventions to yield disproportionate benefits in vulnerable populations.

In the UK Biobank, the metabolomic signature may serve as a potent biological proxy for dietary impact, demonstrating a significant association with incident dementia risk. By reflecting downstream physiological manifestations of food intake, the metabolic score may capture aspects of biological response that complement self-reported measures. Although derived from an NMR-based platform that focuses heavily on lipid metabolism and specific small molecules, the signature successfully identifies key metabolic shifts relevant to cognitive decline. This association remains robust despite the effect from non-dietary factors like adiposity and medication use, suggesting that the score isolates a core biological signal relevant to brain health. Furthermore, because the metabolic signature reflects real-time biological status, its strong correlation with dementia highlights the immediate relevance of metabolic homeostasis in neuroprotection. Future transitions toward mass spectrometry-based approaches are expected to further refine this association by capturing an even broader array of diet-responsive pathways.

This study has several important strengths. It leveraged a large, well-characterized population cohort with long-term follow-up, enabling robust assessment of incident dementia and AD. The discovery and replication of the metabolic signature across two distinct populations, the UK Biobank and the Greek Epirus Health Study, demonstrate cross-population reproducibility and enhance the external validity of our findings. The combination of comprehensive plasma metabolomic profiling and

advanced machine-learning approaches provided objective and biologically meaningful insights into diet-related pathways. However, certain limitations should be considered. Although repeated 24-hour dietary recalls [22] were available for many participants, reliance on self-reported dietary data may have introduced some imprecision in estimating habitual intake. A methodological consideration is that dietary recalls were collected several years after baseline blood sampling and therefore cannot represent a strict baseline exposure. However, this design underscores the utility of baseline metabolomic profiling, which captures biological signals linked to long-term dietary patterns without being affected by the post-baseline timing of the dietary recalls. This strengthens the interpretation of the metabolomic signature as a temporally valid marker for prospective dementia risk. The NMR metabolomics platform provides narrower coverage than mass spectrometry [60] and is largely focused on lipid-related metabolites, potentially limiting the identification of other diet-responsive compounds. Moreover, the use of non-fasting plasma samples in the UK Biobank could have introduced short-term variability in metabolite levels, though the replication of findings in fasting samples from the Epirus Health Study strengthens confidence in their validity. The predominantly White British and non-Mediterranean composition of the UK Biobank cohort may also limit the generalizability of our findings to more ethnically diverse populations. Finally, as an observational analysis, causal inference cannot be established, and residual confounding cannot be excluded despite extensive adjustments.

CONCLUSION

In summary, using NMR-based metabolomic profiling, we identified and externally validated a 26-metabolic plasma signature associated with adherence to the MD. This metabolic signature was inversely associated with the risk of all-cause dementia and reflected a favourable metabolic phenotype for brain health, characterized by higher levels of omega-3 and other unsaturated fatty acids. These findings provide biological insight into diet-related metabolic pathways that may contribute to cognitive resilience and illustrate the potential of metabolomic profiling to complement

traditional dietary assessment in the development of dementia prevention strategies and personalized nutrition approaches.

Funding statement

This study has been funded by the Alzheimer's Association, through the project AARG-NTF-23 1027318 & the action "Flagship actions in interdisciplinary scientific areas with special interest for the connection with the productive fabric", Greece 2.0 - National Recovery and Resilience Plan (TAEDR-0539180).

Role of the funding source

The study funding had no role in the study design, data collection, data analysis, and manuscript preparation and/or publication decisions. The corresponding authors had full access to all the data in the study and had final responsibility for the decisions to submit for publication.

Ethics statement

The analysis was performed under the UK Biobank application number 79696. The UK Biobank has approval from the Northwest Multi-centre Research Ethics Committee. It has also sought approval from the Patient Information Advisory Group in England and Wales to access information that would allow it to invite potential participants.

Disclosure

The authors declare no conflict of interest.

Author Contributors

Conceptualization: C.P., Data acquisition: M.M., Christos Papagiannopoulos, Statistical analyses: M.M., Christos Papagiannopoulos, C.P., Findings interpretation: C.P., I.T., K.K.T., G.M., N.S., J.J., E.P., Writing: M.M., Christos Papagiannopoulos, C.P., Critical revision of the manuscript for

important intellectual content: All authors. None of the authors was involved in the patient data collection.

Supplemental Material

Supplementary Methods

Supplementary Tables 1-10

Supplementary Figures 1-3

REFERENCES

1. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health. 2022 Feb;7(2):e105-e125. doi: 10.1016/S2468-2667(21)00249-8. Epub 2022 Jan 6. PMID: 34998485; PMCID: PMC8810394.
2. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, Brayne C, Burns A, Cohen-Mansfield J, Cooper C, Costafreda SG, Dias A, Fox N, Gitlin LN, Howard R, Kales HC, Kivimäki M, Larson EB, Ogunniyi A, Orgeta V, Ritchie K, Rockwood K, Sampson EL, Samus Q, Schneider LS, Selbæk G, Teri L, Mukadam N. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020 Aug 8;396(10248):413-446. doi: 10.1016/S0140-6736(20)30367-6. Epub 2020 Jul 30. Erratum in: Lancet. 2023 Sep 30;402(10408):1132. doi: 10.1016/S0140-6736(23)02043-3. PMID: 32738937; PMCID: PMC7392084.
3. Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a Mediterranean diet and survival in a Greek population. N Engl J Med. 2003 Jun 26;348(26):2599-608. doi: 10.1056/NEJMoa025039. PMID: 12826634.

4. Martínez-González, Miguel Ángel et al. Cohort profile: design and methods of the PREDIMED study. International journal of epidemiology vol. 41,2 (2012): 377-85. doi:10.1093/ije/dyq250
5. Scarmeas N, Stern Y, Tang MX, Mayeux R, Luchsinger JA. Mediterranean diet and risk for Alzheimer's disease. Ann Neurol. 2006 Jun;59(6):912-21. doi: 10.1002/ana.20854. PMID: 16622828; PMCID: PMC3024594.
6. Singh B, Parsaik AK, Mielke MM, Erwin PJ, Knopman DS, Petersen RC, Roberts RO. Association of mediterranean diet with mild cognitive impairment and Alzheimer's disease: a systematic review and meta-analysis. J Alzheimers Dis. 2014;39(2):271-82. doi: 10.3233/JAD-130830. PMID: 24164735; PMCID: PMC3946820.
7. Shannon OM, Ranson JM, Gregory S, Macpherson H, Milte C, Lentjes M, Mulligan A, McEvoy C, Griffiths A, Matu J, Hill TR, Adamson A, Siervo M, Minihane AM, Muniz-Tererra G, Ritchie C, Mathers JC, Llewellyn DJ, Stevenson E. Mediterranean diet adherence is associated with lower dementia risk, independent of genetic predisposition: findings from the UK Biobank prospective cohort study. BMC Med. 2023 Mar 14;21(1):81. doi: 10.1186/s12916-023-02772-3. PMID: 36915130; PMCID: PMC10012551.
8. Fekete, Mónika et al. The role of the Mediterranean diet in reducing the risk of cognitive impairment, dementia, and Alzheimer's disease: a meta-analysis. GeroScience vol. 47,3 (2025): 3111-3130. doi:10.1007/s11357-024-01488-3
9. Papandreou C, Papagiannopoulos C, Koutsonida M, Kanellopoulou A, Markozannes G, Polychronidis G, Tzakos AG, Fragkiadakis GA, Evangelou E, Ntzani E, Tzoulaki I, Aretouli E, Tsilidis KK. Mediterranean diet related metabolite profiles and cognitive performance. Clin Nutr. 2023 Feb;42(2):173-181. doi: 10.1016/j.clnu.2022.12.012. Epub 2022 Dec 22. PMID: 36599272.
10. Li J, Guasch-Ferré M, Chung W, Ruiz-Canela M, Toledo E, Corella D, Bhupathiraju SN, Tobias DK, Tabung FK, Hu J, Zhao T, Turman C, Feng YA, Clish CB, Mucci L, Eliassen AH, Costenbader KH, Karlson EW, Wolpin BM, Ascherio A, Rimm EB, Manson JE, Qi L,

Martínez-González MÁ, Salas-Salvadó J, Hu FB, Liang L. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk. Eur Heart J. 2020 Jul 21;41(28):2645-2656. doi: 10.1093/eurheartj/ehaa209. PMID: 32406924; PMCID: PMC7377580.

11. Brennan L, Hu FB. Metabolomics-Based Dietary Biomarkers in Nutritional Epidemiology-Current Status and Future Opportunities. Mol Nutr Food Res. 2019 Jan;63(1):e1701064. doi: 10.1002/mnfr.201701064. Epub 2018 May 28. PMID: 29688616.
12. Hu FB, Satija A, Rimm EB, Spiegelman D, Sampson L, Rosner B, Camargo CA Jr, Stampfer M, Willett WC. Diet assessment methods in the nurses' health studies and contribution to evidence-based nutritional policies and guidelines. Am J Public Health 2016;106:1567–1572.
13. Perez-Cornago, Aurora et al. "Description of the updated nutrition calculation of the Oxford WebQ questionnaire and comparison with the previous version among 207,144 participants in UK Biobank." European journal of nutrition vol. 60,7 (2021): 4019-4030. doi:10.1007/s00394-021-02558-4
14. Bailey, Regan L. "Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies." Current opinion in biotechnology vol. 70 (2021): 91-96. doi:10.1016/j.copbio.2021.02.007
15. Lourida I, Soni M, Thompson-Coon J, Purandare N, Lang IA, Ukoumunne OC, Llewellyn DJ. Mediterranean diet, cognitive function, and dementia: a systematic review. Epidemiology. 2013 Jul;24(4):479-89. doi: 10.1097/EDE.0b013e3182944410. PMID: 23680940.
16. Knight A, Bryan J, Murphy K. The Mediterranean diet and age-related cognitive functioning: A systematic review of study findings and neuropsychological assessment methodology. Nutr Neurosci. 2017 Oct;20(8):449-468. doi: 10.1080/1028415X.2016.1183341. Epub 2016 May 18. PMID: 27192034.
17. Frye, Brett M et al. "Mediterranean diet protects against a neuroinflammatory cortical transcriptome: Associations with brain volumetrics, peripheral inflammation, social isolation,

and anxiety in nonhuman primates (Macaca fascicularis).” Brain, behavior, and immunity vol. 119 (2024): 681-692. doi:10.1016/j.bbi.2024.04.016

18. Tanaka T, Talegawkar SA, Jin Y, Candia J, Tian Q, Moaddel R, Simonsick EM, Ferrucci L. Metabolomic Profile of Different Dietary Patterns and Their Association with Frailty Index in Community-Dwelling Older Men and Women. Nutrients. 2022 May 27;14(11):2237. doi: 10.3390/nu14112237. PMID: 35684039; PMCID: PMC9182888.
19. Manou M, Papagiannopoulos C, Chalitsios CV, Asimakopoulos AG, Markozannes G, Bulló M, Tsilidis KK, Papandreou C, Tzoulaki I. Metabolic Signatures of Blood Pressure and Risk of Cardiovascular Diseases. J Am Heart Assoc. 2024 Dec 3;13(23):e036573. doi: 10.1161/JAHA.124.036573. Epub 2024 Nov 22. PMID: 39575750; PMCID: PMC11681602.
20. Soininen, P., Kangas, A. J., Wurtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. CirculationCardiovascular Genetics, 8, 192–206.
21. Wurtz, P., Kangas, A. J., Soininen, P., Lawlor, D. A., Davey Smith, G., & Ala-Korpela, M. (2017). Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: A primer on –Omic technologies. American Journal of Epidemiology, 186, 1084–1096.
22. Greenwood DC, Hardie LJ, Frost GS, Alwan NA, Bradbury KE, Carter M, et al. Validation of the Oxford WebQ Online 24-Hour Dietary Questionnaire Using Biomarkers. Am J Epidemiol. 2019;188:1858–67.
23. Liu B, Young H, Crowe FL, Benson VS, Spencer EA, Key TJ, et al. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. Public Health Nutr. 2011;14:1998–2005.

24. Tong TYN, Wareham NJ, Khaw K-T, Imamura F, Forouhi NG. Prospective association of the Mediterranean diet with cardiovascular disease incidence and mortality and its population impact in a non-Mediterranean population: the EPIC-Norfolk study. BMC Med. 2016;14:135.

25. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. Am J Clin Nutr. 1997;65(4 Suppl):1220S-1228S (discussion 1229S-1231S).

26. Martínez-González MÁ, Corella D, Salas-Salvadó J, Ros E, Covas MI, Fiol M, et al. Cohort profile: design and methods of the PREDIMED study. Int J Epidemiol. 2012;41:377–85.

27. Siervo M, Shannon OM, Llewellyn DJ, Stephan BC, Fontana L. Mediterranean diet and cognitive function: From methodology to mechanisms of action. Free Radic Biol Med. 2021;176:105–17.

28. Gregory S, Ritchie CW, Ritchie K, Shannon O, Stevenson EJ, Muniz-Terrera G. Mediterranean diet score is associated with greater allocentric processing in the EPAD LCS cohort: A comparative analysis by biogeographical region. Front Aging. 2022;3:1012598.

29. Townsend P, Phillimore P, Beattie A. Health and deprivation. *Nurs Stand* 1988; 2:34.

30. Daniel J. Stekhoven, Peter Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics, Volume 28, Issue 1, January 2012, Pages 112–118, https://doi.org/10.1093/bioinformatics/btr597

31. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

32. Benjamini Y, Hochberg Y Controlling the false discovery rate: a practical and powerful approach to multiple testing J R Stat Soc Ser B Methodol, 57 (1995), pp. 289-300

33. Godos J, Micek A, Carota G, et al. Role of Mediterranean diet in the prevention of cognitive decline: Biological mechanisms behind longevity promotion. Mediterranean Journal of Nutrition and Metabolism. 2025;18(4):227-243. doi:10.1177/1973798X251360765

34. Nucci, Daniele et al. "Association between Mediterranean diet and dementia and Alzheimer disease: a systematic review with meta-analysis." Aging clinical and experimental research vol. 36,1 77. 22 Mar. 2024, doi:10.1007/s40520-024-02718-6

35. Youn, Ji-Eun et al. "Association of Mediterranean, high-quality, and anti-inflammatory diet with dementia in UK Biobank cohort." The journal of nutrition, health & aging vol. 29,7 (2025): 100564. doi:10.1016/j.jnha.2025.100564

36. Wang, Xiaojie et al. "Mediterranean diet and dementia: MRI marker evidence from meta-analysis." European journal of medical research vol. 30,1 32. 16 Jan. 2025, doi:10.1186/s40001-025-02276-1

37. Liu, X., Yang, B., Liu, Q. et al. The long-term neuroprotective effect of MIND and Mediterranean diet on patients with Alzheimer's disease. Sci Rep 15, 32725 (2025). https://doi.org/10.1038/s41598-025-17055-5

38. Nishi, Stephanie K et al. "Mediterranean, DASH, and MIND Dietary Patterns and Cognitive Function: The 2-Year Longitudinal Changes in an Older Spanish Cohort." Frontiers in aging neuroscience vol. 13 782067. 13 Dec. 2021, doi:10.3389/fnagi.2021.782067

39. Liu, Y., Gu, X., Li, Y. et al. Interplay of genetic predisposition, plasma metabolome and Mediterranean diet in dementia risk and cognitive function. Nat Med 31, 3790–3800 (2025). https://doi.org/10.1038/s41591-025-03891-5

40. Wei BZ, Li L, Dong CW, Tan CC; Alzheimer's Disease Neuroimaging Initiative; Xu W. The Relationship of Omega-3 Fatty Acids with Dementia and Cognitive Decline: Evidence from

Prospective Cohort Studies of Supplementation, Dietary Intake, and Blood Markers. Am J Clin Nutr. 2023 Jun;117(6):1096-1109. doi: 10.1016/j.ajcnut.2023.04.001. Epub 2023 Apr 5. PMID: 37028557; PMCID: PMC10447496.

41. Kim, Oh Yoen, and Juhyun Song. “Important roles of linoleic acid and α-linolenic acid in regulating cognitive impairment and neuropsychiatric issues in metabolic-related dementia.” *Life sciences* vol. 337 (2024): 122356. doi:10.1016/j.lfs.2023.122356

42. Merra, Giuseppe et al. “Influence of Mediterranean Diet on Human Gut Microbiota.” *Nutrients* vol. 13,1 7. 22 Dec. 2020, doi:10.3390/nu13010007

43. Park, Gwoncheol et al. “A modified Mediterranean-style diet enhances brain function via specific gut-microbiome-brain mechanisms.” Gut microbes vol. 16,1 (2024): 2323752. doi:10.1080/19490976.2024.2323752

44. Ramezani, Matin et al. “Ketone bodies mediate alterations in brain energy metabolism and biomarkers of Alzheimer's disease.” Frontiers in neuroscience vol. 17 1297984. 16 Nov. 2023, doi:10.3389/fnins.2023.1297984

45. Candás-Estébanez, Beatriz et al. “The Impact of the Mediterranean Diet and Lifestyle Intervention on Lipoprotein Subclass Profiles among Metabolic Syndrome Patients: Findings of a Randomized Controlled Trial.” International journal of molecular sciences vol. 25,2 1338. 22 Jan. 2024, doi:10.3390/ijms25021338

46. Hernáez, Álvaro et al. “Mediterranean Diet Improves High-Density Lipoprotein Function in High-Cardiovascular-Risk Individuals: A Randomized Controlled Trial.” Circulation vol. 135,7 (2017): 633-643. doi:10.1161/CIRCULATIONAHA.116.023712

47. Rudajev V, Novotny J. Cholesterol as a key player in amyloid β-mediated toxicity in Alzheimer's disease. Front Mol Neurosci. 2022 Aug 25;15:937056. doi: 10.3389/fnmol.2022.937056. PMID: 36090253; PMCID: PMC9453481.

48. Yang X, Yao K, Zhang M, Zhang W, Zu H. New insight into the role of altered brain cholesterol metabolism in the pathogenesis of AD: A unifying cholesterol hypothesis and new therapeutic approach for AD. Brain Res Bull. 2025 May;224:111321. doi: 10.1016/j.brainresbull.2025.111321. Epub 2025 Mar 29. PMID: 40164234.

49. Sprenger KG, Lietzke EE, Melchior JT, Bruce KD. Lipid and lipoprotein metabolism in microglia: Alzheimer's disease mechanisms and interventions. J Lipid Res. 2025 Oct;66(10):100872. doi: 10.1016/j.jlr.2025.100872. Epub 2025 Aug 11. PMID: 40769380; PMCID: PMC12538436.

50. Wang, Zixuan et al. “Multi-omics integration reveals the impact of mediterranean diet on hepatic metabolism and gut microbiota in mice with metabolic dysfunction-associated steatotic liver disease.” Frontiers in nutrition vol. 12 1644014. 12 Aug. 2025, doi:10.3389/fnut.2025.1644014

51. Miguel-Albarreal, Antonio D et al. “Mediterranean Diet Combined with Regular Aerobic Exercise and Hemp Protein Supplementation Modulates Plasma Circulating Amino Acids and Improves the Health Status of Overweight Individuals.” Nutrients vol. 16,11 1594. 23 May. 2024, doi:10.3390/nu16111594

52. Akbaraly, Tasnime et al. “Association of circulating metabolites with healthy diet and risk of cardiovascular disease: analysis of two cohort studies.” *Scientific reports* vol. 8,1 8620. 5 Jun. 2018, doi:10.1038/s41598-018-26441-1

53. Smith AN, Morris JK, Carbuhn AF, Herda TJ, Keller JE, Sullivan DK, Taylor MK. Creatine as a Therapeutic Target in Alzheimer's Disease. Curr Dev Nutr. 2023 Sep 29;7(11):102011. doi: 10.1016/j.cdnut.2023.102011. PMID: 37881206; PMCID: PMC10594571.

54. Motolese F, Norata D, Iaccarino G, Sapio E, Capone F. The effect of creatinine level on amyloid-β and tau plasma concentrations in a cohort of Alzheimer's disease patients without kidney disease. Behav Brain Res. 2025 Feb 4;477:115289. doi: 10.1016/j.bbr.2024.115289. Epub 2024 Oct 11. PMID: 39396574.

55. Daidone, Mario et al. “Mediterranean diet effects on vascular health and serum levels of adipokines and ceramides.” PloS one vol. 19,5 e0300844. 29 May. 2024, doi:10.1371/journal.pone.0300844

56. Mayneris-Perxachs, Jordi et al. “Effects of 1-year intervention with a Mediterranean diet on plasma fatty acid composition and metabolic syndrome in a population at high cardiovascular risk.” PloS one vol. 9,3 e85202. 20 Mar. 2014, doi:10.1371/journal.pone.0085202

57. Michielsen, Charlotte C J R et al. “Disentangling the Effects of Monounsaturated Fatty Acids from Other Components of a Mediterranean Diet on Serum Metabolite Profiles: A Randomized Fully Controlled Dietary Intervention in Healthy Subjects at Risk of the Metabolic Syndrome.” Molecular nutrition & food research vol. 63,9 (2019): e1801095. doi:10.1002/mnfr.201801095

58. Andreu-Reinón, María Encarnación et al. “Mediterranean Diet and Risk of Dementia and Alzheimer's Disease in the EPIC-Spain Dementia Cohort Study.” Nutrients vol. 13,2 700. 22 Feb. 2021, doi:10.3390/nu13020700

59. Foster, Hamish M E et al. “The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a

prospective analysis of the UK Biobank cohort." *The Lancet. Public health* vol. 3,12 (2018): e576-e585. doi:10.1016/S2468-2667(18)30200-7

60. Buergel T, Steinfeldt J, Ruyoga G, Pietzner M, Bizzarri D, Vojinovic D, Upmeier Zu Belzen J, Loock L, Kittner P, Christmann L, et al. Metabolomic profiles predict individual multidisease outcomes. Nat Med. 2022;28:2309–2320. doi: 10.1038/s41591-022-01980-3

Table 1: Baseline characteristics of 92,299 UK Biobank participants stratified by incident all-cause dementia status.

		All-cause dementia	
	Overall	**Yes**	**No**
Characteristics	**N = 92,299**	**N = 975**	**N = 91,324**
Age (years), mean (SD)	56·3 (7·9)	64·3 (4·2)	56·2 (7·9)
Sex			
Men	43,804 (47·5)	550 (56·4)	43,254 (47·4)

Women	48,495 (52·5)	425 (43·6)	48,070 (52·6)
BMI (kg/m^2), mean (SD)	27·1 (4·6)	27·4 (4·7)	27·1 (4·6)
TDI, mean (SD)	-1·7 (2·8)	-1·6 (2·9)	-1·7 (2·8)
Smoking status			
Never	52,265 (56·6)	469 (48·1)	51,796 (56·8)
Previous	32,784 (35·5)	430 (44·1)	32,354 (35·4)
Current	7,026 (7·6)	69 (7·1)	6,957 (7·6)
Unknown	224 (0·3)	7 (0·7)	217 (0·2)
Education			
Higher	54,901 (59·5)	454 (46·6)	54,447 (59·7)
Lower	36,984 (40·1)	513 (52·6)	36,471 (39·9)
Unknown	414 (0·4)	8 (0·8)	406 (0·4)
Depression	4,815 (5·2)	69 (7·1)	906 (1·0)
Cardiovascular disease	5,413 (5·9)	171 (17·5)	804 (0·9)
Cancer	8,222 (8·9)	119 (12·2)	856 (0·9)
Diabetes	3,881 (4·2)	114 (11·7)	861 (0·9)
APOE4 (*carriers*)	23,433 (25·4)	496 (50·9)	471 (0·5)
MEDAS, mean (SD)	4·2 (2·0)	3·8 (2·1)	4·2 (2·0)

Abbreviations: BMI, body mass index; TDI, Townsend deprivation index; SD, standard deviation. The number of missing values was 112 for TDI, 200 for BMI and 750 for APOE4. All figures are expressed as absolute numbers (and percentages, %) unless otherwise specified.

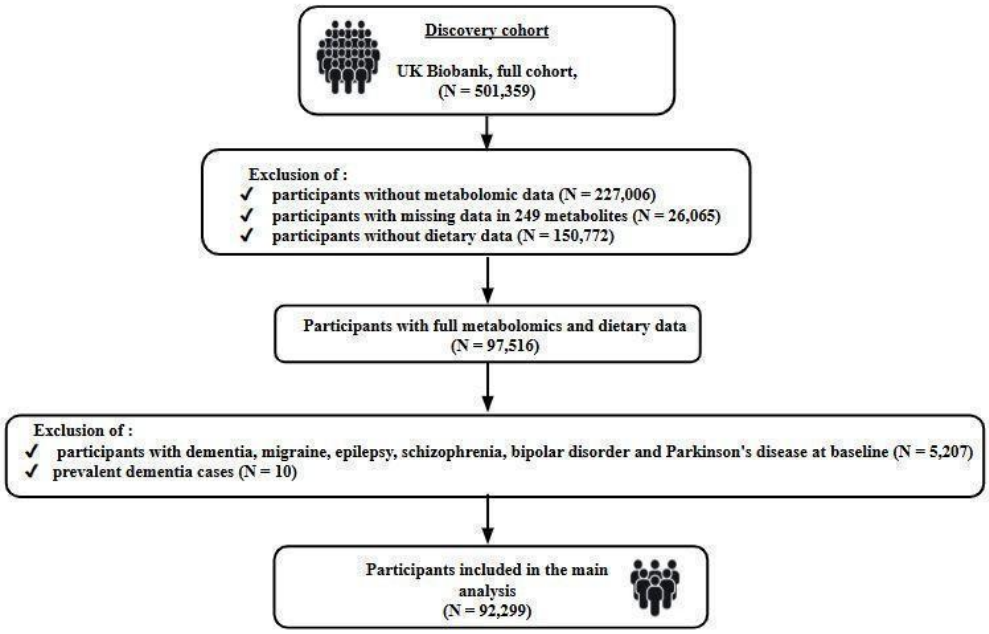


Figure 1: Flowchart of study design.

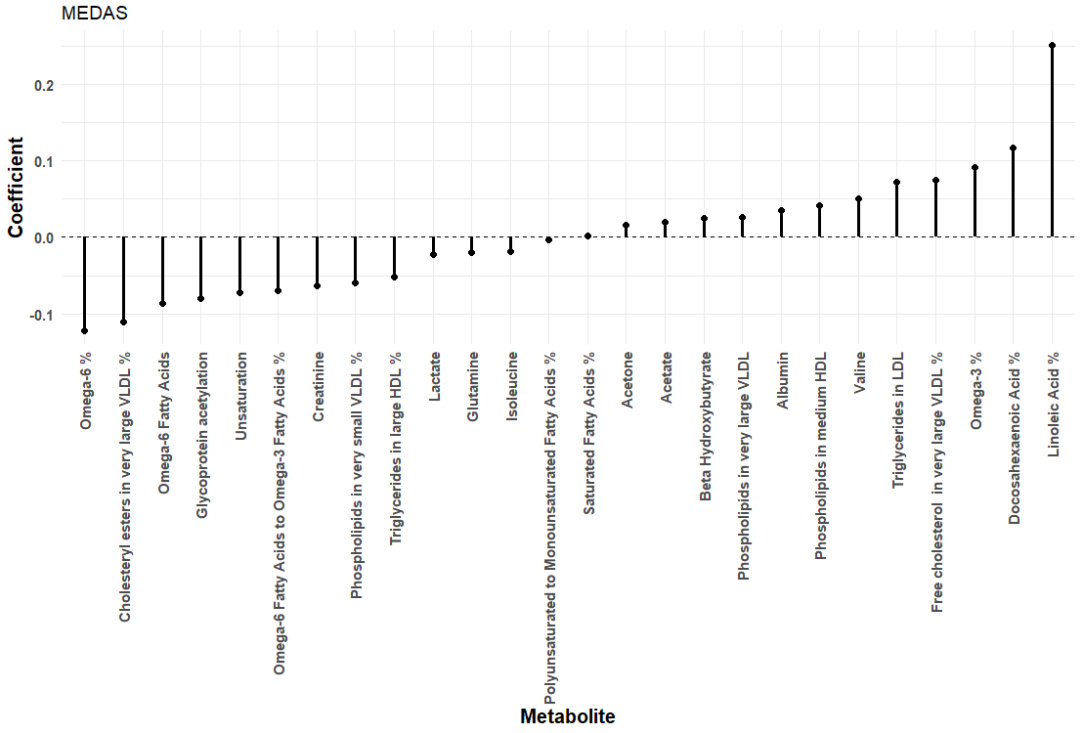


Figure 2. Metabolites ranked from the highest to the lowest MSA-Enet positive and negative regression coefficients for MEDAS. Exposure contrast is per SD/z-score increase of the metabolite (N = 92,299). Abbreviations: HDL, high-density lipoprotein; LDL, low-density lipoprotein; MEDAS, Mediterranean diet adherence screener; VLDL, very low-density lipoprotein; %, Percentage.

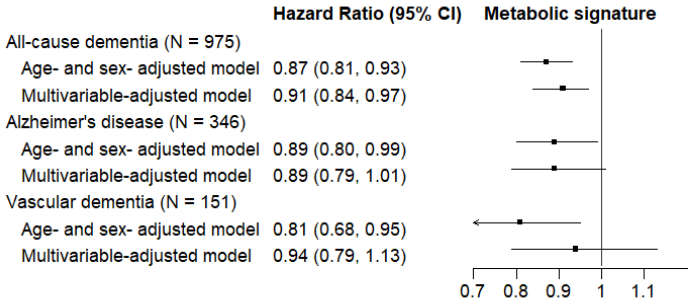


Figure 3: Associations of metabolic signatures with all-cause dementia, Alzheimer's disease and vascular dementia. Abbreviations: CI, confidence interval; Hazard ratio and 95% CI per SD increment in metabolic signature; Multivariable-adjusted model, based on an age and sex-adjusted model, further adjusted for BMI, TDI, diabetes, smoking status, education level, depression, cancer, cardiovascular disease and APOE4 status. The z-score of metabolic signature of MEDAS consists of 26 identified metabolites.

SUPPLEMENTARY MATERIALS

Supplementary Figures

Supplementary Figure 1. Flowchart of participant selection in the Epirus Health Study validation cohort

Supplementary Figure 2. A) Percentage of missing values for the 249 metabolites quantified by Nuclear Magnetic Resonance (NMR) for the discovery population (UK Biobank, N = 92,299). B) Percentage of missing values for the 250 metabolites quantified by Nuclear Magnetic Resonance (NMR) for the external population (Epirus Health Study, N =1,608)

Supplementary Figure 3. Directed Acyclic Graph illustrating the causal relationships between Metabolic signature and all-cause dementia. The minimal sufficient adjustment set includes age, sex, BMI, TDI, diabetes, smoking status, education level, depression, cancer, cardiovascular disease and APOE4 status.

Supplementary Methods

Supplementary Text 1. Data sources

Discovery cohort: The UK Biobank [1] is a resource for examining the determinants of disease in middle and older life. It is an ongoing multi-centre prospective cohort study with over half a million participants. This study's design and methodology have been previously reported. In short, using National Health Service (NHS) patient registers, men and women between the ages of 40 and 70 were recruited from all around England, Scotland, and Wales between 2006 and 2010. Attending one of 22 assessment centres, participants filled out a touchscreen questionnaire, had a verbal interview, and contributed biological samples along with physical function measurements. Participants were then asked to take part in further tests, such as improved nutritional evaluations, imaging, and evaluation of other health-related outcomes. Linkage to electronic health records (cancer, inpatient, death, and primary care) is another feature of UK Biobank that facilitates illness determination.

Validation cohort: The Epirus Health Study (EHS) is a population-based prospective cohort study that was launched in June 2019 and to date over 2,500 participants have been recruited [2, 3]. The overall aim is to improve the general health of the Greek population by shedding light on the complicated aetiology of multiple chronic diseases. The EHS cohort includes individuals who live permanently in Greece's Epirus region between the ages of 25 and 70 and who did not have any signs of an active infection at the time of recruitment. The design and methods of this study have been described in detail elsewhere. The University of Ioannina's Research Ethics Committee gave its approval for the study, which is carried out in compliance with the Helsinki Declaration. Prior to taking part in the study, each subject provided written informed consent.

Supplementary Text 2. Metabolic biomarker measurements by Nuclear Magnetic Resonance (NMR)

EDTA plasma samples from aliquot 3 were assessed using Nightingale Health's NMR-based metabolic biomarker profiling technology. Between June 2019 and June 2022, the measurements were conducted in two stages. For the initial data release, samples were drawn at random from the complete cohort. The samples were produced directly in 96-well plates by UK Biobank. At least 85 µL of plasma was aliquoted into each well using TECAN freedom EVO 150 robotic liquid handlers, which have coefficients of variation in pipetting volume at 0.75% across 8 tips. Batches of 5,000 to 20,000 plasma samples were shipped to Nightingale Health's labs in 96-well plates on dry ice. The platform and experimentation specifics have already been covered in prior descriptions [22, 69]. In a nutshell, EDTA plasma samples were kept at -80°C in a freezer. Prior to processing, frozen materials were gently mixed and centrifuged (3 min, 3400 g, +4°C) to get rid of any potential precipitation. A liquid handler (PerkinElmer Janus Automated Workstation) automatically transferred aliquots of each sample into 3-mm outer-diameter NMR tubes and mixed them in a 1:1 ratio with a phosphate buffer (75mM Na2HPO4 in 80%/20% H2O/D2O, pH 7.4, as well as 0.08% sodium 3-(trimethylsilyl) propionate-2,2,3,3-d4 and 0.04% sodium azide).The measurement consistency inside and between spectrometers for the UK Biobank samples was continuously tracked by Nightingale Health. Each 96-well plate contained two control samples provided by Nightingale Health in order to track consistency between different spectrometers. Furthermore, the UK Biobank supplied two blind duplicate samples for every well plate, the location data was only unlocked following the delivery of the results. The coefficients of variation (CV) targets for the metabolic biomarker profile were pre-established for both the blind duplicates from UK Biobank and the internal control samples from Nightingale Health. The objectives were met for every batch of 25,000 samples that were measured one after the other. The CVs for most metabolic markers are less than 5%. Information about the assessment of metabolic biomarkers by NMR sees the document you can obtain at https://biobank.ndph.ox.ac.uk/ukb/ukb/docs/nmrm_companion_doc.pdf, https://biobank.ndph.ox.ac.uk/ukb/ukb/docs/NMR_companion_phase2.pdf.

Supplementary Text 3. Identification of metabolic signature for MEDAS

We modelled MEDAS as a nonlinear function of 249 metabolites to identify the associated metabolic signature. Nonlinearity was incorporated by including squared terms for each metabolite. Prior to regression, MEDAS was standardized into z-scores. Metabolites were transformed using a rank-based inverse normal transformation, and their squared terms were subsequently computed [4]. The machine learning methodology, including feature selection and hyperparameter tuning were implemented following the literature [5]. Feature selection and metabolic signature were conducted using the discovery population. The external population was reserved exclusively for final evaluation and did not contribute to any machine-learning analyses. We applied the multi-step adaptive elastic-net

(MSA-Enet), a regularization model that effectively addresses the high dimensionality and multicollinearity of metabolomics data [6]. Feature selection was performed using stability selection across 100 bootstrap samples, retaining metabolites appearing in more than 60 iterations [7]. Thereafter, data were split into train (80%) and test (20%) sets and performed grid searching to tune the model using 10-fold cross-validation (CV). The metabolic score was derived from the predictive values (weighted sum) of the tuned model selected as optimal based on the 1000-bootstrapped root mean square error (RMSE). The 1000 bootstrapped Pearson's correlation coefficient was calculated to evaluate the performance of the identified metabolic signature. For external validation, the resulting metabolic score, which was calculated as the weighted sum of the 26 metabolites using the coefficients derived from UK Biobank, was applied to the EHS dataset. We evaluated the performance by calculating RMSE, Pearson correlation coefficient and, R^2.

References

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar 31;12(3):e1001779.
2. Koutsonida M, Koskeridis F, Markozannes G, Kanellopoulou A, Mousas A, Ntotsikas E, Ioannidis P, Aretouli E, Tsilidis KK. Metabolic syndrome and cognitive deficits in the Greek cohort of Epirus Health Study. Neurol Sci. 2023 Oct;44(10):3523-3533.
3. Kanellopoulou A, Koskeridis F, Markozannes G, Bouras E, Soutziou C, Chaliasos K, Doumas MT, Sigounas DE, Tzovaras VT, Panos A, et al. Awareness, knowledge and trust in the Greek authorities towards COVID-19 pandemic: results from the Epirus Health Study cohort. BMC Public Health. 2021 Jun 12;21(1):1125.
4. Wittenbecher C, Guasch-Ferré M, Haslam DE, Dennis C, Li J, Bhupathiraju SN, Lee CH, Qi Q, Liang L, Eliassen AH, et al. Changes in metabolomics profiles over ten years and subsequent risk of developing type 2 diabetes: Results from the Nurses' Health Study. EBioMedicine. 2022 Jan;75:103799.
5. Papagiannopoulos, Christos K et al. "Sex-stratified metabolic signatures of adiposity indices and their associations with clinical biomarkers in the UK Biobank." *EBioMedicine* vol. 119 (2025): 105868. doi:10.1016/j.ebiom.2025.105868
6. Xiao, N., and Q.-S. Xu. 2015. Multi-step adaptive elastic-net: Reducing false positives in high-dimensional variable selection. Journal of Statistical Computation and Simulation 85 (18): 3755–3765. https://doi.org/10.1080/00949655.2015.1016944.

7. Nogueira S, Sechidis K, Brown G. 2018. On the Stability of Feature Selection Algorithms. Journal of Machine Learning Research 18 (174): 1-54. https://jmlr.org/papers/v18/17-514.html

Supplementary Tables

Supplementary Table 1. Median and Interquartile range of concentrations for the 249 metabolites quantified by Nuclear Magnetic Resonance (NMR) for the discovery population (UK Biobank, N = 92,299) and external validation population (Epirus Health Study, N = 1,608)

	UK Biobank (N = 92,299)		Epirus Health Study (N = 1,608)	
Metabolites	**Median**	**IQR**	**Median**	**IQR**
Total-C (mmol/l)	4.623	4.021, 5.248	5.162	4.528, 5.809
Non-HDL-C (mmol/l)	3.276	2.748, 3.845	3.621	3.069, 4.281
Remnant-C (mmol/l)	1.542	1.279, 1.827	1.622	1.362, 1.952
VLDL-C (mmol/l)	0.703	0.551, 0.875	0.696	0.532, 0.877
Clinical-LDL-C (mmol/l)	2.533	2.070, 3.025	2.921	2.418, 3.448
LDL-C (mmol/l)	1.732	1.457, 2.029	2.000	1.700, 2.336
HDL-C (mmol/l)	1.284	1.091, 1.522	1.464	1.284, 1.694
Total Triglycerides (mmol/l)	1.203	0.901, 1.618	1.131	0.878, 1.519
VLDL-TG (mmol/l)	0.828	0.569, 1.183	0.744	0.543, 1.079
LDL-TG (mmol/l)	0.139	0.118, 0.166	0.144	0.121, 0.170
HDL-TG (mmol/l)	0.138	0.113, 0.170	0.124	0.100, 0.153
Total-PL (mmol/l)	2.921	2.620, 3.237	3.136	2.829, 3.424
VLDL-PL (mmol/l)	0.453	0.346, 0.581	0.435	0.325, 0.569
LDL-PL (mmol/l)	0.604	0.515, 0.700	0.683	0.586, 0.785
HDL-PL (mmol/l)	1.527	1.331, 1.750	1.643	1.469, 1.865
Total-CE (mmol/l)	3.361	2.928, 3.809	3.766	3.324, 4.219
VLDL-CE (mmol/l)	0.421	0.331, 0.520	0.420	0.329, 0.530
LDL-CE (mmol/l)	1.268	1.067, 1.486	1.461	1.239, 1.719
HDL-CE (mmol/l)	1.001	0.847, 1.188	1.144	0.999, 1.324
Total-FC (mmol/l)	1.261	1.089, 1.441	1.381	1.205, 1.576
VLDL-FC (mmol/l)	0.281	0.216, 0.357	0.271	0.203, 0.352
LDL-FC (mmol/l)	0.464	0.388, 0.544	0.541	0.461, 0.623
HDL-FC (mmol/l)	0.284	0.242, 0.337	0.322	0.281, 0.373
Total-L (mmol/l)	8.815	7.785, 9.912	9.461	8.431, 10.654
VLDL-L (mmol/l)	2.000	1.504, 2.616	1.894	1.429, 2.508

LDL-L (mmol/l)	2.479	2.103, 2.884	2.826	2.411, 3.290
HDL-L (mmol/l)	2.957	2.572, 3.414	3.239	2.872, 3.690
Total-P (mmol/l)	0.017	0.015, 0.019	0.018	0.017, 0.020
VLDL-P (mmol/l)	<0·001	<0·0001, 0·0001	<0·001	<0·0001, 0·0001
LDL-P (mmol/l)	0.001	0.001, 0.001	0.001	0.001, 0.002
HDL-P (mmol/l)	0.015	0.014, 0.017	0.016	0.015, 0.018
VLDL particle size (nm)	38.609	37.768, 39.507	38.364	37.638, 39.239
LDL particle size (nm)	23.937	23.874, 23.991	23.974	23.904, 24.025
HDL particle size (nm)	9.610	9.491, 9.771	9.672	9.544, 9.824
Phosphoglycerides (mmol/l)	2.272	2.021, 2.539	2.483	2.221, 2.727
TG/PG (ratio)	0.540	0.405, 0.714	0.450	0.363, 0.591
Total cholines (mmol/l)	2.559	2.297, 2.836	2.809	2.536, 3.064
Phosphatidylcolines (mmol/l)	2.090	1.854, 2.343	2.290	2.043, 2.536
Sphingomyelins (mmol/l)	0.449	0.403, 0.497	0.496	0.452, 0.541
ApoB (g/l)	0.839	0.715, 0.976	0.905	0.777, 1.068
ApoA1 (g/l)	1.439	1.290, 1.606	1.547	1.410, 1.707
ApoB/ApoA1 (ratio)	0.581	0.480, 0.700	0.584	0.481, 0.705
Total fatty acids (mmol/l)	11.792	10.407, 13.378	12.522	10.961, 14.132
Unsaturation (degree)	1.361	1.311, 1.409	1.344	1.311, 1.377
Omega-3 (mmol/l)	0.506	0.386, 0.656	0.407	0.312, 0.524
Omega-6 (mmol/l)	4.477	4.053, 4.933	4.902	4.445, 5.420
PUFA (mmol/l)	5.006	4.507, 5.547	5.347	4.811, 5.887
MUFA (mmol/l)	2.746	2.318, 3.290	2.919	2.483, 3.459
SFA (mmol/l)	3.983	3.463, 4.599	4.158	3.640, 4.728
LA (mmol/l)	3.434	3.015, 3.892	3.903	3.477, 4.468
DHA (mmol/l)	0.229	0.185, 0.283	0.206	0.176, 0.242
Omega-3 (%)	4.258	3.426, 5.240	3.237	2.646, 3.894
Omega-6 (%)	38.429	35.939, 40.414	39.847	37.917, 41.333
PUFA (%)	42.911	40.352, 44.933	43.149	41.455, 44.395
MUFA (%)	23.358	21.834, 25.195	23.400	22.293, 24.893
SFA (%)	33.831	32.699, 35.075	33.473	32.646, 34.317
LA (%)	29.279	26.981, 31.38	31.791	29.995, 33.442
DHA (%)	1.957	1.593, 2.376	1.672	1.450, 1.926
PUFA/MUFA (ratio)	1.839	1.607, 2.050	1.842	1.677, 1.986
Omega-6/Omega-3 (ratio)	8.859	7.078, 11.243	12.180	9.928, 15.305
Alanine (mmol/l)	0.292	0.243, 0.349	0.340	0.296, 0.392
Glutamine (mmol/l)	0.549	0.496, 0.604	0.622	0.568, 0.678
Glycine (mmol/l)	0.159	0.129, 0.201	0.238	0.208, 0.279
Histidine (mmol/l)	0.065	0.059, 0.072	0.079	0.071, 0.086
Total BCAA (mmol/l)	0.354	0.307, 0.410	0.393	0.339, 0.452

Isoleucine (mmol/l)	0.048	0.039, 0.059	0.050	0.040, 0.061
Leucine (mmol/l)	0.100	0.085, 0.119	0.116	0.098, 0.137
Valine (mmol/l)	0.206	0.181, 0.235	0.228	0.201, 0.259
Phenylalanine (mmol/l)	0.046	0.040, 0.053	0.067	0.059, 0.076
Tyrosine (mmol/l)	0.061	0.053, 0.071	0.061	0.053, 0.069
Glucose (mmol/l)	3.516	3.074, 3.980	4.563	4.148, 4.992
Lactate (mmol/l)	3.919	3.228, 4.707	2.374	1.993, 2.835
Pyruvate (mmol/l)	0.080	0.063, 0.099	0.028	0.018, 0.041
Citrate (mmol/l)	0.065	0.057, 0.073	0.061	0.055, 0.069
3-Hydroxybutyrate (mmol/l)	0.043	0.030, 0.068	-	-
Acetate (mmol/l)	0.015	0.012, 0.019	0.020	0.014, 0.028
Acetoacetate (mmol/l)	0.010	0.006, 0.015	0.022	0.014, 0.036
Acetone (mmol/l)	0.013	0.011, 0.016	0.016	0.014, 0.020
Creatinine(µmol/l)	0.066	0.058, 0.075	68.445	59.668, 78.859
Albumin(g/l)	39.543	37.477, 41.628	43.189	41.064, 45.956
Glycoprotein acetyls (mmol/l)	0.794	0.723, 0.874	0.806	0.731, 0.886
XXL-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
XXL-VLDL-L (mmol/l)	0.163	0.072, 0.315	0.076	0.028, 0.197
XXL-VLDL-PL (mmol/l)	0.026	0.012, 0.050	0.010	0.003, 0.028
XXL-VLDL-C (mmol/l)	0.045	0.024, 0.077	0.026	0.014, 0.053
XXL-VLDL-CE (mmol/l)	0.026	0.013, 0.043	0.017	0.010, 0.032
XXL-VLDL-FC (mmol/l)	0.020	0.010, 0.034	0.009	0.004, 0.021
XXL-VLDL-TG (mmol/l)	0.092	0.037, 0.188	0.039	0.014, 0.113
XL-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
XL-VLDL-L (mmol/l)	0.179	0.107, 0.278	0.151	0.095, 0.244
XL-VLDL-PL (mmol/l)	0.034	0.020, 0.053	0.027	0.016, 0.046
XL-VLDL-C (mmol/l)	0.051	0.033, 0.071	0.046	0.030, 0.067
XL-VLDL-CE (mmol/l)	0.029	0.020, 0.039	0.028	0.019, 0.039
XL-VLDL-FC (mmol/l)	0.021	0.013, 0.032	0.018	0.011, 0.028
XL-VLDL-TG (mmol/l)	0.094	0.052, 0.154	0.078	0.047, 0.132
L-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
L-VLDL-L (mmol/l)	0.308	0.205, 0.440	0.291	0.196, 0.43
L-VLDL-PL (mmol/l)	0.062	0.038, 0.091	0.056	0.035, 0.086
L-VLDL-C (mmol/l)	0.094	0.065, 0.128	0.087	0.057, 0.124
L-VLDL-CE (mmol/l)	0.050	0.035, 0.067	0.046	0.031, 0.064
L-VLDL-FC (mmol/l)	0.044	0.029, 0.062	0.040	0.026, 0.059
L-VLDL-TG (mmol/l)	0.152	0.101, 0.222	0.146	0.103, 0.217
M-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
M-VLDL-L (mmol/l)	0.564	0.438, 0.711	0.593	0.457, 0.752
M-VLDL-PL (mmol/l)	0.126	0.097, 0.160	0.135	0.102, 0.171

M-VLDL-C (mmol/l)	0.169	0.128, 0.214	0.181	0.142, 0.234
M-VLDL-CE (mmol/l)	0.091	0.067, 0.117	0.100	0.076, 0.127
M-VLDL-FC (mmol/l)	0.078	0.059, 0.098	0.082	0.063, 0.105
M-VLDL-TG (mmol/l)	0.264	0.195, 0.349	0.270	0.202, 0.360
S-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
S-VLDL-L (mmol/l)	0.406	0.326, 0.496	0.400	0.317, 0.504
S-VLDL-PL (mmol/l)	0.096	0.077, 0.117	0.100	0.079, 0.123
S-VLDL-C (mmol/l)	0.154	0.122, 0.190	0.156	0.122, 0.195
S-VLDL-CE (mmol/l)	0.096	0.076, 0.119	0.094	0.072, 0.118
S-VLDL-FC (mmol/l)	0.058	0.046, 0.071	0.061	0.049, 0.076
S-VLDL-TG (mmol/l)	0.153	0.118, 0.195	0.143	0.111, 0.187
XS-VLDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
XS-VLDL-L (mmol/l)	0.352	0.299, 0.412	0.346	0.290, 0.413
XS-VLDL-PL (mmol/l)	0.103	0.087, 0.121	0.101	0.083, 0.120
XS-VLDL-C (mmol/l)	0.182	0.151, 0.215	0.182	0.151, 0.217
XS-VLDL-CE (mmol/l)	0.125	0.103, 0.149	0.126	0.105, 0.152
XS-VLDL-FC (mmol/l)	0.057	0.048, 0.067	0.055	0.046, 0.066
XS-VLDL-TG (mmol/l)	0.067	0.055, 0.081	0.064	0.052, 0.079
IDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
IDL-L (mmol/l)	1.227	1.042, 1.419	1.346	1.147, 1.543
IDL-PL (mmol/l)	0.291	0.250, 0.335	0.319	0.271, 0.362
IDL-C (mmol/l)	0.836	0.700, 0.978	0.933	0.785, 1.075
IDL-CE (mmol/l)	0.618	0.516, 0.723	0.696	0.589, 0.806
IDL-FC (mmol/l)	0.218	0.183, 0.256	0.236	0.199, 0.270
IDL-TG (mmol/l)	0.096	0.082, 0.114	0.097	0.082, 0.115
L-LDL-P (mmol/l)	0.001	0.001, 0.001	0.001	0.001, 0.001
L-LDL-L (mmol/l)	1.580	1.343, 1.834	1.806	1.548, 2.092
L-LDL-PL (mmol/l)	0.356	0.303, 0.411	0.397	0.342, 0.459
L-LDL-C (mmol/l)	1.129	0.951, 1.319	1.314	1.120, 1.526
L-LDL-CE (mmol/l)	0.833	0.703, 0.974	0.971	0.826, 1.128
L-LDL-FC (mmol/l)	0.295	0.247, 0.346	0.343	0.293, 0.397
L-LDL-TG (mmol/l)	0.093	0.080, 0.110	0.097	0.082, 0.114
M-LDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
M-LDL-L (mmol/l)	0.616	0.511, 0.729	0.703	0.583, 0.831
M-LDL-PL (mmol/l)	0.161	0.134, 0.189	0.185	0.156, 0.218
M-LDL-C (mmol/l)	0.423	0.349, 0.503	0.483	0.400, 0.576
M-LDL-CE (mmol/l)	0.303	0.248, 0.364	0.344	0.281, 0.411
M-LDL-FC (mmol/l)	0.119	0.099, 0.141	0.140	0.117, 0.162
M-LDL-TG (mmol/l)	0.031	0.026, 0.038	0.032	0.027, 0.039
S-LDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001

S-LDL-L (mmol/l)	0.282	0.242, 0.327	0.316	0.272, 0.367
S-LDL-PL (mmol/l)	0.088	0.077, 0.101	0.099	0.087, 0.113
S-LDL-C (mmol/l)	0.179	0.152, 0.210	0.204	0.173, 0.237
S-LDL-CE (mmol/l)	0.130	0.109, 0.152	0.146	0.123, 0.170
S-LDL-FC (mmol/l)	0.050	0.042, 0.058	0.058	0.050, 0.067
S-LDL-TG (mmol/l)	0.014	0.012, 0.018	0.014	0.012, 0.018
XL-HDL-P (mmol/l)	<0·0001	<0·0001, <0·0001	<0·0001	<0·0001, <0·0001
XL-HDL-L (mmol/l)	0.150	0.113, 0.205	0.175	0.134, 0.233
XL-HDL-PL (mmol/l)	0.068	0.047, 0.100	0.080	0.057, 0.113
XL-HDL-C (mmol/l)	0.075	0.058, 0.098	0.088	0.071, 0.112
XL-HDL-CE (mmol/l)	0.052	0.039, 0.072	0.064	0.051, 0.085
XL-HDL-FC (mmol/l)	0.022	0.019, 0.027	0.023	0.020, 0.028
XL-HDL-TG (mmol/l)	0.007	0.005, 0.009	0.006	0.005, 0.008
L-HDL-P (mmol/l)	0.001	0.001, 0.002	0.002	0.001, 0.002
L-HDL-L (mmol/l)	0.587	0.414, 0.836	0.725	0.532, 0.982
L-HDL-PL (mmol/l)	0.296	0.212, 0.413	0.360	0.266, 0.476
L-HDL-C (mmol/l)	0.262	0.173, 0.392	0.342	0.241, 0.473
L-HDL-CE (mmol/l)	0.203	0.133, 0.304	0.266	0.187, 0.370
L-HDL-FC (mmol/l)	0.059	0.040, 0.088	0.076	0.054, 0.105
L-HDL-TG (mmol/l)	0.029	0.022, 0.037	0.026	0.019, 0.034
M-HDL-P (mmol/l)	0.004	0.003, 0.004	0.004	0.004, 0.005
M-HDL-L (mmol/l)	1.026	0.886, 1.179	1.105	0.981, 1.250
M-HDL-PL (mmol/l)	0.482	0.421, 0.549	0.513	0.458, 0.577
M-HDL-C (mmol/l)	0.488	0.412, 0.575	0.547	0.477, 0.630
M-HDL-CE (mmol/l)	0.403	0.341, 0.472	0.451	0.394, 0.515
M-HDL-FC (mmol/l)	0.086	0.070, 0.103	0.098	0.084, 0.114
M-HDL-TG (mmol/l)	0.052	0.041, 0.064	0.046	0.035, 0.057
S-HDL-P (mmol/l)	0.010	0.009, 0.011	0.010	0.010, 0.011
S-HDL-L (mmol/l)	1.157	1.059, 1.260	1.211	1.120, 1.308
S-HDL-PL (mmol/l)	0.659	0.602, 0.720	0.684	0.632, 0.739
S-HDL-C (mmol/l)	0.445	0.407, 0.486	0.481	0.445, 0.518
S-HDL-CE (mmol/l)	0.330	0.301, 0.361	0.356	0.327, 0.386
S-HDL-FC (mmol/l)	0.115	0.105, 0.126	0.124	0.115, 0.134
S-HDL-TG (mmol/l)	0.051	0.041, 0.063	0.046	0.037, 0.057
XXL-VLDL-PL (%)	15.719	14.495, 16.869	13.242	8.890, 15.259
XXL-VLDL-C (%)	26.468	23.089, 32.281	30.500	24.622, 42.057
XXL-VLDL-CE (%)	14.613	12.223, 18.531	19.185	14.699, 28.692
XXL-VLDL-FC (%)	11.695	10.507, 13.900	11.228	9.516, 13.805
XXL-VLDL-TG (%)	57.834	51.606, 61.855	55.692	45.377, 62.944
XL-VLDL-PL (%)	18.769	17.808, 19.602	17.773	16.232, 18.687

XL-VLDL-C (%)	28.051	24.297, 32.775	29.456	25.335, 33.848
XL-VLDL-CE (%)	16.207	13.171, 20.087	18.153	14.703, 21.807
XL-VLDL-FC (%)	11.813	10.984, 12.771	11.222	10.475, 12.090
XL-VLDL-TG (%)	53.273	48.412, 57.418	53.360	48.540, 57.551
L-VLDL-PL (%)	19.812	18.424, 20.759	19.134	17.389, 20.106
L-VLDL-C (%)	30.008	27.436, 32.725	28.383	26.058, 30.859
L-VLDL-CE (%)	16.128	14.000, 18.408	15.130	13.283, 17.180
L-VLDL-FC (%)	13.959	13.286, 14.617	13.322	12.639, 14.008
L-VLDL-TG (%)	50.619	47.586, 53.53	52.973	49.832, 56.066
M-VLDL-PL (%)	22.393	21.210, 23.447	22.417	21.416, 23.456
M-VLDL-C (%)	30.539	25.860, 34.840	31.094	27.202, 34.811
M-VLDL-CE (%)	16.664	13.156, 19.961	17.403	14.360, 20.278
M-VLDL-FC (%)	13.849	12.658, 14.913	13.709	12.726, 14.687
M-VLDL-TG (%)	47.059	41.763, 52.863	46.460	41.796, 51.357
S-VLDL-PL (%)	23.715	22.440, 25.008	24.653	23.548, 25.934
S-VLDL-C (%)	38.297	35.137, 41.192	39.017	35.946, 41.578
S-VLDL-CE (%)	23.813	21.932, 25.526	23.524	21.644, 25.148
S-VLDL-FC (%)	14.431	13.029, 15.759	15.337	14.171, 16.572
S-VLDL-TG (%)	37.971	33.860, 42.365	36.214	32.622, 40.239
XS-VLDL-PL (%)	29.183	28.652, 29.766	28.973	28.373, 29.516
XS-VLDL-C (%)	51.962	48.798, 54.474	52.791	50.453, 54.993
XS-VLDL-CE (%)	35.806	32.930, 38.116	36.826	34.646, 38.920
XS-VLDL-FC (%)	16.151	15.757, 16.458	15.948	15.652, 16.201
XS-VLDL-TG (%)	18.874	16.675, 21.630	18.294	16.411, 20.483
IDL-PL (%)	23.865	23.267, 24.447	23.652	23.120, 24.109
IDL-C (%)	68.288	66.330, 69.824	69.259	67.853, 70.381
IDL-CE (%)	50.338	48.757, 51.644	51.692	50.426, 52.810
IDL-FC (%)	17.830	17.120, 18.487	17.443	16.916, 17.960
IDL-TG (%)	7.876	6.766, 9.397	7.187	6.361, 8.228
L-LDL-PL (%)	22.476	21.987, 22.997	21.975	21.660, 22.364
L-LDL-C (%)	71.435	70.446, 72.293	72.563	71.701, 73.304
L-LDL-CE (%)	52.799	52.001, 53.509	53.606	52.826, 54.363
L-LDL-FC (%)	18.701	17.888, 19.409	18.944	18.390, 19.455
L-LDL-TG (%)	5.941	5.181, 6.947	5.367	4.750, 6.121
M-LDL-PL (%)	26.171	25.630, 26.680	26.541	26.022, 26.983
M-LDL-C (%)	68.774	67.795, 69.480	68.914	68.078, 69.536
M-LDL-CE (%)	49.262	48.051, 50.378	48.820	47.473, 50.101
M-LDL-FC (%)	19.493	18.327, 20.505	20.051	19.025, 20.955
M-LDL-TG (%)	5.068	4.414, 5.997	4.598	4.111, 5.279
S-LDL-PL (%)	31.183	30.147, 32.294	31.297	30.402, 32.185

S-LDL-C (%)	63.543	62.277, 64.588	64.206	63.168, 65.019
S-LDL-CE (%)	45.908	44.609, 47.165	45.674	44.688, 46.677
S-LDL-FC (%)	17.784	16.534, 18.710	18.567	17.695, 19.169
S-LDL-TG (%)	5.095	4.306, 6.204	4.448	3.930, 5.220
XL-HDL-PL (%)	45.933	42.007, 49.023	46.091	42.309, 49.09
XL-HDL-C (%)	49.322	47.121, 52.279	49.991	47.735, 53.000
XL-HDL-CE (%)	35.094	33.595, 36.634	36.876	35.549, 38.617
XL-HDL-FC (%)	14.509	12.628, 16.785	13.201	11.827, 15.113
XL-HDL-TG (%)	4.348	3.054, 6.325	3.565	2.722, 4.973
L-HDL-PL (%)	49.87	48.433, 51.780	49.078	48.112, 50.353
L-HDL-C (%)	45.301	41.753, 47.822	47.383	45.285, 48.962
L-HDL-CE (%)	35.212	32.093, 37.347	36.920	35.051, 38.383
L-HDL-FC (%)	10.232	9.560, 10.706	10.509	10.042, 10.897
L-HDL-TG (%)	4.728	3.410, 6.772	3.475	2.642, 4.694
M-HDL-PL (%)	46.890	46.198, 47.792	46.277	45.731, 46.877
M-HDL-C (%)	47.964	45.872, 49.679	49.622	48.117, 50.903
M-HDL-CE (%)	39.654	37.851, 41.081	40.801	39.493, 41.932
M-HDL-FC (%)	8.336	7.891, 8.782	8.790	8.459, 9.174
M-HDL-TG (%)	5.148	4.058, 6.387	4.165	3.249, 5.155
S-HDL-PL (%)	57.016	56.233, 57.817	56.497	55.825, 57.223
S-HDL-C (%)	38.583	37.345, 39.683	39.743	38.658, 40.639
S-HDL-CE (%)	28.680	27.453, 29.737	29.537	28.445, 30.466
S-HDL-FC (%)	9.884	9.590, 10.219	10.199	9.927, 10.540
S-HDL-TG (%)	4.435	3.661, 5.286	3.758	3.139, 4.493

Abbreviations: -C, cholesterol; -TG, triglycerides; -PL, Phospholipids; -CE, Cholesteryl esters; -FC, Free cholesterol; -L, Total lipids; -P, Lipoprotein particle concentrations; XXL-, Chylomicrons and extremely large; XL-, Very large; L-, Large; M-, Medium; S-, Small; XS- Very small; LA, Linoleic Acid; DHA, docosahexaenoic acid; FA, fatty acids; HDL, high-density lipoprotein; IDL, intermediate-density lipoproteins; LDL, low-density lipoprotein; MUFA, monounsaturated fatty acid; PUFA, polyunsaturated fatty acid; SFA, saturated fatty acid; VLDL, very low-density lipoprotein; VHDL, very high-density lipoprotein, %, Percentage; IQR, interquartile range.

Supplementary Table 2. Components and scoring of the MEDAS Mediterranean diet adherence scales

Food component	Contributing foods from the Oxford WebQ	MEDAS[1]	
		Servings required for 0 points	Servings required for 1 point
Olive oil	Type of fat/ oil used for cooking (20090, 103980)	Non-consumption	Consumption
Vegetables	Carrot (104170), spinach (104300), broccoli (104140), cabbage/ kale (104160), sprouts (104310), courgette (104200), cauliflower (104180), parsnip (104270), turnip/ swede (104360), leek (104230), onion (104260), garlic (104220), mushroom (104250), sweet pepper (104290), side salad (104090), lettuce (104240), cucumber (104210), celery (104190), watercress (104370), fresh tomato (104340), tin tomato (104350), sweet corn (104320), beetroot (104130), avocado (104100), mixed vegetables (104060), vegetable pieces (104070), butternut squash (104150), other veg (104380), olives (102490), coleslaw (104080), guacamole (20088), vegetables from canned soup (102540, 20108), vegetables from homemade soup (102620, 20109)	<2/d	≥2/d
Fruit	Apple (104450), pear (104560), orange (104530), satsuma (104540), grapefruit (104490), banana (104460), grape (104500), melon (104520), peach/ nectarine (104550), plum (104580), berry (104470), dried fruit (104430), stewed fruit (104410), mixed fruit (104440), prune (104420), cherry (104480), mango (104510), pineapple (104570), other fruit (104590), fruit added to cereal (100880), grapefruit juice (100200), orange juice (100190)	<3/d	≥3/d
Red meat	Beef (103020), pork (103030), lamb (103040), red meat from canned soup (102540, 20108), red meat from homemade soup (102620, 20109), bacon (103070), ham (103080), sausage (103010), liver (103090), meat from Scotch egg (102970)	>1/d	<1/d
Butter, margarine or cream	Butter/ margarine on potato (104040), baguettes with butter/ margarine (101350, 20099), baps with butter/ margarine (101390, 20100), bread rolls with butter/ margarine (101430, 20101), bread slices with butter/ margarine (101310, 20098), crackers/ crispbread with butter/ margarine (101470, 20102), oatcakes with butter/ margarine (101510, 20103), other bread with butter/ margarine (101550, 20104), butter/ margarine used in cooking (20090), cream (20088)	>1/d	<1/d
Sweetened or carbonated drinks	Fizzy drinks (100170), low calories drinks (100160), squash intake (100180)	>1/d	<1/d
Wine	Red wine (100590), rose wine (100630), white wine (100670)	<7/wk	≥7/wk

Legumes	Peas (104280), green beans (104120), broad beans (104110), baked beans (104000), pulses (104010), tofu (103270), hummus (20088), pulses from canned soup (102540, 20108), pulses from homemade soup (102620, 20109)	<3/wk	≥3/wk
Seafood	Battered fish (103180), breaded fish (103170), white fish (103190), oily fish (103160), shellfish (103220), other fish (103230), tinned tuna (103150), prawn (103200), lobster/ crab (103210), fish from canned soup (102540, 20108), fish from homemade soup (102620, 20109)	<3/wk	≥3/wk
Sweets or pastries	Chocolate biscuit (102350), chocolate covered biscuit (102340), chocolate bar (102260), chocolate sweets (102310), chocolate raisins (102300), dark chocolate (102290), milk chocolate (102280), white chocolate (102270), sweet biscuits (102360), cakes (102190), cheesecake (102220), doughnut (102200), fruitcake (102180), Danish pastry (102060), sponge pudding (102210), milk based pudding (102140), other milk based pudding (102150), other desert intake (102230), soya desert intake (102170), sweets (102330), diet sweets (102320), other sweets (102380), ice cream (102120)	>2/wk	<2/wk
Nuts	Unsalted nuts (102440), salted nuts (102430), unsalted peanuts (102420), salted peanuts (102410), peanut butter (20088)	<3/wk	≥3/wk
White meat	Poultry (103060), breaded poultry (103050), white meat from canned soup (102540, 20108), white meat from homemade soup (102620, 20109)	Less white meat than red meat	More white meat than red meat
Sofrito	Tomato-based sauce (103310, 20088)	<2/wk	≥2/wk

Supplementary Table 3. ICD-9 and ICD-10 codes used for ascertainment of dementia diagnosis

Dementia sub-type	ICD-9	ICD-10
All-cause dementia (other codes)	290.2, 290.3, 291.2, 294.1, 331.2, 331.5	A81.0, F02, F02.1, F02.2, F02.3, F02.4, F02.8, F03, F05.1, F10.6, G31.1, G31.8
Alzheimer's disease	331.0	F00, F00.0, F00.1, F00.2, F00.9, G30, G30.0, G30.1, G30.8, G30.9
Vascular dementia	290.4	F01, F01.0, F01.1, F01.2, F01.3, F01.8, F01.9, I67.3

Abbreviations: ICD, international classification of diseases

Supplementary Table 4. Antidiabetic drugs as coded in UK Biobank (Data-field 20003)

1140857494 glibornuride	1140874724 daonil 5mg tablet
1140857496 glutril 25mg tablet	1140874726 semi-daonil 2·5mg tablet
1140857500 glymidine	1140874728 euglucon 2·5mg tablet
1140857502 gondafon 500mg tablet	1140874732 malix 2·5mg tablet
1140857506 pramidex 500mg tablet	1140874736 diabetamide 2·5mg tablet
1140874646 glipizide	1140874744 gliclazide
1140874650 glibenese 5mg tablet	1140874746 diamicron 80mg tablet
1140874652 minodiab 2·5mg tablet	1140883066 insulin product
1140874664 tolazamide	1140884600 metformin
1140874666 tolanase 100mg tablet	1141152590 glimepiride
1140874674 tolbutamide	1141153254 troglitazone
1140874678 glyconon 500mg tablet	1141153262 romozin 200mg tablet
1140874680 rastinon 500mg tablet	1141156984 amaryl 1mg tablet
1140874686 glucophage 500mg tablet	1141171646 pioglitazone
1140874690 orabet 500mg tablet	1141171652 actos 15mg tablet
1140874706 chlorpropamide	1141177600 rosiglitazone
1140874712 diabinese 100mg tablet	1141177606 avandia 4mg tablet
1140874716 glymese 250mg tablet	1141189090 rosiglitazone 1mg / metformin 500mg tablet
1140874718 glibenclamide	1141189094 avandamet 1mg / 500mg tablet

Supplementary Table 5. Summary of all relevant UK Biobank Data-fields

Variable	Designation
Demographic	
Age (years)	21003
Sex	31
Townsend Deprivation Index	22189
Education	6138
Ethnic background	21000
Lifestyle factors	
Physical activity	22032
Alcohol status	20117
Smoking status	20116
Sleep duration (hours/day)	1160
Body mass index (kg/m^2)	21001
Genetic	
ApoE4	affy16020316, affy16020324
Genetic sex	22001
Medication and diseases	
Medication for cholesterol, blood pressure or diabetes	6177
Medication for cholesterol, blood pressure, diabetes or take exogenous hormones	6153
Treatment/medication code	20003
Non-cancer illness, self-reported	20002
Cancer code, self-reported	20001
Type 2 diabetes diagnosis	2443
Glycated haemoglobin (HbA1c)	30750
Diagnoses - ICD10	41270
Diagnoses – ICD9	41271
Date of first in-patient diagnosis - ICD10	41280
Date of first in-patient diagnosis – ICD9	41281
Date of death	40000
Underlying (primary) cause of death: ICD10	40001

Contributory (secondary) causes of death: ICD10	40002

Supplementary Table 6. Baseline characteristics of Epirus Health Study participants (N = 1,608)

Characteristics	Overall N = 1,608
Age (years), mean (SD)	46·8 (11·1)
Sex	
Men	658 (40·9)
Women	950 (59·1)
Smoking status	
Never	737 (45·8)
Previous	326 (20·3)
Current	545 (33·9)
Alcohol status	
Never	207 (12·9)
Less than once/month	480 (29·8)
1-3 times/month	245 (15·2)
1-2 times/week	476 (29·6)
3-4 times/week	120 (7·5)
Almost everyday	80 (5·0)
Education	
Primary school	10 (0·6)
Junior High school	105 (6·5)
High school	461 (28·7)
University degree	654 (40·7)
Master's degree	273 (17·0)
Doctor of philosophy	104 (6·5)
BMI (kg/m^2), mean (SD)	26·8 (4·8)
SBP (mm Hg), median (IQR)	115·0 (108·0 - 125·0)
DBP (mm Hg), median (IQR)	75·0 (67·0 - 83·0)
Cancer	28 (1·7)
Diabetes	55 (3·4)
MEDAS, mean (SD)	7·2 (1·7)

Abbreviations: BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; SD, standard deviation; IQR, interquartile range. All figures are expressed as absolute numbers (and percentages, %) unless otherwise specified.

Supplementary Table 7: Metabolites ranked from the highest to the lowest MSA-Enet regression coefficients for MEDAS (N = 92,299)

MEDAS			
Metabolite	**β**	**Metabolite**	**β**
Linoleic Acid to Total Fatty Acids, LA (%)	0·251	Omega-6 Fatty Acids to Total Fatty Acids (%)	-0·122
Docosahexaenoic Acid to Total Fatty Acids, DHA (%)	0·117	Cholesteryl esters to total lipids ratio in very large VLDL (%)	-0·111
Omega-3 Fatty Acids to Total Fatty Acids (%)	0·091	Omega-6 Fatty Acids (mmol/l)	-0·087
Free cholesterol to total lipids ratio in very large VLDL (%)	0·074	Glycoprotein acetylation, GlycA (mmol/l)	-0·080
Triglycerides in LDL (mmol/l)	0·072	Degree of Unsaturation	-0·072
Valine (mmol/l)	0·050	Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio (%)	-0·070
Phospholipids in medium HDL (mmol/l)	0·041	Creatinine (mmol/l)	-0·063
Albumin	0·035	Phospholipids to total lipids ratio in very small VLDL (%)	-0·059
Phospholipids in very large VLDL (mmol/l)	0·026	Triglycerides to total lipids ratio in large HDL (%)	-0·052
Beta Hydroxybutyrate (mmol/l)	0·025	Lactate (mmol/l)	-0·023
Acetate (mmol/l)	0·020	Glutamine (mmol/l)	-0·019
Acetone (mmol/l)	0·017	Isoleucine (mmol/l)	-0·018
Saturated Fatty Acids to Total Fatty Acids, SFA (%)	0·002	Polyunsaturated Fatty Acids to Monounsaturated Fatty Acids ratio (%)	-0·004

Abbreviations: HDL, high-density lipoprotein, LDL, low-density lipoprotein, MEDAS, Mediterranean diet adherence screener, MSA-Enet, multi-step adaptive elastic net, TG, triglycerides, VLDL, very-low density lipoprotein· Exposure contrast is per SD/z-score increase of the metabolite.

Supplementary Table 8. Associations of metabolic signature with all-cause dementia, Alzheimer's disease and vascular dementia.

	Metabolic signature*	
	HR (95% CI)[1]	**P-value**
All-cause dementia (N = 975)		
Age- and sex- adjusted model	0·87 (0·81, 0·93)	$3{\cdot}15 \times 10^{-5}$
Multivariable-adjusted model[3]	0·91 (0·84, 0·97)	$6{\cdot}66 \times 10^{-3}$
AD (N = 346)		
Age- and sex- adjusted model	0·89 (0·80, 0·99)	$3{\cdot}84 \times 10^{-2}$
Multivariable-adjusted model[2]	0·89 (0·79, 1·01)	$6{\cdot}61 \times 10^{-2}$
VD (N = 151)		
Age- and sex- adjusted model	0·81 (0·68, 0·95)	$1{\cdot}25 \times 10^{-2}$
Multivariable-adjusted model[2]	0·94 (0·79, 1·13)	$5{\cdot}21 \times 10^{-1}$

Abbreviations: CI, confidence interval, HR, hazard ratio, *: The metabolic signature consists of the 26 identified metabolites presented in Supplementary Table 7 , [1]:HR and 95% CI per SD increment in metabolic signature, [2]: Based on an age and sex-adjusted model, further adjusted for BMI, TDI, diabetes, smoking status, education level, depression, cancer, cardiovascular disease and APOE4 status.

Supplementary Table 9. Stratified associations of metabolic signature with all-cause dementia.

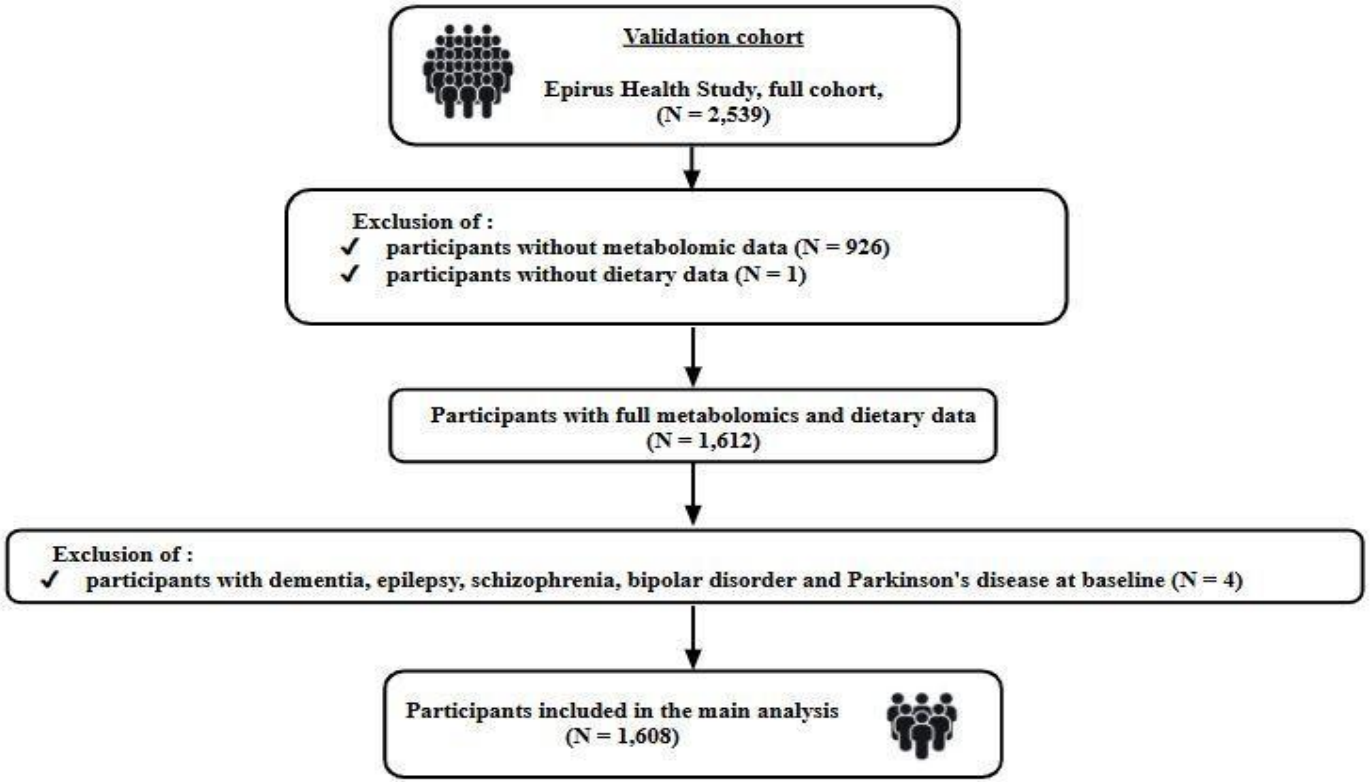
Modifier	Level	Metabolic Signature* HR (95% CI) [1]	Metabolic Signature p interaction (FDR)
Age	<57 years	0.72 (0.51, 1.01)	NS
Age	≥57 years	0.93 (0.87, 1.00)	
BMI	<25 kg/m^2	0.92 (0.82, 1.03)	NS
BMI	≥25 kg/m^2	0.89 (0.82, 0.97)	
TDI	<-2.43	0.98 (0.88, 1.08)	**<0.001**
TDI	≥−2.43	0.84 (0.77, 0.93)	
Sex	Women	0.92 (0.83, 1.02)	NS
Sex	Men	0.90 (0.82, 0.98)	
Education	High	0.92 (0.83, 1.01)	NS
Education	Low	0.89 (0.81, 0.98)	
APOE4	Carrier	0.95 (0.86, 1.04)	NS
APOE4	Non-carrier	0.86 (0.79, 0.95)	
Smoking	Current	0.87 (0.69, 1.09)	NS
Smoking	Former	0.94 (0.85, 1.04)	
Smoking	Never	0.88 (0.80, 0.97)	

Abbreviations: CI, confidence interval; HR, hazard ratio; NS, non-significant; *: The metabolic signature consists of the 26 identified metabolites presented in Supplementary Table 7, [1]:HR and 95% CI per SD increment in metabolic signature. Age, BMI, and TDI were dichotomised at the median. Estimates are from multivariable Cox models fitted in multiply imputed data and adjusted for age, sex, BMI, TDI, diabetes, smoking status, education, depression, cancer, cardiovascular disease, and APOE4 carrier status, except when the corresponding variable was used as an effect modifier. Interaction p-values are from likelihood ratio tests comparing models with and without the exposure×modifier term and are adjusted for multiple testing using the Benjamini–Hochberg FDR.

Supplementary Table 10. Sensitivity analyses for associations between metabolic signature with all-cause dementia, Alzheimer's disease and vascular dementia.

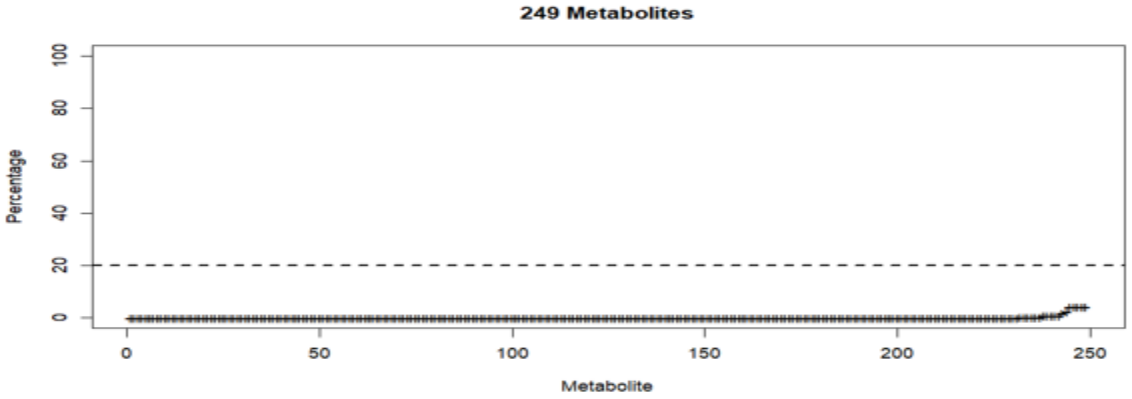
	Metabolic signature*	
	HR (95% CI)[1]	**P-value**
Sensitivity analysis 1		
All-cause dementia (N = 474)		
Age- and sex- adjusted model	0·94 (0·85, 1·03)	$1{\cdot}79 \times 10^{-1}$
Multivariable-adjusted model[2]	0·97 (0·88, 1·08)	$5{\cdot}94 \times 10^{-1}$
AD (N = 160)		
Age- and sex- adjusted model	1·06 (0·90, 1·25)	$4{\cdot}93 \times 10^{-1}$
Multivariable-adjusted model[2]	1·04 (0·87, 1·25)	$6{\cdot}41 \times 10^{-1}$
VD (N = 73)		
Age- and sex- adjusted model	0·93 (0·73, 1·18)	$5{\cdot}58 \times 10^{-1}$
Multivariable-adjusted model[2]	1·09 (0·84, 1·41)	$5{\cdot}34 \times 10^{-1}$
Sensitivity analysis 2		
All-cause dementia (N = 973)		
Age- and sex- adjusted model	0·87 (0·82, 0·93)	$4{\cdot}91 \times 10^{-5}$
Multivariable-adjusted model[2]	0·91 (0·85, 0·98)	$8{\cdot}76 \times 10^{-3}$
AD (N = 345)		
Age- and sex- adjusted model	0·89 (0·80, 0·99)	$4{\cdot}17 \times 10^{-2}$
Multivariable-adjusted model[2]	0·89 (0·79, 1·01)	$6{\cdot}79 \times 10^{-2}$
VD (N = 151)		
Age- and sex- adjusted model	0·81 (0·68, 0·95)	$1{\cdot}24 \times 10^{-2}$
Multivariable-adjusted model[2]	0·94 (0·79, 1·13)	$5{\cdot}29 \times 10^{-1}$
Sensitivity analysis 3		
All-cause dementia (N = 925)		
Age- and sex- adjusted model	0·87 (0·82, 0·93)	$8{\cdot}73 \times 10^{-5}$
Multivariable-adjusted model[2]	0·91 (0·85, 0·98)	$1{\cdot}23 \times 10^{-2}$
AD (N = 330)		
Age- and sex- adjusted model	0·89 (0·80, 1·00)	$4{\cdot}52 \times 10^{-2}$
Multivariable-adjusted model[2]	0·90 (0·79, 1·02)	$9{\cdot}17 \times 10^{-2}$
VD (N = 138)		
Age- and sex- adjusted model	0·84 (0·71, 1·00)	$5{\cdot}30 \times 10^{-2}$
Multivariable-adjusted model[2]	0·99 (0·82, 1·19)	$8{\cdot}91 \times 10^{-1}$

Abbreviations: CI, confidence interval, HR, hazard ratio, [*]: The metabolic signature consists of the 26 identified metabolites presented in Supplementary Table 7 , [1]: HR and 95% CI per SD increment in metabolic signature, [2]: Based on an age and sex-adjusted model, further adjusted for BMI, TDI, diabetes, smoking status, education level, depression, cancer, cardiovascular disease and APOE4 status.
Sensitivity analysis 1: excluded participants with less than twice dietary assessment (N = 54,693).
Sensitivity analysis 2: excluded participants with less than 1 year of follow-up (N = 92,297).
Sensitivity analysis 3: excluded participants with less than 5 years of follow-up (N = 92,249).

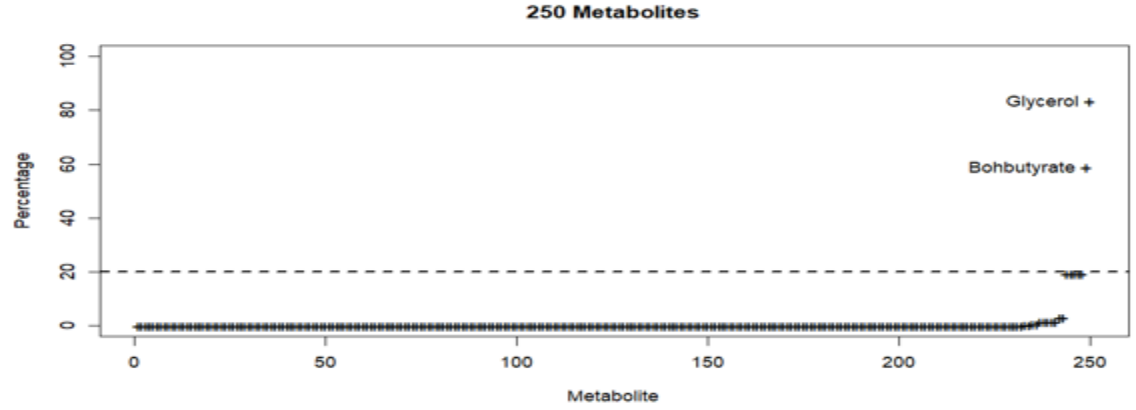


Supplementary Figure 1. Flowchart of participant selection in the Epirus Health Study validation cohort

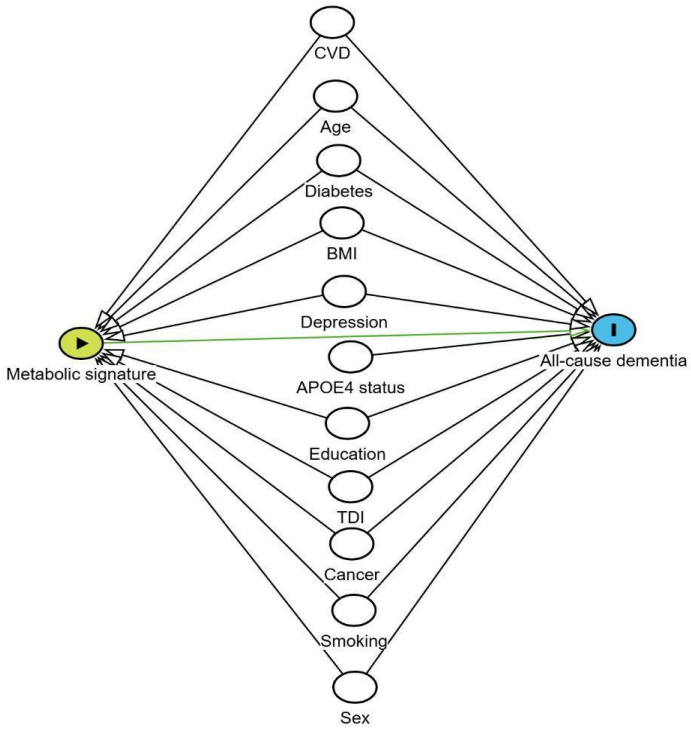
A.



B.



Supplementary Figure 2. A) Percentage of missing values for the 249 metabolites quantified by Nuclear Magnetic Resonance (NMR) for the discovery population (UK Biobank, N = 92,299). B) Percentage of missing values for the 250 metabolites quantified by Nuclear Magnetic Resonance (NMR) for the external population (Epirus Health Study, N =1,608)



Supplementary Figure 3. Directed Acyclic Graph illustrating the causal relationships between Metabolic signature and all-cause dementia. The minimal sufficient adjustment set includes age, sex, BMI, TDI, diabetes, smoking status, education level, depression, cancer, cardiovascular disease and APOE4 status.

BioData Mining

RESEARCH Open Access

The goldmine of GWAS summary statistics: a systematic review of methods and tools

Panagiota I. Kontou[1] and Pantelis G. Bagos[2*]

*Correspondence:
pbagos@compgen.org

[1] Department of Mathematics, University of Thessaly, 35131 Lamia, Greece
[2] Department of Computer Science and Biomedical Informatics, University of Thessaly, 35131 Lamia, Greece

Abstract

Genome-wide association studies (GWAS) have revolutionized our understanding of the genetic architecture of complex traits and diseases. GWAS summary statistics have become essential tools for various genetic analyses, including meta-analysis, fine-mapping, and risk prediction. However, the increasing number of GWAS summary statistics and the diversity of software tools available for their analysis can make it challenging for researchers to select the most appropriate tools for their specific needs. This systematic review aims to provide a comprehensive overview of the currently available software tools and databases for GWAS summary statistics analysis. We conducted a comprehensive literature search to identify relevant software tools and databases. We categorized the tools and databases by their functionality, including data management, quality control, single-trait analysis, and multiple-trait analysis. We also compared the tools and databases based on their features, limitations, and user-friendliness. Our review identified a total of 305 functioning software tools and databases dedicated to GWAS summary statistics, each with unique strengths and limitations. We provide descriptions of the key features of each tool and database, including their input/output formats, data types, and computational requirements. We also discuss the overall usability and applicability of each tool for different research scenarios. This comprehensive review will serve as a valuable resource for researchers who are interested in using GWAS summary statistics to investigate the genetic basis of complex traits and diseases. By providing a detailed overview of the available tools and databases, we aim to facilitate informed tool selection and maximize the effectiveness of GWAS summary statistics analysis.

Keyword: GWAS, Summary statistics, Software, Database, Systematic review

Background

Genome-wide association studies (GWAS) enable the simultaneous testing of thousands of genetic variants, usually SNPs, across the genome in order to find variants associated with a trait or a disease [1]. The GWAS methodology, so far, has generated many robust associations for various traits and diseases and has revolutionized our understanding of the genetic architecture of complex traits. With increasing sample sizes, new sequencing technologies and the accumulation of large biobanks it is expected that our ability to investigate the effects of human genetic variation in complex traits will increase in the

near future [2]. In the first years of the development of the field, efforts were oriented towards the statistical aspects of the analysis [3], which involved thousands of SNPs simultaneously, including the methodology for multiple testing and quality control. This task was successful and enabled the discovery of associations replicated in subsequent studies, and in several cases, validated experimentally and functionally using a wide variety of methods [4]. However, it was soon clear that most variants discovered via GWAS have small overall effects on disease susceptibility [5]. Thus, it became evident that integrating data from multiple sources and developing reliable bioinformatics tools was a necessary step in order to address the complexity of the underlying genetic basis of common human diseases [5].

Soon after the publication of the first GWAS it also became evident that, at least theoretically, individuals could be identified in such cohorts even if only the summary statistics are available [6]. This led to imposing strict control access for sharing individual patients' data (IPD) from GWAS. Subsequent works found that privacy attacks are possible in theory but unsuccessful and unconvincing in real practice. For instance, even sharing 1,000 SNPs for datasets with more than 500 individuals generally leads to a low power of the "attack" [7]. A more thorough investigation is given in [8]. In practice, however, not all studies share their data, at least when it comes to the studies published in the first decade of GWAS. It has been estimated that the proportion is only 13%, which increased from 3% in 2010 to 23% in 2017 [9]. On the contrary, researchers sharing their summary data has been shown to receive on average 81.8% more citations, an effect that probably is related, at least partially, to the usability of the data in downstream analyses [10]. Summary statistics do not only offer the additional protection of privacy, but also offer significant advantages in computational cost when using the data in downstream analyses, which does not scale with the number of participants in the study [11]. Thus, it is of no surprise that during the last years a large variety of methods have been developed to perform a so-called post-GWAS analysis using the summary results of a single study, or of several studies, and in most cases integrating data from other sources [11]. The majority of these methods use the summary data in the form of per-allele SNP effect sizes (log odds ratios or betas) along with their standard errors, or equivalently the z-scores (per-allele effect sizes divided by their standard errors). These methods seek to go a step further from the simple analysis, or re-analysis of a study, and aim to improve our understanding about the functional role of the identified variants [12]. The most important factors that played significant role in the development of such methods, in this so-called post-GWAS era, is the linkage disequilibrium (LD) information from a population reference panel such as HapMap or 1000 Genomes Project, the gene expression variation in the form of eQTL, and the integration of functional information on biological pathways [13–15].

The methods developed so far cover a broad range of different types of analysis, either in the study of a single trait or in the combined analysis of multiple traits. For a single trait, we may have methods for meta-analysis [16, 17], methods for inferring heritability [18, 19], gene-based tests [20], methods for Gene Set (or Pathway) Analysis [21], or methods for fine-mapping causal variants [22]. Regarding the analysis of multiple traits there is also a variety of methods [23], ranging from those that estimate the genetic correlation between traits [24], the joint analysis of multiple traits [25], or the

methods that try to estimate causality between traits such as Mendelian Randomization [26], transcriptome-wide association studies [27], or colocalization [28]. Of course, the data standards [29] used to facilitate these analyses and the databases that the results are stored in, are also of great importance for the community.

In order to provide a comprehensive overview of the currently available software tools and databases for GWAS summary statistics we performed a systematic review following the PRISMA guidelines [30]. We conducted a comprehensive search of the literature to identify relevant software tools and databases. We categorized the tools and databases by their functionality, in categories related to data, single-trait analysis, and multiple-trait analysis, along with their sub-categories mentioned in the previous paragraph. We also compared the tools and databases based on their features, limitations, and user-friendliness. Our review identified a wide range of software tools and databases for GWAS summary statistics analysis, each with unique strengths and limitations. We provide descriptions of the key features of each tool and database, including their input/output formats, data types, and computational requirements. We also discuss the overall usability and applicability of each tool for different research scenarios. This comprehensive review will serve as a valuable resource for researchers who are interested in using GWAS summary statistics to investigate the genetic basis of complex traits and diseases. By providing a detailed overview of the available tools and databases, we aim to facilitate informed tool selection and maximize the effectiveness of using GWAS summary statistics.

The systematic review

In order to collect all the available published papers, we performed a systematic review of the literature following the PRISMA guidelines [30]. The search was performed in PubMed (https://pubmed.ncbi.nlm.nih.gov) with the following query: *("Summary Statistics" OR "Summary Data" OR "Summary Association Statistics" OR "Summary Association Data") AND (GWAS OR genomewide OR genome-wide*). The abstracts initially, and then the full articles were scrutinized in order to collect the necessary information. The inclusion criteria state that methods, software tools and databases, suitable for the analysis of GWAS summary data are suitable for inclusion. Methods papers that do not report software, or software pages not currently available are excluded. Additional searches were performed in the reference lists of the identified articles in order to identify additional studies that were missing. In many cases multiple articles regarding a single tool were found, so we kept only one. We decided to include reports deposited in preprint servers like medRxiv/bioRxiv, but some of these papers were eventually published in peer-review journals, so in such cases we retained only the latter reference. Tools regarding Polygenic Risk Scores (PRSs) and visualization were excluded. For all included tools we recorded the URL, the PMID, and the main functionality/es along with comments regarding its main methodological features. The initial search identified 2942 articles (22/12/2023).

In total we identified 305 tools and databases (Fig. 1). We classified them in three broad categories: *data*, *tools for single traits* and *tools for multiple traits*, along with the various sub-categories. The total breakdown is given in Table 1. Several tools may perform different tasks and thus they can be considered for more than one category;

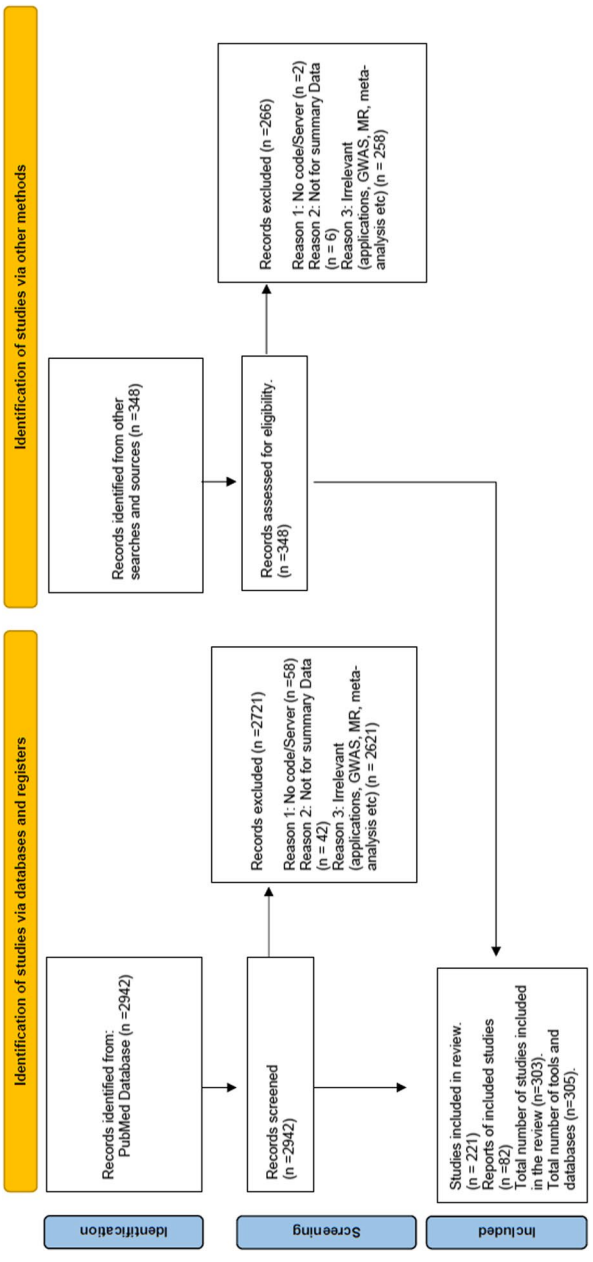


Fig. 1 PRISMA Flow diagram for systematic review

Table 1 The broad categories and the sub-categories of tools and databases

Category	Type	N	
Data	Database	17	5.57%
	QC	13	4.26%
	Reconstruction	6	1.97%
	Imputation	10	3.28%
Single trait	Meta-analysis	29	9.51%
	Heritability	18	5.90%
	Gene-based tests	30	9.84%
	GSA	29	9.51%
	Fine-mapping	25	8.20%
Multiple trait	Genetic correlation	14	4.59%
	Pleiotropy	38	12.46%
	MR	31	10.16%
	Colocalization	16	5.25%
	TWAS	29	9.51%

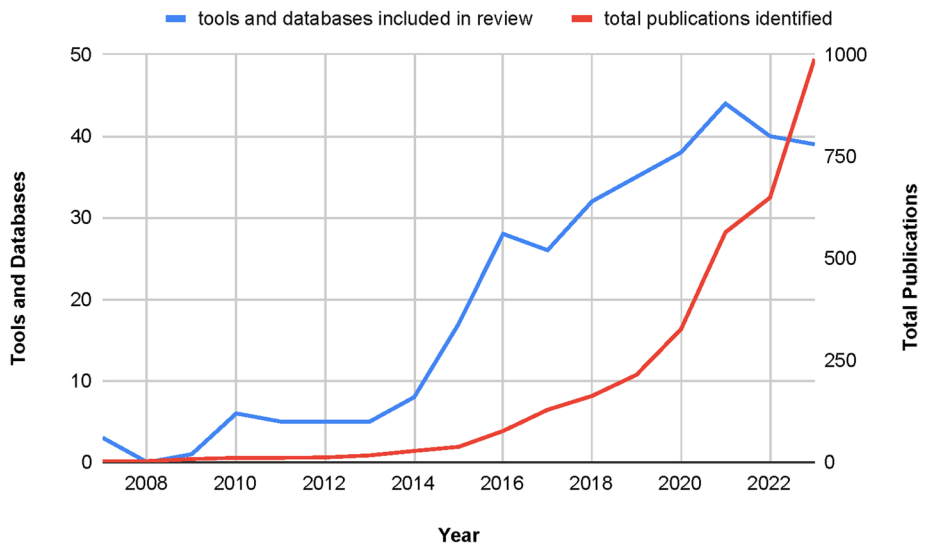


Fig. 2 Number of Tools and Databases included in the review Published Per Year

so, we classified them to the one most closely related to the primary goal of the analysis they claim to perform. Other tools do not fit exactly to the general description of the category, but we nevertheless classified them to the most similar one. The largest sub-category consists of the tools for pleiotropy analysis, whereas the smallest one is related to reconstruction of genotypes and effect sizes. Most tools are written in R (56.4%) with the largest proportion being in the multiple traits category, followed by Python (12.5%) and C/C++ (8.2%) (Fig. 2). Apart from the publicly available databases only a handful of tools are offered as webservers (6.95%). Most of the tools were published after 2015 (Fig. 3). Nearly 60% of the tools and databases were published in: Bioinformatics, American Journal of Human Genetics, Nature Genetics, Nature Communications, Nucleic Acids Research and PloS Genetics (Fig. 4). In the following

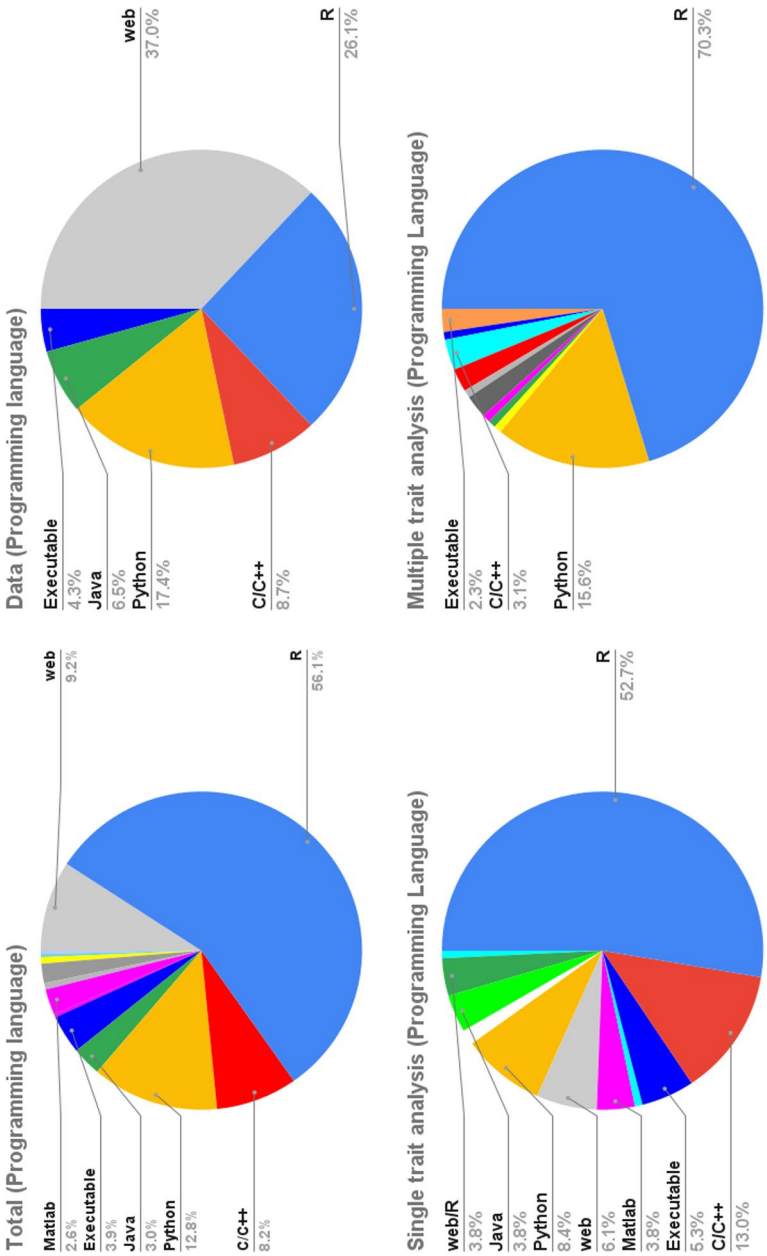


Fig. 3 The programming languages used in the various categories of identified tools

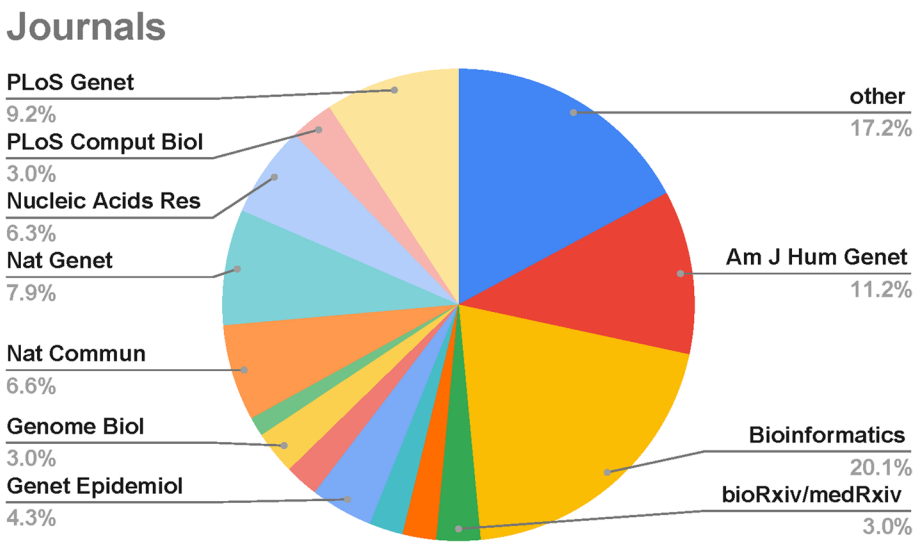


Fig. 4 The journals in which the studies including in the review were published

sections we proceed with the detailed description of the various tools identified, classified in the different categories and sub-categories. The complete list of identified tools along with the relevant information is given in Supplementary Table 1.

The data

Firstly, we are going to present the tools dedicated to the data themselves. We include here tools for *quality control* of GWA summary statistics, tools for *imputation and genotype reconstruction* as well as the publicly available *databases* of summary results.

Standards and quality control

The need for sharing and re-using GWAS summary statistics has been an issue for the community during the last years. Generally, it is acceptable that the minimum information ("mandatory") contained in GWAS summary statistics should include: the chromosome and the base-pair location, the *p*-value of the association, the risk allele and the other allele, the risk allele frequency, and an estimate of the effect size (odds ratio or beta) along with its standard error [29]. Other important summary statistics that nevertheless termed as "encouraged" ones include the sample size, the variant ID, the rsID, the confidence interval of the effect size and so on. Such specifications were considered for the GWAS-SSF format [31], which was developed to meet the requirements settled by the community. GWAS-SSF consists of a tab-separated data file with well-defined fields and an accompanying metadata file. Most repositories and programs use some variant of the GWAS-SSF. However, such tabular formats in several cases lead to ambiguity or incomplete storage of information, or other times lack essential metadata. This leads to poor performance and increased risk of possible errors in downstream analyses. To address these issues, an adaptation of the well-known variant call format [32] was developed, capable of storing GWAS summary statistics which was called GWAS-VCF along with software tools to apply it in downstream analyses [33]. The VCF contains a file header with metadata and a main file containing variant-level (one locus per row

with one or more alternative alleles/variants) and sample-level (one sample per column) information. This way, the VCF was adapted to include GWAS-specific metadata utilizing the sample column to store variant-trait association data. The GWAS-VCF is the standard used by the MRC-IEU OpenGWAS database [34] and it comes with appropriate tools to map GWAS summary statistics to VCF with on-the-fly harmonization (https://github.com/mrcieu/gwas2vcf).

Despite these efforts, not all available data are in line with the standards, especially when dealing with data from older studies. Thus, there is a need for additional tools to harmonize the data, as well as to identify and correct errors. Tools belonging to the former class were developed early and were focused mainly on harmonizing data in preparation of a meta-analysis. These include QCGWAS [35], GWAtoolbox [36] and EasyQC [37]. GEAR [38] is very interesting in that it incorporates ideas from population genetics which allow verification of the genetic origin and geographic location of each cohort and identifying significant sample overlap. More recent tools like MungeSumstats [39] and GWASlab [40] perform standardization and quality control handling the most common formats, SumStatsRehab [41] can be used for data validation, restoration of missing data, correction of errors or formatting, and GWASinspector [42] provides extensive QC reports and perform harmonization being compatible with recent reference panels and by handling insertion/deletion and multi-allelic variants. The latter class of methods, additionally, leverages information from the LD among SNPs. One such tool is GQS [43] which identifies suspicious regions and prevents erroneous interpretations by comparing the significance of the association for each SNP to its LD value for the reported index SNP. Similar functionalities are offered by DENTIST [44] which uses LD to detect and eliminate errors and disagreements between GWAS data and the LD reference panel. EXTminus23andMe [45] evaluates the quality of summary statistics after data removal and the suitability of the down sampled summary statistics for typical follow-up genetic analyses.

Databases

The publicly available biological databases played and continue to play a central role in bioinformatics and in biological research in general [46–48]. The same is the case for databases related to human research [49] and in particular those involved in GWAS [50]. The databases we identified can be roughly divided in two categories: databases that contain summary statistics from GWAS and databases that contain important secondary analyses on those data with some of the methods that we will describe in later sections.

Regarding the databases of the first category, NCBI's dbGAP [51] was developed to contain the results of studies investigating the interaction of genotype and phenotype, which include GWAS. One of the dbGAP's primary objectives was to house individual level GWAS data, but the database also contains summary data as well. Summary statistics are generally available to the public, whereas access to IPD requires varying levels of authorization. The NHGRI-EBI GWAS Catalog [52], which was established in 2008 is considered for years the central repository of GWAS summary statistics. It is a high-quality curated collection of all published GWAS and as of 2023–12-20, contains 6,680 publications, 566,798 top associations and 66,825 full summary statistics (Fig. 5). The database played an important role in the community efforts leading to the development

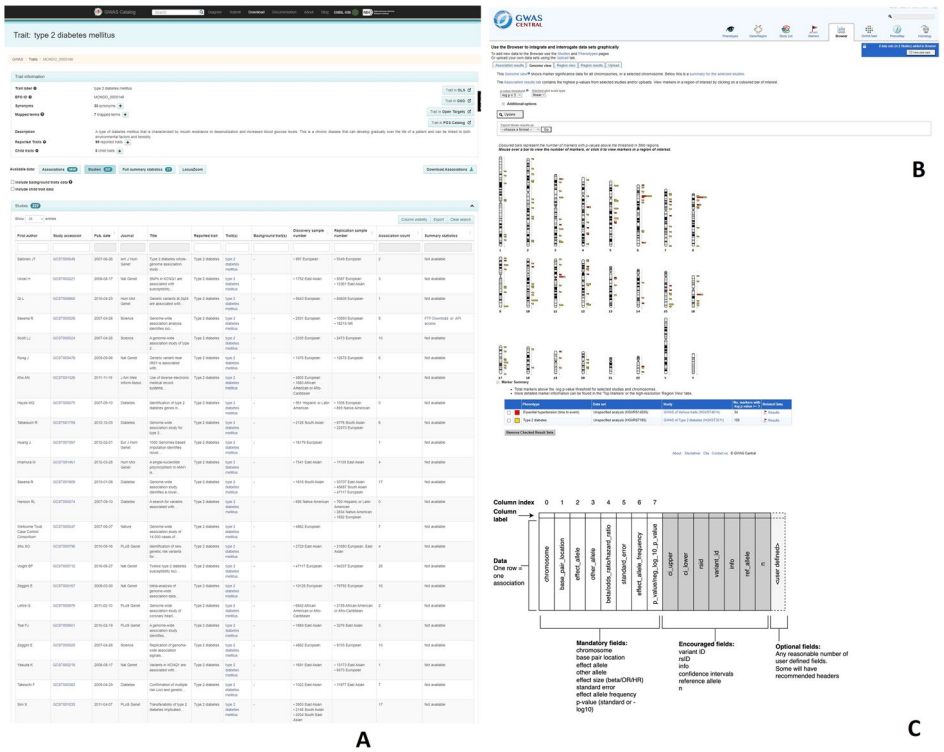


Fig. 5 A snapshot of the data. **A** A view of the Type 2 Diabetes Mellitus studies deposited in NHGRI-EBI GWAS Catalog. **B** Type 2 Diabetes Mellitus studies contained in GWAS Central, depicting the significant hits in the chromosomes. **C** The SFF format

of GWAS-SSF format. GWAScentral [53] previously known as the Human Genome Variation (HGV) database of Genotype-to-Phenotype information is a database that contains over 72.5 million *P*-values for over 5,000 studies, with over 7.4 million unique genetic markers involved in more than 1,700 unique phenotypes. The database contains data from several sources (including NHGRI-EBI GWAS Catalog, OpenGWAS, Japanese GWASdb, dbGaP, WTCCC and so on). The IEU MRC OpenGWAS [34] is a new addition and contains 346 million genetic associations from 50,037 GWAS summary datasets. It contains complete data from various consortia and the UK Biobank and comes with a lot of tools for harmonizing the data and storing them in the GWAS-VCF format. At the time of writing there are 4,126 binary traits, 725 metabolites, 3,371 proteins, 3,143 brain imaging phenotypes, and 3,217 other continuous phenotypes. In addition to the complete GWAS summary data, it also contains independent top hits for every dataset, totaling 116,918 independent signals in which 7,109 datasets have at least one hit. GeneATLAS [54] and GBE [55] contain associations from the UK Biobank cohort. GeneATLAS currently contains data for 452,264 individuals, 778 traits and 30 million variants, whereas GBE contains summary statistics from over 750,000 individuals combining data from the UK Biobank, the Million Veterans Program and the Biobank Japan. GTEx [56] and QTLbase [57] are the primary resources for xQTL data. The GTEx project has been expanded over time, and currently contains data of genetic associations for gene expression and splicing in 838 individuals in 49 tissues. QTLbase, similarly, contains genome-wide QTL summary statistics for many molecular traits across 95 tissue/

cell types and multiple conditions. Contains tens of millions of significant genotype-molecular trait associations under different conditions. Other resources of this category, related to various large consortia (GIANT, WTCC, PGC etc.) as well as other biobanks (FinnGen etc.) can be found in Supplementary Table 2.

The second category contains databases of important secondary analyses performed on GWAS summary statistics with some of the methods that we describe in detail in later sections, such as gene-based tests, heritability analysis, TWAS, colocalization and so on. TSEA-DB [58] and PCGA [59] use information from gene-expression in various tissues to perform tissue or cell-type enrichment analysis of the GWAS association statistics. webTWAS [60] and COLOCdb [61] also use information on eQTL but in different fashion. webTWAS currently contains data for over 1,389 full GWAS for which it calculates the causal genes using single tissue expression imputation (using MetaXcan and FUSION), or cross-tissue expression imputation (using UTMOST). COLOCdb on the other hand is the most comprehensive colocalization analysis by integrating publicly available GWASs with different types of xQTL and different algorithms (COLOC, SMR). GWAS ATLAS [62] contains results of 4,756 GWAS from 473 unique studies across 3,302 unique traits accompanied by useful information obtained from downstream analysis. Each study is accompanied by MAGMA results (see also "gene-based tests"), SNP heritability estimation and genetic correlations with other traits in the database. GWASROCS [63], on the other hand, contains a large and comprehensive set of SNP-derived AUROCs and heritabilities. Currently includes 579 simulated populations (corresponding to 219 traits) and SNP data (odds ratio, risk allele frequency, and p-values) for 2,886 unique SNPs. Phenome-wide association studies (PheWAS) invert the idea of a GWAS by searching for phenotypes associated with specific variants across the range of thousands of human phenotypes, or the "phenome [64–66]. Thus, it is expected that a PheWAS will need large databases of GWAS results. PhenoScanner [67] is the most complete such database with publicly available results from over 65 billion associations and more than 150 million unique genetic variants. Similar functionalities are offered also by OpenGWAS, GWAS ATLAS and PheWAS Catalog [68]. Lastly, we need to mention LD Hub [69], a centralized database of publicly available GWAS results for 173 diseases/traits which offers a web interface that automates the LD score regression (LDSC) analysis pipeline (see also "Genetic correlation").

Imputation and genotype reconstruction

Although some of the methods for quality control mentioned previously can correct errors and alter the data, the methods used for imputation go one step further. As expected, imputation methods were developed initially for individual data for handling studies genotyped with different platforms [70–72]. Such methods can infer missing genotypes using LD information from reference samples genotyped using denser arrays or sequencing. Genotype imputation increases the coverage of SNPs and thus can be used to increase statistical power, increase the accuracy of fine-mapping and harmonize the data in order to facilitate meta-analysis [70]. Several factors can influence the imputation accuracy: the sample size, the suitability of the reference panel for the particular sample, the genotyping chip and the allele frequency [71]. In general, however, these methods are time-consuming since they process individuals one at a time, and

thus methods that impute directly the summary statistics were developed. These methods utilize only the information provided in the sample regarding the studied population (*p*-value, z-score or odds-ratio/beta) and require additional information regarding the LD structure. Nearly all methods perform a kind of multiple regression assuming the multivariate normal distribution for the test statistics and utilizing the theoretical result pointing that the correlation of such test statistics equals the correlation of the corresponding variables [73], that is the genotype correlation, available through the reference panel. Such methods include FAPI [74], ImpG [75], RAISS [76], DIST [77] and SSimp [78] with most of the differences lying in the choice of the reference panel and the exact details of the mathematical methods used to handle matrix inversions in the multivariate normal. DISSCO [79] uses a similar framework but allows for covariates. Such methods may perform poorly in cases where the sample has a different LD structure compared to the reference panel. Thus, extensions such as DISTMIX [80] and ARDISS [81] were developed to handle mixed ethnicity cohorts, improving the imputation performance. Adapt-Mix [82] estimates the correlation structure in both admixed and non-admixed individuals using simulated and real data and allows the use of this matrix with other imputation methods. Other methods such LS-meta [83] and LSimputing [84] offer additional advantages; LS-meta imputes both genetic and environmental components using information from additional omics-trait association summary data, whereas LSimputing implements a non-parametric method that allows for nonlinear SNP-trait associations and predictions in case a sample of IPD is available. Using the same principles, simGWAS [84] allows simulation of whole GWAS summary data, without generating individual data as an intermediate step.

Genotype reconstruction methods take a different approach. Given the summary statistics for a SNP (either directly measured or imputed), one can reconstruct the genotype counts that produced it. This will offer many advantages, since with the reconstructed genotypes the researchers could perform additional analyses using other statistical methods suitable for grouped data and test different hypotheses [85]. For instance, one can calculate grouped Polygenic Risk Scores (PRS) [85], perform logistic regression for grouped data [85, 86], perform multivariate meta-analysis [87], or implement robust tests for association that is expected to work better when the underlying model of inheritance deviates from the additive which is usually assumed [88, 89]. The details and the success of the reconstruction depend heavily on available summary statistics. As one can easily understand, *p*-values and z-scores cannot be used, and one must rely on available effect sizes such as the odds ratio (OR). When the OR, the standard error and the sample size is given, methods are available in epidemiology that allow the reconstruction of the allelic 2X2 table [90]. If z-scores, confidence intervals or *p*-values are available one can use them to obtain the standard error. React [85] uses an equivalent method relying on solving a system of nonlinear equations. If the allele frequency in one group (usually the controls) is also known, the allelic counts may easily be obtained with a simple calculation. In all cases the accuracy of the reconstruction may depend on the precision of the available summary statistics. After the allelic 2X2 table is reconstructed, it is straightforward to obtain the genotype counts, assuming HWE (which as one might expect adds another source of potential bias). MetaSustract [91] is a tool that recreates analytically the results of the validation cohort from meta-analysis summary statistics, allowing the

researchers to compute meta-analysis summary statistics that are independent of the validation cohort, without requiring access to the IPD. Spkmt [92] works in similar fashion but in families; it can be used to derive the summary statistics of one parent from the data of the offspring and the other parent. Finally, we need to mention two tools that work in somewhat different modes. OATH [93] is used to reproduce reported results from a GWAS and recover underreported results from other alternative models with a different combination of nuisance parameters, whereas LMOR [94] performs transformations from the genetic effects estimated under the Linear Mixed Model to the Odds Ratio that only rely on summary statistics.

Analysis of a single trait

In this section we are going to present the various types of methods and tools dedicated to the analysis of a single trait. These include tools for *meta-analysis*, tools for the estimation of *heritability*, tools for implementing *gene-based tests*, *gene set* methods and *fine mapping* methods.

Meta-analysis

One of the most obvious uses of GWAS summary data is to combine them and perform a meta-analysis. Meta-analysis is the statistical procedure used to combine evidence from multiple studies in order to increase statistical power and it is a methodology widely used in medical research for decades [95]. A meta-analysis can be performed with various methods [16] using IPD or summary data; the former offers many advantages, but the latter is far more easy to be performed taking into account the various restrictions imposed on sharing GWAS IPD and the difficulties in the logistics of such a project [17]. Moreover, given the large samples usually encountered in GWAS it has been shown, both theoretically and empirically, that meta-analysis using summary statistics has the same efficiency as the joint analysis of IPD [96]. A compromise between these two extremes arises when a research group has access to individual-level genotype data of a limited sample size and wants to integrate these with existing summary data available in the databases. Such methods are in use in epidemiology for years [97] and several tools have been developed especially for handling GWAS data, for instance IGESS [98], metaGIM [99] and LEP [100]. PolyGIM [101] can be applied with or without IPD and uses polytomous logistic regression to investigate disease subtype heterogeneity in situations when only summary data is available.

Regarding summary-data meta-analysis of GWAS, the most commonly used methods includes standard methods, such as combining *p*-values, z-statistics or effects sizes like Odds Ratio (for binary traits) or mean differences (for continuous traits) using fixed or random effects models [16, 102]. These statistical methods are straightforward to implement, and are available in general purpose statistical packages such as STATA and R. However, there are several specialized tools that facilitate the process and provide integration with useful bioinformatics or visualization functions. Such widely used tools include METAL [103], GWAMA [104] and PLINK [105]. Other tools are oriented to more specialized cases offering advanced options. For instance, YAMAS performs meta-analysis including missing SNPs identified with LD without performing imputation [106] and rareMETALS [107] uses a partial correlation

based score to perform meta-analysis in the presence of large amounts of missing values. There is also a class of tools which focus on the replication of GWAS and the combined analysis of data from primary and replication studies. Such tools include rfdr [108] and Jlfdr [109] which control for False Discovery Rate (FDR), Rrate [110], which determines the sample size of the replication study and checks the consistency between the primary and the replication study, and MAJAR [111] which jointly test prognostic and predictive effects in meta-analysis without the need of using an independent cohort. metaGAP [112] is an online tool for calculating the statistical power of a meta-analysis of GWAS (Fig. 6). METACARPA works with overlapping or related samples, even when details of the overlap or relatedness are unknown [113], MAGENTA [114] performs meta-analysis with gene set enrichment analysis (GSEA), whereas GWASmeta [115] and MetABF [116] work in a bayesian framework calculating the Approximate Bayes Factor (ABF). Other tools offer more advanced options such as meta-analysis with multiple traits (see also "multiple traits"), like nGWAMA [117], metaCCA [118], CPASSOC [119], metaUSAT [120] and CPBayes [114] (and its extension GCPBayes [121]), and others are designed for meta-analysis under different genetic models, like GWAR [89] which uses robust methods (like MIN2 or MAX) in order to handle the uncertainty in the underlying genetic model, or like the simulation tool [122] which implements an alternate strategy for the additive genetic model simulating data for the individual studies. Finally, we need to mention sPLINK [123] which performs privacy-aware GWAS on distributed datasets, and XPEB [124] which is an empirical Bayes approach designed to improve the power GWAS in minority

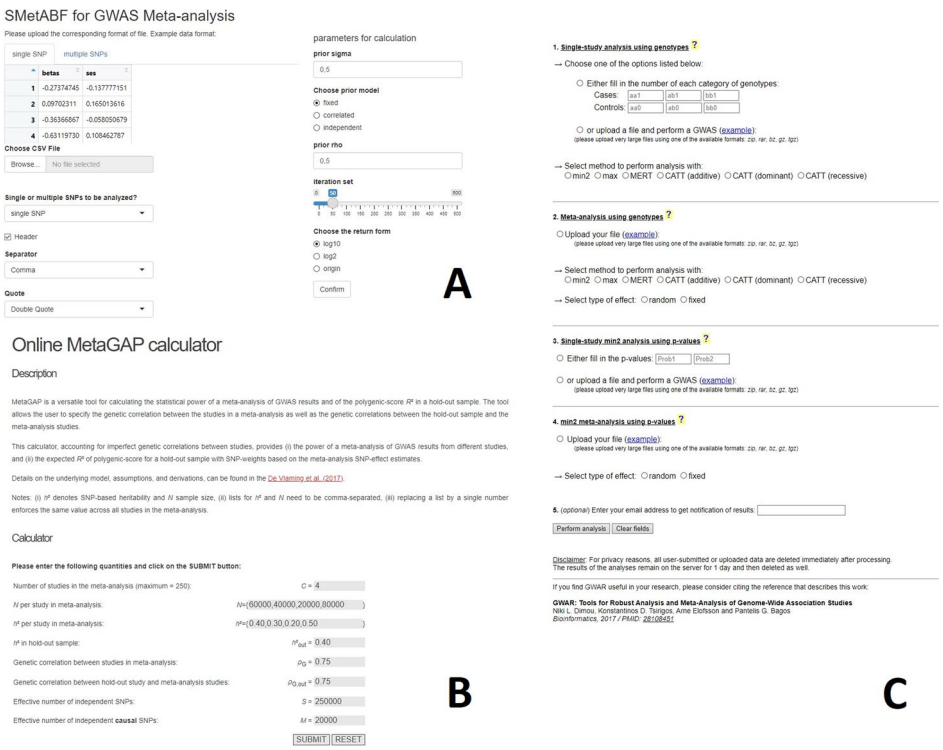


Fig. 6 Tools for meta-analysis. **A** GWASmeta (SMetABF) for performing Bayesian meta-analysis. **B** The MetaGAP power calculator. **C** GWAR for robust analysis and meta-analysis of GWAS

populations by exploiting information from GWASs performed in populations of different origin.

Inferring heritability

Heritability is generally defined as the fraction of phenotypic variation explained by genetic variation. Heritability is a dimensionless parameter of the population, and it was introduced by Sewall Wright and Ronald Fisher in the previous century. Traditionally, heritability is estimated using family-based designs such as twin studies. However, there are controversies regarding the various methodologies for estimation and interpretation of the results [125]. Despite all these, heritability is an important aspect of research in modern genetics, and regarding the prediction of disease risk from genomic data [126]. The technological advancements have facilitated the development of methods that use large samples of unrelated, or related, individuals. Thus, family-based designs using genomic data (trio-genome-wide complex trait analysis, and so on) have emerged. Such methods are discussed and compared in [127]. Of course, heritability can also be estimated via the results obtained in a traditional GWAS using unrelated individuals. The gap between these estimates and those obtained from classical heritability estimation methods has been termed the "missing heritability problem" and it is an important open question in current research [128]. Recent reviews of the methods that use GWAS data, are given in [18, 19] focusing on their modeling assumptions, their similarities, and their applicability.

One of the first and simplest methods to calculate heritability from allele frequency, odds ratio and prevalence of the disease was implemented in the SumVg package [129]. This method, however, utilizes only the significant SNPs. The same authors extended the method later in order to allow calculation using the z-statistics from the whole GWAS sample [130]. A disadvantage of this method is that LD is not taken care of, and highly correlated SNPs need to be filtered manually. AVENGEME [131] is a tool that treats causal effect sizes as fixed effects and models the genotypes as random correlated variables. HESS [132] which was presented later built upon the same ideas and can be viewed as a weighted sum of the squares of the projection of effect sizes onto the eigenvectors of the LD matrix at the particular locus, with weights inversely proportional to the corresponding eigenvalues. LD Score Regression (LDSC) has been frequently applied to summary statistics from GWAS and one of its functionalities is to estimate the SNP heritability of a trait [133]. LDER [134] extends LDSC making full use of the information from the LD matrix providing more accurate estimates, whereas s-LDSC [135] is an extension suitable for partitioning heritability. SumHer [136] presented later and offers the same functionalities, with the main difference being that it allows for different so called “heritability models”. According to these, a SNP with high MAF is expected to contribute more to the total heritability compared to one with low MAF, whereas on the other hand, a SNP in a region of low LD is expected to contribute more compared to one in a region of high LD. On the contrary, LDSC estimates are obtained by assuming that all SNPs contribute equally. HEELS [137] is a new tool using REML to produce accurate and precise local heritability estimates and RSS, is a multiple regression-based fine-mapping tool (see “Fine-mapping”), can also calculate SNP heritability from the regression model. VarExp [138] and GxESum [139] are methods for

estimating the phenotypic variance explained by genome-wide gene-environment (GxE) interactions. There are also tools like GWIZ [63] and SummaryAUC [140] that calculate the Receivers Operator's Characteristic (ROC) curve and the associated Area Under the Curve (AUC). GWIZ generates ROC curves and the AUC using simulations and then estimates heritability using the square of the Somers' rank correlation D. SummaryAUC on the other hand approximates the AUC of a PRS and its variance. HAMSTA [141] is a tool that, among others, estimates heritability explained by local ancestry using data from admixture mapping studies. Estimating the Effect size distribution is also a related important concept. GENESIS [142] uses LD and a Likelihood-based approach to estimate effect-size distributions. It also allows predictions regarding yield of future GWAS with larger sample sizes. GWEHS [143] calculates the distribution of effect sizes of SNPs, as well as their contribution to trait heritability. Furthermore, it performs predictions for the change in the effect size as well as in the heritability when new variants are identified. FMR [144] is a method-of-moments for calculating the effect-size distribution and GWAS-Causal-Effects-Model [145] is a random effects model for estimating the causal variants and their effect size distribution. Finally, there are tools to implicate gene-expression in heritability analysis: MESC [146] which estimates the proportion of heritability mediated by gene expression levels using linkage disequilibrium (LD) scores and eQTL, and GCSC [147] which uses results from a TWAS (see "TWAS and Colocalization") in the so-called gene co-regulation score regression, to identify gene sets enriched for disease heritability.

Gene-based tests

Historically, association tests are oriented towards single variants, and this was the case for both traditional association studies as well as for GWAS. However this approach has some limitations that were noted earlier and a call for a shift towards gene-based tests was made [148]. Gene-based tests aggregate individual variant associations within a gene, providing a more comprehensive assessment of the gene's overall contribution to a trait or disease. This approach helps prioritize genes with multiple associated variants, enhancing the biological relevance of findings, and it has proven to be useful particularly in case of low frequency variants [148]. There are plenty of different methods for combining the association statistics or p-values within a gene, ranging from simple Fisher's method or the minimum p-value approach, to more advanced methods like the Burden Test (BT) [149] or quadratic tests like SKAT [150] with variations in power [151]. Nevertheless, there is a consensus regarding the importance of incorporating LD information of the nearby variants into the methods for controlling the type I error rate at the desired level [20].

VEGAS, GATES, fastBAT and GCTA are among the oldest tools available for summary data, which remain efficient and widely used. SKAT (Sequence Kernel Association Test) is a well-known regression method for testing association between variants and traits adjusting for covariates. As a score-based variance-component test, it calculates p-values analytically by fitting the null model containing only the covariates [150]. The original SKAT method uses only IPD, but later implementations like metaSKAT or SKAT-O have been extended to handle summary data. GCTA and VEGAS also use the multivariate normal framework adjusting the estimates for LD using a reference panel

[152, 153]. Of note, GCTA also offers methods for conditional analysis (see "Fine mapping"), and same also holds for KGG [154], whereas VEGAS's new version allows for mixed ethnicity populations. GATES [155], on the other hand, uses an extended Simes procedure that integrates functional information and association evidence to combine *p*-values, whereas fastBAT [156] offers fast analytical *p*-value computations. The gene analysis in MAGMA (Multi-marker Analysis of GenoMic Annotation) is based on a multiple linear principal components' regression model to account for LD and uses an F-test to compute the overall gene *p*-value [157]. Its extension, nMAGMA, extends the lists of genes that can be annotated by integrating local signals, long-range regulation signals, and tissue-specific gene networks. It also provides tissue-specific risk signals, which are useful for understanding disorders with multi-tissue origins [158]. H-MAGMA [159] and eMAGMA [160] are two other extensions. The former integrates 3D chromatin configuration, whereas the latter leverages significant tissue-specific cis-eQTL information to assign SNPs to putative genes. EPIC [161] and GAMBIT [162] also utilize functional data for gene-based analysis; the former using cell-type-specific gene expression data obtained from single-cell RNA sequencing and the latter using coding and tissue-specific regulatory annotations. Such methods share several features in common with TWAS methods (see respective section). AgglomerativLD [163] also captures LD between SNPs of nearby genes, which induces correlation of the gene-based test statistics. DOT [164] is one of the few methods that applies a decorrelation-based approach before combining SNP-level statistics or *p*-values. Tools like GPA [165], oTFisher [166], TS [167] and aSPU [168] implement some type of so-called adaptive tests (AT), that is, they account for possibly varying association patterns across SNPs, whereas some modern tools like MKATR [169], COMBAT [170], MCA [171], OWC [172], FST [173], ACAT [174], HYST [175], GBJ [176] and sumFREGAT [177] perform analysis with multiple statistical methods and test and combine the results. Notably, tools like aSPU [168], snpGeneSets [178], Pascal/PascalX [179, 180], MAGMA, chromMAGMA [181] and FUMA [182], also offer the option of performing gene-set analysis after performing the gene-based analysis (see next section), whereas HSVS-M [168, 183] tests the association of a gene with multiple correlated traits.

Gene Set analysis

Gene set analysis (GSA), or Pathway Analysis, extends the concept of gene-based methods by jointly analyzing groups of functionally related genes and identifying biological pathways enriched with trait-associated genes. By considering the collective impact of multiple genes within a pathway, researchers can obtain a clearer picture of the underlying biological mechanisms influencing the phenotype under investigation. The first applications of such methods borrowed ideas from the microarray data analysis literature, and since then they became widespread in analysis of GWAS [184]. Any GSA method needs to address some issues. Firstly, how to handle SNPs of the same gene; secondly, how to define the appropriate gene-set or pathway, and finally how to combine the effects from multiple SNPs/genes within the same set/pathway [185]. Thus, the choices made by different methods can be very diverse leading to a wide variety of different approaches. For instance, some methods operate with SNP-level statistics (effect sizes, z, or *p*-values) assigning the SNP to the closest gene (usually within a range of ± 20 K

bases), whereas others take as input a gene-level statistic or simply a gene list obtained by a gene-based method (of course, several tools allow for both a gene-based and a GSA approach). Regarding the choice of set there is a plethora of databases containing biological pathways (KEGG, PANTHER etc.), or other types of gene-set representation like PPI interactions, ontologies and so on [186]. Finally, regarding the statistical method used to aggregate evidence there is also a wide range of different methods that handle with different approaches the gene set size and gene length, the LD patterns and the presence of overlapping genes within pathways, or apply different statistical approaches such as those using the so-called competitive null hypothesis, or those using the self-containing one [14, 187]. A tutorial regarding the use of such methods is given in [21].

Among the most easily used and frequently cited are the tools that utilize a webserver. FUMA [182] and iGSE4GWAS [188] are tools specialized in GWAS and use SNP-level statistics as inputs, differing in the subsequent analyses: FUMA uses MAGMA for gene-based testing and allows for ORA and Kologorov-Smirnov test (GSEA), whereas iGSE4GWAS maps the most significant SNP to a gene and then performs an improved GSEA with label permutation to obtain accurate p-values. Tools like Enrichr [189], g:Profiler [190], DAVID [191], WebGestalt [192] and PANTHER [193] are general purpose enrichment tools that provide functionalities for different types of omics data (Fig. 7). They accept gene or SNP-list as input and provide Application Programming Interface (API) ensuring interoperability, whereas for the statistical analysis they all use some version of ORA and/or GSEA (WebGestalt also uses Network Topology-based Analysis). A major feature of these tools is that they incorporate a large number of biological and pathway databases, with g:Profiler and Enrichr offering the most complete collection. GSA-SNP2 is one of the first methods to be developed for GWAS and has seen several improvements regarding the calculation of the combined gene score and the execution time, being among the fastest methods [194]. aSPUpath2 [195] and GIGSEA [196] are two methods that integrate expression data (eQTL) in the pathway analysis. The former uses an adaptive test that extends the aSPU methodology based on

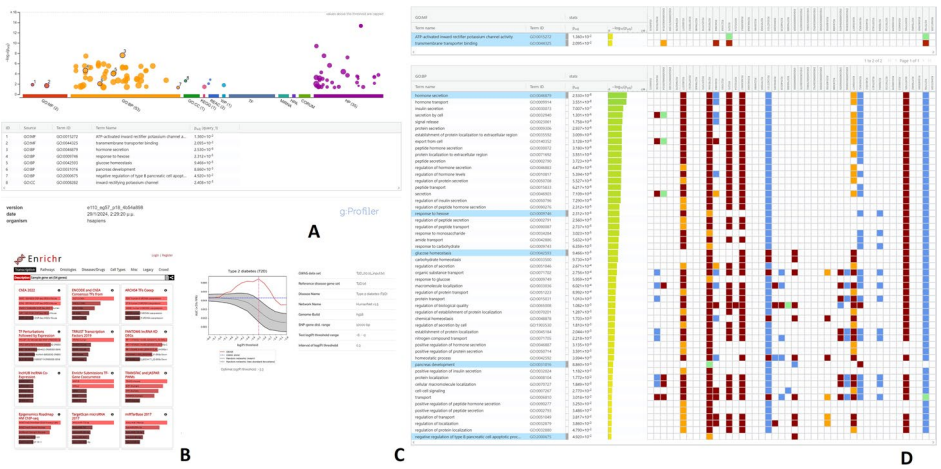


Fig. 7 Enrichment. **A** Summary view in g:Profiler of the significant SNPs for Type 2 Diabetes Mellitus. **B** Enrichr results for the same set. **C** Output of GWAB for Type 2 Diabetes Mellitus SNPs. **D** Detailed results from g:Profiler

chi-square, whereas the latter uses a regression-based approach coupled with permutations to calculate accurate p-values. In a similar fashion, deTS [197] and PGCA perform tissue-specific enrichment analysis (TSEA) for detecting tissue-specific genes and for enrichment test of different forms of query data. Other methods use different definitions of the gene-sets, in some cases utilizing additional information. For instance, dmGWAS [198] integrates PPI networks and uses a search method to identify subnetworks. Compared with standard pathway methods it offers to the users the flexibility in the definition of a gene set and can utilize local PPI information. GEMB [199] defines the gene-sets using gene weights from model predictions and gene ranks from GWAS, and GENOMICper [200] uses permutations of the identified SNPs by rotation with respect to the genomic locations. GWAB [201] uses network connections to reprioritize candidate genes by integrating the GWAS and network data, whereas GenToS [202] searches for trait-associated variants in existing human GWAS. We also need to mention PAPA [203] which is a flexible tool for pleiotropic pathway analysis. As we already mentioned, aSPU, snpGeneSets, PascalX/PASCAL and MAGMA/chromMAGMA are gene-based methods that also perform GSA, whereas MAGENTA is a tool that performs meta-analysis and subsequently GSA (see "meta-analysis"). Lastly, we need to mention Inferno [204] and Mergeomics [205] which are webservers offering a variety of options, extending typical GSA applications. Inferno integrates a variety of functional genomics sources to identify causal noncoding variants using COLOC, WebGestalt, LDSC and MetaXcan. Mergeomics uses summary statistics of multi-omics association studies (GWAS, EWAS, TWAS, PWAS, etc.) and performs correction for LD, GSEA, meta-analysis and identification of regulators of disease-associated pathways and networks.

Fine-mapping

While GWAS can identify broad genomic regions associated with the trait, it doesn't pinpoint the exact causal variant within those regions. Fine mapping, working in the opposite direction of that of the gene-based approaches, is a process aimed at narrowing down and identifying causal variants, that is the specific genetic variants responsible for the observed associations between genomic regions and traits of interest. The plethora of statistical methods and study designs makes it difficult to choose an optimal approach. The different approaches that have been proposed to perform fine-mapping can be divided in three broad categories: heuristic methods that select SNPs based on LD patterns, conditional or penalized regression models that perform variable selection, and Bayesian methods that calculate posterior probabilities or Bayes Factors. Based on theoretical and empirical evidence it seems that Bayesian methods have superior performance [22]. Several factors may influence the performance of fine-mapping approaches, including the true number of causal SNPs in a region and their effect sizes, the local LD structure, the sample size, and the SNP density [22, 206]. Functional annotations are also of great importance leading to the so-called functionally informed fine-mapping (FIFM) methods [206]. The hypothesis of a single causal variant is also very restrictive, and several methods have been developed to allow multiple causal variants in a region as well as to incorporate additional layers of functional annotations, like eQTL [207]. Moreover, methods for fine-mapping of multiple datasets have been proposed, either exploiting

different LD patterns across ethnic groups or borrowing information between different traits [207].

As we already noted Bayesian methods seem to have superior performance [22] and thus it is of no surprise that most of the currently available methods operate in a Bayesian framework calculating Posterior Inclusion Probabilities (PIP) and/or Bayes Factors (BFs) in various settings: PAINTOR [208], DAP [209], fgwas [210], FINEMAP [211], flashfm [212], FINMOM [213], CARMA [214] and CAVIAR/CAVIARBF [215]. MsCAVIAR [216] is an extension of the latter method leveraging information from multiple studies, useful in trans-ethnic fine mapping. Similarly, XMAP [217] performs cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. BEATRICE [218] is a unique method that combines a hierarchical Bayesian model with a deep learning-based inference procedure, whereas RIVIERA-beta [219] performs Bayesian fine-mapping using Epigenomic Reference Annotation. On a different level, PolyFun/PolyLoc [220] do not perform fine-mapping per se but are used for estimating the prior causal probabilities of SNPs, which can then be used by other Bayesian fine-mapping methods. SusieR [221], BVS-PICA [222] and JAM [223], operate also in a Bayesian regression framework performing variable selection and penalized regression. Other regression-based methods, like SOJO [224] and ANNORE [225] work in a frequentist framework and perform lasso-type and differential shrinkage via random effects, respectively, whereas GSR utilizes a gene score regression approach [226] and RSS performs multiple regression utilizing the so-called summary statistics likelihood [227]. AHIUT [228] performs an intersection–union test based on a joint/conditional regression model with all the SNPs in a region. Lastly, we need to mention PICS2 [229], which performs probabilistic identification of causal SNPs and is the only of the methods that is available as a web-server, and echocolatoR [230] which requires minimal input from users and integrates a suite of fine-mapping tools to identify consensus variants, test enrichment and visualize the results.

Analysis of multiple traits

In this section we analyze methods developed for handling multiple traits. Depending on the type of data and the purpose of the analysis the methods can be divided into *pleiotropy* methods, methods that calculate the *genetic correlation*, methods for *mendelian randomization, transcriptome-wide association* and *colocalization* methods.

Pleiotropy

Pleiotropy is the phenomenon in which a single variant influences several traits [231]. Such methods are of great importance in genetic research and several methods have been developed during the last years. A major goal of such methods is to increase the statistical power over single trait methods. Imagine for instance a variant that produces a near-significant effect when analyzed separately for two or three traits. A method that can combine these estimates may produce significant results. Another application of a joint analysis would be to identify variants that influence both traits, or variants that influence only one of them. When all the relevant variants are considered, one can also estimate the kind of relationship between the traits (see "genetic correlation"). A review of the statistical methods to detect pleiotropy in complex traits can be found in [25].

Usually, the methods that allow for multiple trait analysis are oriented toward quantitative traits like BMI, SBP, DBP and so on, that traditionally are measured on a single cohort, resulting in the existence of cross-trait correlation that needs to be taken into account in the analysis. However, there are also methods for performing the same analysis with summary estimates derived from different cohorts, as well as methods that allow for binary traits with the case–control design, using overlapped or non-overlapped controls.

All methods base their inference on the assumption that the z-statistics follow a multivariate normal distribution (MVN) and perform different types of tests and/or different procedures to estimate or approximate the correlation structure. ACA [232] one of the first methods, estimates the traits covariance from a subset of the phenotypic data or from published studies, p_ACT [233] integrates the MVN using the trait correlation, PAT [234] uses a likelihood-ratio test, and PLEI [235] uses the union-intersection testing method, but in addition to the likelihood ratio test, it also applies generalized estimating equations under the working independence model; it can be applied for both marginal analysis and conditional analysis. USAT [236] uses a score-based test, JaSPU [237] uses an adaptive test which is robust to violations of the MVN assumptions and MTAR [238] uses a Principal Components (PC)-based test. BMASS [239] on the other hand is a Bayesian multivariate method, whereas TWT [240], MTAFS [241] and EBMMT [242], which are among the newer tools, perform a Cauchy Combined Test (CCT) to handle the correlation structure and obtain accurate p-values. SHAHER [243] uses a linear combination of traits by maximizing the proportion of its genetic variance explained by the shared variants and allows both shared and unshared variants to be effectively analyzed and HIPO [244] performs heritability-informed power optimization for conducting multi-trait association analysis. HOPS [245] computes a horizontal pleiotropy score by removing correlations between traits caused by vertical pleiotropy and normalizing effect sizes across all traits and PDR [246] performs a pleiotropic decomposition regression to identify shared components and their underlying genetic variants. We also need to mention methods like MTAG [247] and PLEIO [248] which use LDSC and apart from sample overlap also allow data from multiple studies, something that can be considered meta-analysis and methods like MSKAT [249], multiSKAT [250], MGAS [251], MAIUP [252] and MTAR (multi-trait analysis of rare variants) [253] which are gene-based methods specialized for multiple traits. Finally, methods like iMAP [254] and graphGPA2 [255] use graphical models and are capable of performing analysis of large number of traits.

On the other hand, there are several methods that assume independence of the studied samples. Most of them are designed for larger analyses of many traits from multiple studies, for instance PolarMorphism [256], JASS [257], gwas-pw [258] and FactorGo [259], sumDAG [260], combGWAS [261] and GCPBayes pipeline [262]. GCPBayes_pipeline uses the functionality of GCPBayes to perform cross-phenotype gene-set analysis between two traits. gwas-pw is used for the joint analysis of two GWAS in order to identify variants influencing both traits. PolarMorphism is based on a transform from Cartesian to polar coordinates and reports a per variant degree of 'sharedness' across traits, whereas FactorGo provides scalable variational factor analysis model that is computationally efficient for large number of traits. JASS provides interactive exploration

and visualization of the results of comparison of many traits through a web interface (Fig. 8 A-C), sumDAG goes one step further and constructs phenotype networks by using a Gaussian linear model and a directed acyclic graph, and combGWAS identifies susceptibility variants for comorbid disorders and calculate genetic correlations. EPS [263] and GPA [264] differ in integrating Pleiotropy and functional annotation from eQTL.

Genetic correlation

Genetic correlation is related to pleiotropy and describes the relationship between two traits, that is, the extent to which the genetic variants influencing one trait overlap with the genetic variants associated with the other. It thus can quantify the overall genetic similarity and provide insights into the polygenic genetic architecture of complex traits [23]. As we already saw, analyzing simultaneously multiple traits may increase power in case of horizontal pleiotropy; an additional potential application is to use the estimated correlation in order to establish causality between traits in case of vertical pleiotropy (see also next sections). Since heritability is the proportion of the phenotypic variance explained by genotypic variation it is of no surprise that genetic correlation (or, the genetic covariance) is related to the traits' heritabilities. Thus, several of the methods for estimating heritability discussed earlier, like HESS and SumHer can also calculate the correlation between traits. The most commonly used method, however, for calculating genetic correlation is LDSC (LD Score Regression). The method originally developed for distinguishing polygenicity from bias by examining the relationship between test statistics and LD score, but it is also used for estimating heritability and genetic correlation [133]. LDSC is also available through the LD Hub server. PCGC-s [265] is an adaptation of stratified LDSC for case–control studies and can also estimate genetic heritability,

Fig. 8 Analysis of multiple traits. **A** JASS analysis for Type 2 Diabetes Mellitus (T2DM), Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (SBP), indicating the pairwise genetic correlations between the traits. **B** Manhattan Plot from JASS for the combined analysis of the three traits. **C** Pairwise analysis of the SNPs identified as significant in the univariate analysis and in the combined analysis. **D** Two-sample Mendelian Randomization analysis for the association of SBP and T2DM obtained by MR-BASE

genetic correlation, and functional enrichment. Another popular tool is GNOVA [266] which calculates annotation-stratified covariance using the method of moments and allows for sample overlap. Its extension, SUPERGNOVA [267] identifies global and local genetic correlations that could provide new insights into the shared genetic basis of many phenotypes. Local correlations, among others, can be also computed using LAVA [268]. HDL [269] is a likelihood-based method which produces more precise estimates. A recent comparison found that LDSC and GNOVA are more similar and robust to LD and sample overlap compared to HDL. HDL provides biased estimates of the genetic covariance in most cases and could not distinguish genetic from non-genetic correlation. Moreover, HDL restricts the users to using the built-in reference panel, and its performs poorly when the number of shared SNPs between reference panel and GWAS is small [24]. Other tools provide somewhat different types of analyses. For instance Popcorn [270] estimates transethnic genetic correlation, GECKO [271] estimates both genetic and environmental covariances, PhenoSD [272] uses LDSC for estimating phenotypic correlations and then performs correction for multiple testing using the spectral decomposition of matrices, whereas LPM [273] is a latent probit model scalable to hundreds of annotations and phenotypes that integrates functional annotations. ccGWAS [274] is a tool for comparing two different disorders with small genetic correlation providing a case-case association test, and RHOGE [275] estimates the genetic correlation between two traits as a function of predicted gene expression effect. LOGOdetect [276] uses scan statistics with an LD score-weighted inner product of local z-scores to identify small segments that harbor local genetic correlation between two traits. DONUTS [277] is a unique method since it operates on summary statistics from families.

Mendelian randomization

Mendelian Randomization (MR) is a method suggested in the pre-GWAS era to investigate causal relationships between two traits, usually a phenotype and a disease [278] using genotype–trait associations to make inferences about environmentally modifiable causes of the traits. In technical terms, MR uses genetic variants as instrumental variables [279] to mimic the random assignment of exposures in a randomized controlled trial, similar to the way Mendel's laws of inheritance dictate the random assortment of alleles during gamete formation. By utilizing the natural randomization of genetic inheritance, MR aims to minimize biases introduced by confounding factors that usually affect observational studies when investigating the association of two traits. Usually, we are interested in a disease and some other intermediate phenotype, or another disease. For instance, the MR approach may involve the relationship between hypertension and BMI, or between hypertension and diabetes. Traditionally MR was performed with one sample (1SMR) using a single variant (usually referred to IPD methods), and subsequently multivariate methods for MR meta-analysis were developed [280]. With the emergence of GWAS these methods evolved to the most commonly used two-sample MR (2SMR) methods that utilize summary data estimates from several variants regarding the genotype–phenotype and genotype-disease association from different samples [26, 281]. To establish connection with the previous sections, MR seeks to analyze correlated traits [282] and to provide evidence for causation, in other words to distinguish vertical from horizontal pleiotropy.

Several standard methods for MR in GWAS with summary data have been made available during the last years: the inverse-variance weighted method (IVW), the various types of median estimators (simple or weighted) and the MR-Egger regression approach. IVW gives consistent estimates only if all the genetic variants in the analysis are valid instruments. The median estimator is consistent even when up to 50% of the information comes from invalid instrumental variables, whereas MR-Egger performs equally well but provides somewhat less precise estimates [283]. These methods are readily available in standard packages like TwoSampleMR [284] and MR [285]. The functionalities of TwoSampleMR are also offered, at least partially, through the webserver of MRBASE [284], which is the only method available as such (see Fig. 8, D). BWMR [286] is a tool that performs MR in a Bayesian framework. Besides the issue of weak instruments which is of importance, most modern methods also aim to perform the MR analysis accounting or correcting for horizontal pleiotropy. For instance, pIVW [287] is an extension of the IVW that accounts simultaneously for weak instruments and balanced horizontal pleiotropy and MRmix [288] uses a mixture approach allowing a fraction of the instruments to have pleiotropic effect on the outcome. Similarly, MRcML [289], MR-LDP [290], MR-Corr2 [291] and MR-PRESSO [292] provide functionalities to account for horizontal pleiotropy, whereas IMRP [293] takes a different approach and searches iteratively for horizontally pleiotropic variants and causal effects. MR-APSS [294] differs in that it performs MR accounting for both pleiotropy and sample structure which seems to be another important confounder (and includes population stratification, cryptic relatedness, and sample overlap); MRlap [295] considers both weak instrument bias and winner's curse, accounting for sample overlap. MR.CUE [296] and TS_LMM [297] offer additional functionality for handling variability of the estimates. LCV [298] is a method that estimates causal associations between traits avoiding confounding by genetic correlation, whereas OMR [299] uses information from all GWAS SNPs for causal inference and JAM-MR [300] performs variable selection and causal effect estimation in MR. CS [301], BiDirectCausal [302], MRCI [303] and LHC-MR [304] constitute another important class of methods since they can identify bidirectional causal effects. Another important extension is offered by methods like MR2 [305], MV-MR [306], MRBEE [307], MVMR-cML [308] and adOMICs [309] which extend the MR framework in the multivariate setting allowing more than one exposures or outcomes, as well as MR-BMA [310] which go one step further performing multivariate MR in a Bayesian framework. Finally, other methods like hJAM [311], MR.RAPS [312] and MRPEA [313] offer more advanced options. hJAM unifies the framework of MR and TWAS and can be applied to correlated instruments and multiple intermediates, MR.RAPS uses a three-sample genome-wide design with many independent genetic instruments across the genome to handle many weak genetic instruments and pleiotropy, whereas MRPEA uses pathway association MR analysis approach using data of environmental exposures.

Colocalization and TWAS

As we already described, the MR approach involves the combination of two types of data, a genotype-disease association, and a genotype–phenotype association. If the phenotype involves gene-expression, that is the result of an eQTL study, then we have two distinct but fundamentally related methods, the Transcriptome-wide association study (TWAS)

and the colocalization approach (Fig. 9). TWAS is based on the idea that genetic variants can influence gene expression, which subsequently can affect complex traits or diseases [27]. Thus, the approach uses information from eQTL to identify associations between predicted gene expression levels and complex traits/diseases [314]. Even though there are several different methods, the resemblance to MR is obvious; in fact several methods like SMR that uses a single variant [315], GSMR that uses multiple variants [310], and PMR [316] which can account for correlated instruments, horizontal pleiotropy, and can accommodate both single traits and multiple correlated outcomes, all use the term MR, whereas the authors of TScML [317], which uses two-stage constrained maximum likelihood, which is an extension of 2SLS, explicitly state that can be used for both MR and TWAS analyses. FUSION and S-PrediXcan are the oldest and most widely known methods. FUSION is the current implementation of the first TWAS method [318], whereas S-PrediXcan [319] is the summary-data version of PrediXcan. Xu et al. [320] noted that PrediXcan and TWAS can be viewed as a special case of general association testing with multiple SNPs in a GLM and proposed the so-called sum of powered score (SPU) test implemented in aSPU-TWAS [320]. A subsequent evaluation has shown that

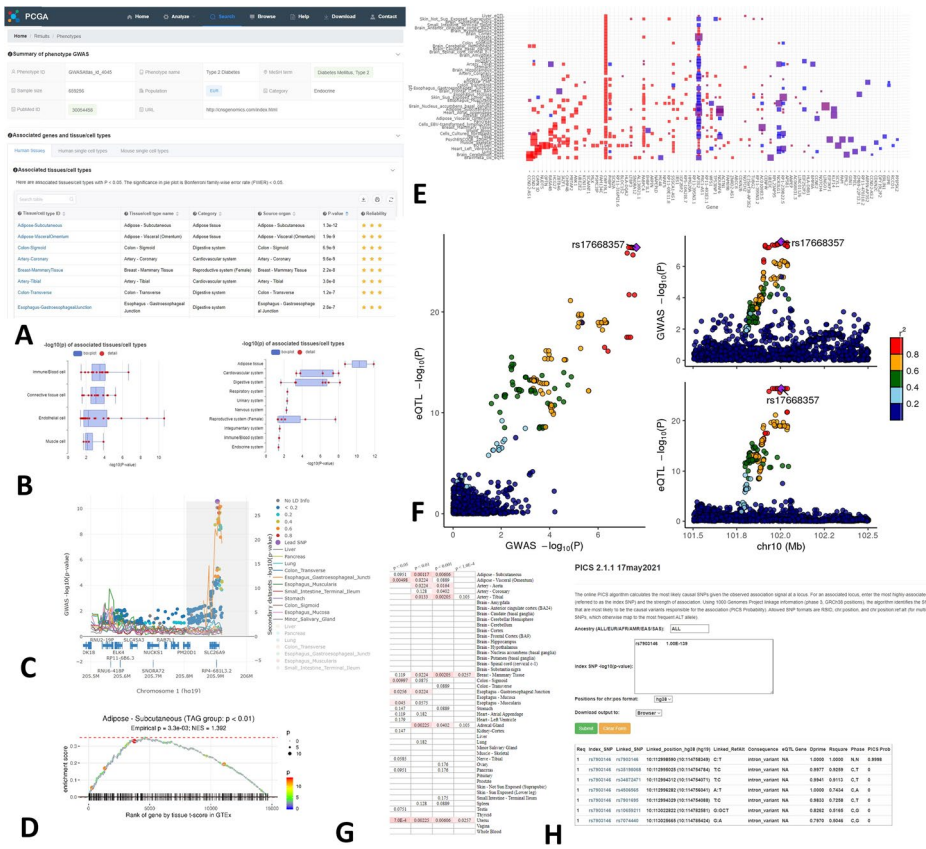


Fig. 9 Incorporation of eQTL data. **A** Overview of the gene-expression patterns in T2DM obtained by PCGA. **B** Top associated tissues and cells for T2DM (PCGA). **C** An example of colocalization output perform by LocusFocus. **D** TSEA-DB view of the analysis of significant SNPs involved in T2DM. **E** Heat-map for the tissues involved in T2DM significant hits obtained by COLOC. **F** Plots of the genome-wide significant hits obtained from GWAS and eQTL (COLOC). **G** Heat-map for the tissues involved in T2DM (TSEA-DB). **H** Example of fine-mapping regarding a SNP indicated in T2DM obtained by PICS2

the original TWAS statistic is equivalent to an LD-aware version of standard MR [321]. iFunMed [322] and sMIST [323] formulate the problem within the framework of mediator analysis, and similarly PTWAS [324] applies principles from instrumental variables analysis. Comm-S* [325] uses a variational Bayesian EM algorithm and a likelihood ratio test to assess expression-trait association. Its extension Tiss-Comm [326] leverages the co-regulation of genetic variations across different tissues explicitly via a unified probabilistic model and also detects the tissue-specific role of candidate target genes in complex traits. Similar multi-tissue approaches are followed by fQTL [327], sCCA [328] and UTMOST [329]. Primo [330], and OPERA [331] extend further the integration by allowing different types of xQTL data (eQTL, pQTL, mQTL etc.) to allow estimation under different conditions, whereas SUMMIT [332] uses a large eQTL summary-level dataset, penalized regression and Cauchy Combination Test and HMAT [333] aggregates TWAS association tests obtained across multiple gene expression prediction models using the harmonic mean *P*-value combination (HMP). BGW [334] and ARCHIE [335] are two methods that utilize trans-regulated eQTLs. Other tools use combination of methods, like TIGAR [336] which combines DPR and PrediXcan, whereas others, like JEPEG-MIX2-P [337] or FOCUS [338], perform TWAS using pathway information, or use LD to perform fine-mapping over the gene–trait association signals obtained from TWAS, respectively. Even though the various methods discussed here have different modeling assumptions and many were initially developed to answer different biological questions, a recent technical review of the TWAS methods showed that all can be viewed as versions of the two-sample MR analysis [339]. Indeed, several recent tools like MRLocus [340], TWMR [341], and Mr.MtRobin [342] make explicit use of the MR methodology and jargon in order to perform a sophisticated TWAS. MRLocus performs first a colocalization step to each nearly-LD-independent eQTL, and then performs an MR analysis step across eQTLs. TWMR performs a multi-gene multi-instrument MR approach to identify genes whose expression influence the phenotype. Finally, Mr.MtRobin uses multi-tissue eQTL and a reverse regression random slope mixed model to infer whether a gene is associated with a complex trait. As we have already noticed, webTWAS, apart from the database, also offers a webserver for accessing S-PrediXcan, SMR and UTMOST with user supplied datasets.

Another method that also uses GWAS results along with eQTL data is colocalization. Colocalization approaches are used to assess whether two different traits or diseases share a common causal genetic variant or set of variants at a specific genomic locus [13]. Colocalization analysis identifies genetic variants that show significant association in both GWAS and eQTL studies. However, unlike TWAS, it does not perform gene expression prediction and gene-trait association tests, but it focuses on the colocalized SNPs [28]. TWAS and colocalization are related approaches but not identical, since it has been shown that may give different results under different conditions (for instance in case of horizontal pleiotropy) and thus it has been suggested that they should be used complementary [28, 343]. COLOC was one of the first methods for colocalization and has seen several improvements [344, 345] (see also Fig. 9). The latest version uses SuSiE and allows evidence for association at multiple causal variants to be evaluated simultaneously, while at the same time separating the statistical support for each variant conditional on the causal signal being considered. MOLOC [346] is multiple-trait version

of COLOC, operating in a Bayesian framework that integrates GWAS summary data with multiple xQTL data to identify regulatory effects, HyPrColoc [347] is a deterministic Bayesian method that detects colocalization across large numbers of traits, and SS2 [348] operates across any number of gene-tissue pairs allowing for sample overlap. LLR [349] works for colocalizing genetic risk variants in multiple GWAS and phenotypes, whereas POEMcoloc [350] is an approximation to the COLOC method that can be applied when limited data are available. SparkINFERNO [351], PwCoCo [352] and ColocQuiaL [353] are pipelines offering additional functionalities, all using COLOC. eCAVIAR is another popular method [354] that uses a probabilistic model that accounts for more than one causal variant at a given locus. MSG [355] increases the power using a spliced gene approach and SharePro [356] integrates LD modeling and colocalization assessment to account for multiple causal variants in colocalization analysis. PESCA [357] uses estimates of LD that are ancestry-matched, in order to infer proportions of population-specific and shared causal variants in two populations. These estimates are then used as priors in an empirical Bayes framework for colocalization and test for enrichment of these causal variants in loci of interest. Lastly, we have to mention the methods that operate as webservers offering ease of use. Sherlock [358] which is also one of the oldest methods, uses a database of eQTL associations from different tissues to identify genetic signatures that match those for specific genes. Unlike other methods it incorporates information from both cis- and trans- eQTL SNPs. LocusFocus [359] is a web-based colocalization tool that tests colocalization using the Simple Sum method to identify relevant genes and tissues for a particular GWAS locus in the presence of high linkage disequilibrium and/or allelic heterogeneity. Regarding the analysis of eQTL data, ezQTL [360] is a webserver performing various tasks like data quality control for variants matched between different datasets, LD visualization, and colocalization analysis using eCAVIAR and HyPrColoc, whereas BAGEA [361] uses a variational Bayes framework to model cis-eQTLs using directed and undirected genomic annotations.

Conclusions

Summary statistics offer protection of privacy over IPD, as well as significant advantages in computational cost, which does not scale with the number of individuals in the study [11]. Naturally, in the post-GWAS era it is expected that a large number of methods would be developed to perform analysis using the summary results of GWAS [11]. The particular methods, integrating data from multiple sources such as LD, gene expression and biological pathways, aim to provide biological insight and improve our understanding about the functional role of identified variants [12–15]. One thing which we should emphasize is the fact that GWAS summary statistics are not mere replacements for IPD. Of course, some types of analysis can be applied using both summary data or IPD, like meta-analysis, heritability analysis, fine-mapping and so on. In such cases the summary data methods greatly enhance the applicability and the ease of use overcoming the limitations of IPD mentioned earlier. However, methods for other types of analysis, and particularly those that use multiple datasets, like TWAS, colocalization or Mendelian Randomization were designed having in mind the summary data and the integration of data from multiple sources. This is exactly the spirit of the so-called post-GWAS analysis that brought bioinformatics into a central role in genetics research [11]. Most of the

"success stories" in GWAS during the last years can be attributed to the development and the application of such methods in identifying new variants, in functional annotation, causal discovery or even in medical applications [2, 12, 362].

In this work we conducted, for the first time in the literature, a systematic review in order to identify software tools and databases dedicated to GWAS summary data analysis. We categorized the tools and databases by their functionality, in categories related to data, single-trait analysis, and multiple-trait analysis, along with their sub-categories which we analyzed and reviewed. We also compared the tools and databases based on their features, limitations, and user-friendliness. Our review identified a wide range of tools, each with unique strengths and limitations. We provided descriptions of the key features of each tool and database, including their input/output formats, data types, and computational requirements. We also discussed the overall usability and applicability of each tool for different research scenarios. We identified families of related tools for performing different or complementary tasks, for instance the CAVIAR tools (CAVIAR, CAVIARBF, msCAVIAR, eCAVIAR), the EpiXcan tools (S-MultiXcan, S-PrediXcan), the LDAK programs (SumHer, GBAT), the MAGMA tools (nMAGMA, H-MAGMA, eMAGMA) and so on. We need to emphasize that in many cases a tool, originally developed for IPD, is later adapted to handle summary data, whereas in other cases a tool is succeeded by a newer version with added capabilities. For instance, the original PrediXcan method uses only IPD, but it is now considered deprecated. S-PrediXcan and S-MultiXcan are later versions that are designed to be used with summary data. The same is the case regarding SKAT. The original method uses only IPD, but later implementations like metaSKAT or SKAT-O allow for summary data as well. At the same time, it is of importance that there are several tools that combine different functionalities. For instance there are tools that can perform meta-analysis and GSA (MAGENTA), gene-based methods that also offer functionalities for conditional analysis (GCTA), methods for analysis of multiple traits with gene-based tests (multiSKAT, MSKAT), methods that can be seen both as methods for multiple-traits or as meta-analysis (PLEIO, PASCAL), methods that perform both GSA and gene-based tests (aSPU, snpGeneSets, PascalX, PASCAL,MAGMA, FUMA). Of course, there are several single-purpose methods that use and combine different statistical tests or different methods (OWC, MCA, TWT, EBMMT, COMBAT, sumFREGAT, MKATR), and we may not forget methods like LDSC, with its variants, which was originally developed for distinguishing polygenicity from bias, but it is also used for estimating heritability and genetic correlation being integrated in many other tools and pipelines.

As we already mentioned, the tools and databases included in the study were those with a functioning URL. In many publications identified through the literature search the URL was not working. In some situations, we recovered a valid link by performing google searches, or by identifying the authors' websites, but in many cases, this was not enough. Similarly, several tools deposited in CRAN had been removed or archived. This kind of problem is something already known in the scientific community for years [363–365]. However, there is more to it. Even for the tools included in the review we could not verify without proper testing that they all work seamlessly, especially for the older ones [366]. Operating systems evolve, programming languages change, and with these the dependencies of each software also change. Even though

there are available best practices [367], it is not always realistic to expect complex software to work forever without maintenance. Even for some of the tools having valid URLs, for instance deposited on GitHub, or on personal web pages, we found statements by the authors indicating that the software is no longer maintained and that it is not easy to provide technical support. It is clear that more advanced solutions should be pursued. For instance, among the tools we identified the majority are written in R and Python, but only a handful is available as a webserver: ten of the tools for GSA, three tools for colocalization, two tools for meta-analysis, and one for pleiotropy analysis, MR and fine-mapping. Of course, several of the secondary databases we identified also provide the functionality of performing the analysis using data provided by the user (webTWAS, TSEA-DB, PCGA), but even counting these the proportion of web-tools is rather low (< 10%). Web servers and web services have become of high relevance to the field of bioinformatics during the last 20 years [368], so it is expected to have an increasing number of relevant webservers in the near future as relevant tools are available to facilitate the incorporation of existing applications [369–372]. On the other hand, some tools may be too computationally demanding, so other solutions must be found. Container-based applications [373, 374] such as Docker can simplify maintenance procedures and add to the reproducibility of research [375]. Community efforts such as udocker [376] may promote usability of complex software tools by non-experts in multi-user environments.

As data accumulates it is unavoidable to head to analyses on an even larger scale. Traditionally the large-scale analysis of many gene-disease associations is modeled by the so-called diseasome [377, 378] using graph theoretic methods [379, 380]. The gene-disease network is composed of pairwise associations obtained from public databases and is a bipartite network [379] consisting of two separate sets of nodes and the interactions between nodes belonging to the different sets. The projection to the one or the other of the sets may lead to the gene–gene or the disease-disease projected networks that inform us about the associations between members of the same set (for instance, two diseases are connected if they share common genes, and so on). Such methods are available for years, but they treat the associations as fixed inputs to the graph. As data accumulate and even more complex statistical methods are developed that allow cross-trait comparisons and combined analyses of multiple traits, along with the integration of different types of data such as xQTL, it is tempting to speculate that a fusion of these two traditions may come, in which the statistical formalism of the tools presented in this review will merge with the graph theoretic approaches developed in the systems biology literature. For instance, we may see network approaches leading to causal analyses (similar to MR) that consider simultaneously all the diseases and traits for which we have GWAS summary data, or similar approaches that integrate xQTL data of various types, different tissues and so on.

We hope that this comprehensive review will serve as a valuable resource for researchers who are interested in using GWAS summary statistics to investigate the genetic basis of complex traits and diseases, as well as to methodologists that develop and test relevant methods. We provided a detailed overview of the available tools and databases, and we hope that this work will facilitate informed tool selection and will maximize the effectiveness of using GWAS summary statistics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00385-x.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
The authors would like to thank the anonymous reviewers whose comments and constructive criticism helped in improving the quality of the manuscript.

Authors' contributions
PK: Investigation, Methodology, Data Curation, Visualization. PB: Conceptualization, Supervision, Investigation, Methodology, Data Curation, Visualization. PK and PB wrote parts of the manuscript and have read and approved the final manuscript.

Funding
This work is funded by the project "Bridging big omic, genetic and medical data for Precision Medicine implementation in Greece" (TAEDR-0539180) which is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union –NextGenerationEU.

Availability of data and materials
The data collected in this study are available in Supplementary Material. Supplementary Table 1 contains the list with the identified tools along with the URLs, the references and the descriptions. Supplementary Table 2 contains the list with the additional datasets identified in various consortia.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 February 2024 Accepted: 27 August 2024
Published online: 05 September 2024

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.
2. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023;110(2):179–94.
3. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8–28.
4. Alsheikh AJ, Wollenhaupt S, King EA, Reeb J, Ghosh S, Stolzenburg LR, et al. The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. BMC Med Genomics. 2022;15(1):74.
5. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010;26(4):445–55.
6. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4(8): e1000167.
7. Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet. 2011;12(10):730–6.
8. Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. Deterministic identification of specific individuals from GWAS results. Bioinformatics. 2015;31(11):1701–7.
9. Thelwall M, Munafo M, Mas-Bleda A, Stuart E, Makita M, Weigert V, et al. Is useful research data usually shared? An investigation of genome-wide association study summary statistics. PLoS One. 2020;15(2): e0229578.
10. Reales G, Wallace C. Sharing GWAS summary statistics results in more citations. Commun Biol. 2023;6(1):116.
11. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18(2):117–27.

12. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018;102(5):717–30.
13. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020;11:424.
14. Chimusa ER, Dalvie S, Dandara C, Wonkam A, Mazandu GK. Post genome-wide association analysis: dissecting computational pathway/network-based approaches. Brief Bioinform. 2019;20(2):690–700.
15. Ishigaki K. Beyond GWAS: from simple associations to functional insights. Semin Immunopathol. 2022;44(1):3–14.
16. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res. 2012;40(9):3777–84.
17. Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR, International Meta-analysis of HIVHG. Commentary: meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol. 2002;156(3):204–10.
18. Tang M, Wang T, Zhang X. A review of SNP heritability estimation methods. Brief Bioinform. 2022;23(3).
19. Zhu H, Zhou X. Statistical methods for SNP heritability estimation and partition: A review. Comput Struct Biotechnol J. 2020;18:1557–68.
20. Cinar O, Viechtbauer W. A Comparison of Methods for Gene-Based Testing That Account for Linkage Disequilibrium. Front Genet. 2022;13: 867724.
21. Mooney MA, Wilmot B. Gene set analysis: A step-by-step guide. Am J Med Genet B Neuropsychiatr Genet. 2015;168(7):517–27.
22. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491–504.
23. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20(10):567–81.
24. Zhang Y, Cheng Y, Jiang W, Ye Y, Lu Q, Zhao H. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. Brief Bioinform. 2021;22(5).
25. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017;7(11).
26. Boehm FJ, Zhou X. Statistical methods for Mendelian randomization in genome-wide association studies: A review. Comput Struct Biotechnol J. 2022;20:2338–51.
27. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592–9.
28. Hukku A, Sampson MG, Luca F, Pique-Regi R, Wen X. Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. Am J Hum Genet. 2022;109(5):825–37.
29. MacArthur JAL, Buniello A, Harris LW, Hayhurst J, McMahon A, Sollis E, et al. Workshop proceedings: GWAS summary statistics standards and sharing. Cell Genom. 2021;1(1).
30. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372: n71.
31. Hayhurst J, Buniello A, Harris L, Mosaku A, Chang C, Gignoux CR, et al. A community driven GWAS summary statistics standard. bioRxiv. 2023:2022.07.15.500230.
32. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
33. Lyon MS, Andrews SJ, Elsworth B, Gaunt TR, Hemani G, Marcora E. The variant call format provides efficient and robust storage of GWAS summary statistics. Genome Biol. 2021;22(1):32.
34. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. bioRxiv. 2020:2020.08.10.244293.
35. van der Most PJ, Vaez A, Prins BP, Munoz ML, Snieder H, Alizadeh BZ, et al. QCGWAS: A flexible R package for automated quality control of genome-wide association results. Bioinformatics. 2014;30(8):1185–6.
36. Fuchsberger C, Taliun D, Pramstaller PP, Pattaro C. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. Bioinformatics. 2012;28(3):444–5.
37. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. Nat Protoc. 2014;9(5):1192–212.
38. Chen GB, Lee SH, Robinson MR, Trzaskowski M, Zhu ZX, Winkler TW, et al. Across-cohort QC analyses of GWAS summary statistics from complex traits. Eur J Hum Genet. 2016;25(1):137–46.
39. Murphy AE, Schilder BM, Skene NG. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. Bioinformatics. 2021;37(23):4593–6.
40. He Y, Koido M, Shimmori Y, Kamatani Y. GWASLab: a Python package for processing and visualizing GWAS summary statistics. 2023.
41. Matushyn M, Bose M, Mahmoud AA, Cuthbertson L, Tello C, Bircan KO, et al. SumStatsRehab: an efficient algorithm for GWAS summary statistics assessment and restoration. BMC Bioinformatics. 2022;23(1):443.
42. Ani A, van der Most PJ, Snieder H, Vaez A, Nolte IM. GWASinspector: comprehensive quality control of genome-wide association study results. Bioinformatics. 2021;37(1):129–30.
43. Awasthi S, Chen CY, Lam M, Huang H, Ripke S, Altar CA. GWAS quality score for evaluating associated regions in GWAS analyses. Bioinformatics. 2023;39(1).
44. Chen W, Wu Y, Zheng Z, Qi T, Visscher PM, Zhu Z, et al. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. Nat Commun. 2021;12(1):7117.
45. Williams CM, Poore H, Tanksley PT, Kweon H, Courchesne-Krak NS, Londono-Correa D, et al. Guidelines for Evaluating the Comparability of Down-Sampled GWAS Summary Statistics. Behav Genet. 2023;53(5–6):404–15.
46. Baxevanis AD, Bateman A. The Importance of Biological Databases in Biological Discovery. Curr Protoc Bioinformatics. 2015;50:1–8.
47. Ison J, Rapacki K, Menager H, Kalas M, Rydza E, Chmura P, et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 2016;44(D1):D38-47.
48. Rigden DJ, Fernandez XM. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Res. 2020;48(D1):D1–8.

49. Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. Genomics Proteomics Bioinformatics. 2015;13(1):55–63.
50. Hassani-Pak K, Rawlings C. Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. J Integr Bioinform. 2017;14(1).
51. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39(10):1181–6.
52. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–12.
53. Beck T, Rowlands T, Shorter T, Brookes AJ. GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies. Nucleic Acids Res. 2023;51(D1):D986–93.
54. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50(11):1593–9.
55. McInnes G, Tanigawa Y, DeBoever C, Lavertu A, Olivieri JE, Aguirre M, et al. Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. Bioinformatics. 2019;35(14):2495–7.
56. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30.
57. Huang D, Feng X, Yang H, Wang J, Zhang W, Fan X, et al. QTLbase2: an enhanced catalog of human quantitative trait loci on extensive molecular phenotypes. Nucleic Acids Res. 2023;51(D1):D1122–8.
58. Dai Y, Hu R, Manuel AM, Liu A, Jia P, Zhao Z. CSEA-DB: an omnibus for human complex trait and cell type associations. Nucleic Acids Res. 2021;49(D1):D862–70.
59. Xue C, Jiang L, Zhou M, Long Q, Chen Y, Li X, et al. PCGA: a comprehensive web server for phenotype-cell-gene association analysis. Nucleic Acids Res. 2022;50(W1):W568–76.
60. Cao C, Wang J, Kwok D, Cui F, Zhang Z, Zhao D, et al. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res. 2022;50(D1):D1123–30.
61. Pan S, Kang H, Liu X, Li S, Yang P, Wu M, et al. COLOCdb: a comprehensive resource for multi-model colocalization of complex traits. Nucleic Acids Res. 2024;52(D1):D871–81.
62. Watanabe K, Stringer S, Frei O, Umicevic Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019;51(9):1339–48.
63. Patron J, Serra-Cayuela A, Han B, Li C, Wishart DS. Assessing the performance of genome-wide association studies for predicting disease risk. PLoS ONE. 2019;14(12): e0220215.
64. Bastarache L, Denny JC, Roden DM. Phenome-Wide Association Studies. JAMA. 2022;327(1):75–6.
65. Verma A, Ritchie MD. Current Scope and Challenges in Phenome-Wide Association Studies. Curr Epidemiol Rep. 2017;4(4):321–9.
66. Wang L, Zhang X, Meng X, Koskeridis F, Georgiou A, Yu L, et al. Methodology in phenome-wide association studies: a systematic review. J Med Genet. 2021;58(11):720–8.
67. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019;35(22):4851–3.
68. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31(12):1102–10.
69. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 2017;33(2):272–9.
70. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
71. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511.
72. Naj AC. Genotype Imputation in Genome-Wide Association Studies. Curr Protoc Hum Genet. 2019;102(1): e84.
73. Dickhaus T, Stange J, Demirhan H. On an extended interpretation of linkage disequilibrium in genetic case-control association studies. Stat Appl Genet Mol Biol. 2015;14(5):497–505.
74. Kwan JS, Li MX, Deng JE, Sham PC. FAPI: Fast and accurate P-value Imputation for genome-wide association study. Eur J Hum Genet. 2016;24(5):761–6.
75. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics. 2014;30(20):2906–14.
76. Julienne H, Shi H, Pasaniuc B, Aschard H. RAISS: robust and accurate imputation from summary statistics. Bioinformatics. 2019;35(22):4837–9.
77. Lee D, Bigdeli TB, Williamson VS, Vladimirov VI, Riley BP, Fanous AH, et al. DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. Bioinformatics. 2015;31(19):3099–104.
78. Rueger S, McDaid A, Kutalik Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. 2018;14(5): e1007371.
79. Xu Z, Duan Q, Yan S, Chen W, Li M, Lange E, et al. DISSCO: direct imputation of summary statistics allowing covariates. Bioinformatics. 2015;31(15):2434–42.
80. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: direct imputation of summary statistics for unmeasured SNPs. Bioinformatics. 2013;29(22):2925–7.
81. Togninalli M, Roqueiro D, Investigators CO, Borgwardt KM. Accurate and adaptive imputation of summary statistics in mixed-ethnicity cohorts. Bioinformatics. 2018;34(17):i687–96.
82. Park DS, Brown B, Eng C, Huntsman S, Hu D, Torgerson DG, et al. Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. Bioinformatics. 2015;31(12):i181–9.
83. Ren J, Lin Z, Pan W. Integrating GWAS summary statistics, individual-level genotypic and omic data to enhance the performance for large-scale trait imputation. Hum Mol Genet. 2023;32(17):2693–703.
84. Ren J, Lin Z, He R, Shen X, Pan W. Using GWAS summary data to impute traits for genotyped individuals. HGG Adv. 2023;4(3): 100197.

85. Yang Z, Paschou P, Drineas P. Reconstructing SNP allele and genotype frequencies from GWAS summary statistics. Sci Rep. 2022;12(1):8242.
86. Bagos PG, Nikolopoulos GK. A method for meta-analysis of case-control genetic association studies using logistic regression. Stat Appl Genet Mol Biol. 2007;6:Article17.
87. Bagos PG. A unification of multivariate methods for meta-analysis of genetic association studies. Stat Appl Genet Mol Biol. 2008;7(1):Article31.
88. Bagos PG. Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis. Stat Appl Genet Mol Biol. 2013;12(3):285–308.
89. Dimou NL, Tsirigos KD, Elofsson A, Bagos PG. GWAR: robust analysis and meta-analysis of genome-wide association studies. Bioinformatics. 2017;33(10):1521–7.
90. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. Stat Med. 2006;25(13):2299–322.
91. Nolte IM. Metasubtract: an R-package to analytically produce leave-one-out meta-analysis GWAS summary statistics. Bioinformatics. 2020;36(16):4521–2.
92. Woolf B, Sallis HM, Munafò MR, Gill D. Deriving GWAS summary estimates for paternal smoking in UK biobank: a GWAS by subtraction. BMC Res Notes. 2023;16(1):159.
93. Niu YF, Ye C, He J, Han F, Guo LB, Zheng HF, et al. Reproduction and In-Depth Evaluation of Genome-Wide Association Studies and Genome-Wide Meta-analyses Using Summary Statistics. G3 (Bethesda). 2017;7(3):943–52.
94. Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. Genetics. 2018;208(4):1397–408.
95. Forero DA, Lopez-Leon S, González-Giraldo Y, Bagos PG. Ten simple rules for carrying out and writing meta-analyses. PLoS Comput Biol. 2019;15(5): e1006922.
96. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 2010;34(1):60–6.
97. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med. 2008;27(11):1870–93.
98. Dai M, Ming J, Cai M, Liu J, Yang C, Wan X, et al. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. Bioinformatics. 2017;33(18):2882–9.
99. Fu S, Deng L, Zhang H, Qin J, Yu K. Integrative analysis of individual-level data and high-dimensional summary statistics. Bioinformatics. 2023;39(4).
100. Dai M, Wan X, Peng H, Wang Y, Liu Y, Liu J, et al. Joint analysis of individual-level and summary-level GWAS data by leveraging pleiotropy. Bioinformatics. 2019;35(10):1729–36.
101. Fu S, Purdue MP, Zhang H, Qin J, Song L, Berndt SI, et al. Improve the model of disease subtype heterogeneity by leveraging external summary data. PLoS Comput Biol. 2023;19(7): e1011236.
102. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379–89.
103. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.
104. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 2010;11:288.
105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
106. Meesters C, Leber M, Herold C, Angisch M, Mattheisen M, Drichel D, et al. Quick, "imputation-free" meta-analysis with proxy-SNPs. BMC Bioinformatics. 2012;13:231.
107. Jiang Y, Chen S, McGuire D, Chen F, Liu M, Iacono WG, et al. Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. PLoS Genet. 2018;14(7): e1007452.
108. Jiang W, Yu W. Jointly determining significance levels of primary and replication studies by controlling the false discovery rate in two-stage genome-wide association studies. Stat Methods Med Res. 2018;27(9):2795–808.
109. Jiang W, Yu W. Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies. Bioinformatics. 2017;33(4):500–7.
110. Jiang W, Xue JH, Yu W. What is the probability of replicating a statistically significant association in genome-wide association studies? Brief Bioinform. 2017;18(6):928–39.
111. Xie Y, Zhai S, Jiang W, Zhao H, Mehrotra DV, Shen J. Statistical assessment of biomarker replicability using MAJAR method. Stat Methods Med Res. 2023;32(10):1961–72.
112. de Vlaming R, Okbay A, Rietveld CA, Johannesson M, Magnusson PK, Uitterlinden AG, et al. Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. PLoS Genet. 2017;13(1): e1006495.
113. Province MA, Borecki IB. A correlated meta-analysis strategy for data mining "OMIC" scans. Pac Symp Biocomput. 2013:236–46.
114. Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 2010;6(8).
115. Sun J, Lyu R, Deng L, Li Q, Zhao Y, Zhang Y. SMetABF: A rapid algorithm for Bayesian GWAS meta-analysis with a large number of studies included. PLoS Comput Biol. 2022;18(3): e1009948.
116. Trochet H, Pirinen M, Band G, Jostins L, McVean G, Spencer CCA. Bayesian meta-analysis across genome-wide association studies of diverse phenotypes. Genet Epidemiol. 2019;43(5):532–47.
117. Baselmans BML, Jansen R, Ip HF, van Dongen J, Abdellaoui A, van de Weijer MP, et al. Multivariate genome-wide analyses of the well-being spectrum. Nat Genet. 2019;51(3):445–51.
118. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016;32(13):1981–9.
119. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet. 2015;96(1):21–36.

120. Ray D, Boehnke M. Methods for meta-analysis of multiple traits using GWAS summary statistics. Genet Epidemiol. 2018;42(2):134–45.
121. Baghfalaki T, Sugier PE, Truong T, Pettitt AN, Mengersen K, Liquet B. Bayesian meta-analysis models for cross cancer genomic investigation of pleiotropic effects using group structure. Stat Med. 2021;40(6):1498–518.
122. John M, Lencz T, Malhotra AK, Correll CU, Zhang JP. A simulations approach for meta-analysis of genetic association studies based on additive genetic model. Meta Gene. 2018;16:143–64.
123. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, et al. sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. Genome Biol. 2022;23(1):32.
124. Coram MA, Candille SI, Duan Q, Chan KH, Li Y, Kooperberg C, et al. Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach. Am J Hum Genet. 2015;96(5):740–52.
125. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet. 2013;14(2):139–49.
126. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era–concepts and misconceptions. Nat Rev Genet. 2008;9(4):255–66.
127. Barry CS, Walker VM, Cheesman R, Davey Smith G, Morris TT, Davies NM. How to estimate heritability: a guide for genetic epidemiologists. Int J Epidemiol. 2023;52(2):624–32.
128. Zaitlen N, Kraft P. Heritability in the genome-wide association era. Hum Genet. 2012;131(10):1655–64.
129. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol. 2011;35(5):310–7.
130. So HC, Li M, Sham PC. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genet Epidemiol. 2011;35(6):447–56.
131. Palla L, Dudbridge F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. Am J Hum Genet. 2015;97(2):250–9.
132. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet. 2016;99(1):139–53.
133. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5.
134. Song S, Jiang W, Zhang Y, Hou L, Zhao H. Leveraging LD eigenvalue regression to improve the estimation of SNP heritability and confounding inflation. Am J Hum Genet. 2022;109(5):802–11.
135. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228–35.
136. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat Genet. 2019;51(2):277–84.
137. Li H, Mazumder R, Lin X. Accurate and efficient estimation of local heritability using summary statistics and the linkage disequilibrium matrix. Nat Commun. 2023;14(1):7954.
138. Laville V, Bentley AR, Privé F, Zhu X, Gauderman J, Winkler TW, et al. VarExp: estimating variance explained by genome-wide GxE summary statistics. Bioinformatics. 2018;34(19):3412–4.
139. Shin J, Lee SH. GxEsum: a novel approach to estimate the phenotypic variance explained by genome-wide GxE interaction based on GWAS summary statistics for biobank-scale data. Genome Biol. 2021;22(1):183.
140. Song L, Liu A, Shi J. SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. Bioinformatics. 2019;35(20):4038–44.
141. Chan TF, Rui X, Conti DV, Fornage M, Graff M, Haessler J, et al. Estimating heritability explained by local ancestry and evaluating stratification bias in admixture mapping from summary statistics. Am J Hum Genet. 2023;110(11):1853–62.
142. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018;50(9):1318–26.
143. López-Cortegano E, Caballero A. GWEHS: A Genome-Wide Effect Sizes and Heritability Screener. Genes (Basel). 2019;10(8).
144. O'Connor LJ. The distribution of common-variant effect sizes. Nat Genet. 2021;53(8):1243–9.
145. Holland D, Frei O, Desikan R, Fan CC, Shadrin AA, Smeland OB, et al. Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. PLoS Genet. 2020;16(5):e1008612.
146. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet. 2020;52(6):626–33.
147. Siewert-Rocks KM, Kim SS, Yao DW, Shi H, Price AL. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. Am J Hum Genet. 2022;109(3):393–404.
148. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet. 2004;75(3):353–62.
149. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.
150. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
151. Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. Genet Epidemiol. 2008;32(6):560–6.
152. Lee D, Williamson VS, Bigdeli TB, Riley BP, Fanous AH, Vladimirov VI, et al. JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. Bioinformatics. 2015;31(8):1176–82.
153. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75, s1–3.
154. Li M, Jiang L, Mak TSH, Kwan JSH, Xue C, Chen P, et al. A powerful conditional gene-based association approach implicated functionally important genes for schizophrenia. Bioinformatics. 2019;35(4):628–35.

155. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011;88(3):283–93.
156. Bakshi A, Zhu Z, Vinkhuyzen AA, Hill WD, McRae AF, Visscher PM, et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Sci Rep. 2016;6:32894.
157. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4): e1004219.
158. Yang A, Chen J, Zhao XM. nMAGMA: a network-enhanced method for inferring risk genes from GWAS summary statistics and its application to schizophrenia. Brief Bioinform. 2021;22(4).
159. Sey NYA, Pratt BM, Won H. Annotating genetic variants to target genes using H-MAGMA. Nat Protoc. 2023;18(1):22–35.
160. Gerring ZF, Mina-Vargas A, Gamazon ER, Derks EM. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. Bioinformatics. 2021;37(16):2245–9.
161. Wang R, Lin DY, Jiang Y. EPIC: Inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. PLoS Genet. 2022;18(6): e1010251.
162. Quick C, Wen X, Abecasis G, Boehnke M, Kang HM. Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. PLoS Genet. 2020;16(12): e1009060.
163. Yurko R, Roeder K, Devlin B, G'Sell M. An approach to gene-based testing accounting for dependence of tests among nearby genes. Brief Bioinform. 2021;22(6).
164. Vsevolozhskaya OA, Shi M, Hu F, Zaykin DV. DOT: Gene-set analysis by combining decorrelated association statistics. PLoS Comput Biol. 2020;16(4): e1007819.
165. Zhang J, Zhao Z, Guo X, Guo B, Wu B. Powerful statistical method to detect disease-associated genes using publicly available genome-wide association studies summary data. Genet Epidemiol. 2019;43(8):941–51.
166. Chen X, Zhang H, Liu M, Deng HW, Wu Z. Simultaneous detection of novel genes and SNPs by adaptive p-value combination. Front Genet. 2022;13:1009428.
167. Zhang J, Guo X, Gonzales S, Yang J, Wang X. TS: a powerful truncated test to detect novel disease associated genes using publicly available gWAS summary data. BMC Bioinformatics. 2020;21(1):172.
168. Kwak IY, Pan W. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. Bioinformatics. 2017;33(1):64–71.
169. Guo B, Wu B. Statistical methods to detect novel genetic variants using publicly available GWAS summary data. Comput Biol Chem. 2018;74:76–9.
170. Wang M, Huang J, Liu Y, Ma L, Potash JB, Han S. COMBAT: A Combined Association Test for Genes Using Summary Statistics. Genetics. 2017;207(3):883–91.
171. Shao Z, Wang T, Qiao J, Zhang Y, Huang S, Zeng P. A comprehensive comparison of multilocus association methods with summary statistics in genome-wide association studies. BMC Bioinformatics. 2022;23(1):359.
172. Zhang J, Liang X, Gonzales S, Liu J, Gao XR, Wang X. A gene based combination test using GWAS summary data. BMC Bioinformatics. 2023;24(1):2.
173. He Z, Xu B, Lee S, Ionita-Laza I. Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. Am J Hum Genet. 2017;101(3):340–52.
174. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019;104(3):410–21.
175. Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. Am J Hum Genet. 2012;91(3):478–88.
176. Sun R, Lin X. Genetic Variant Set-Based Tests Using the Generalized Berk-Jones Statistic with Application to a Genome-Wide Association Study of Breast Cancer. J Am Stat Assoc. 2020;115(531):1079–91.
177. Berrandou TE, Balding D, Speed D. LDAK-GBAT: Fast and powerful gene-based association testing using summary statistics. Am J Hum Genet. 2023;110(1):23–9.
178. Mei H, Li L, Jiang F, Simino J, Griswold M, Mosley T, et al. snpGeneSets: An R Package for Genome-Wide Study Annotation. G3 (Bethesda). 2016;6(12):4087–95.
179. Krefl D, Brandulas Cammarata A, Bergmann S. PascalX: a Python library for GWAS gene and pathway enrichment tests. Bioinformatics. 2023;39(5).
180. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput Biol. 2016;12(1): e1004714.
181. Nameki R, Shetty A, Dareng E, Tyrer J, Lin X, Pharoah P, et al. chromMAGMA: regulatory element-centric interrogation of risk variants. Life Sci Alliance. 2022;5(10).
182. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826.
183. Yang Y, Basu S, Zhang L. A Bayesian hierarchically structured prior for gene-based association testing with multiple traits in genome-wide association studies. Genet Epidemiol. 2022;46(1):63–72.
184. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007;81(6):1278–83.
185. Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends Genet. 2014;30(9):390–400.
186. Pers TH. Gene set analysis for interpreting genetic studies. Hum Mol Genet. 2016;25(R2):R133–40.
187. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics. 2011;98(1):1–8.
188. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res. 2010;38(Web Server issue):W90–5.
189. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7.

190. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51(W1):W207–12.
191. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–21.
192. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47(W1):W199-w205.
193. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 2021;49(D1):D394-d403.
194. Yoon S, Nguyen HCT, Yoo YJ, Kim J, Baik B, Kim S, et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res. 2018;46(10): e60.
195. Wu C, Pan W. Integrating eQTL data with GWAS summary statistics in pathway-based analysis with application to schizophrenia. Genet Epidemiol. 2018;42(3):303–16.
196. Zhu S, Qian T, Hoshida Y, Shen Y, Yu J, Hao K. GIGSEA: genotype imputed gene set enrichment analysis using GWAS summary level data. Bioinformatics. 2019;35(1):160–3.
197. Pei G, Dai Y, Zhao Z, Jia P. deTS: tissue-specific enrichment analysis to decode tissue specificity. Bioinformatics. 2019;35(19):3842–5.
198. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 2011;27(1):95–102.
199. Cochran AL, Nieser KJ, Forger DB, Zöllner S, McInnis MG. Gene-set Enrichment with Mathematical Biology (GEMB). Gigascience. 2020;9(10).
200. Cabrera CP, Navarro P, Huffman JE, Wright AF, Hayward C, Campbell H, et al. Uncovering networks from genome-wide association studies via circular genomic permutation. G3 (Bethesda). 2012;2(9):1067–75.
201. Shim JE, Bang C, Yang S, Lee T, Hwang S, Kim CY, et al. GWAB: a web server for the network-based boosting of human genome-wide association data. Nucleic Acids Res. 2017;45(W1):W154–61.
202. Hoppmann AS, Schlosser P, Backofen R, Lausch E, Köttgen A. GenToS: Use of Orthologous Gene Information to Prioritize Signals from Human GWAS. PLoS ONE. 2016;11(9): e0162466.
203. Wen Y, Wang W, Guo X, Zhang F. PAPA: a flexible tool for identifying pleiotropic pathways using genome-wide association study summaries. Bioinformatics. 2016;32(6):946–8.
204. Amlie-Wolf A, Tang M, Mlynarski EE, Kuksa PP, Valladares O, Katanic Z, et al. INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res. 2018;46(17):8740–53.
205. Ding J, Blencowe M, Nghiem T, Ha SM, Chen YW, Li G, et al. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Res. 2021;49(W1):W375-w87.
206. Wang QS, Huang H. Methods for statistical fine-mapping and their applications to auto-immune diseases. Semin Immunopathol. 2022;44(1):101–13.
207. Hutchinson A, Asimit J, Wallace C. Fine-mapping genetic associations. Hum Mol Genet. 2020;29(R1):R81–8.
208. Kichaev G, Roytman M, Johnson R, Eskin E, Lindström S, Kraft P, et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics. 2017;33(2):248–55.
209. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am J Hum Genet. 2016;98(6):1114–29.
210. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94(4):559–73.
211. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–501.
212. Hernández N, Soenksen J, Newcombe P, Sandhu M, Barroso I, Wallace C, et al. The flashfm approach for fine-mapping multiple quantitative traits. Nat Commun. 2021;12(1):6147.
213. Karhunen V, Launonen I, Järvelin MR, Sebert S, Sillanpää MJ. Genetic fine-mapping from summary data using a nonlocal prior improves the detection of multiple causal variants. Bioinformatics. 2023;39(7).
214. Yang Z, Wang C, Liu L, Khan A, Lee A, Vardarajan B, et al. CARMA is a new Bayesian model for fine-mapping in genome-wide association meta-analyses. Nat Genet. 2023;55(6):1057–65.
215. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics. 2015;200(3):719–36.
216. LaPierre N, Taraszka K, Huang H, He R, Hormozdiari F, Eskin E. Identifying causal variants by fine mapping across multiple studies. PLoS Genet. 2021;17(9): e1009733.
217. Cai M, Wang Z, Xiao J, Hu X, Chen G, Yang C. XMAP: Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. Nat Commun. 2023;14(1):6870.
218. Ghosal S, Schatz MC, Venkataraman A. BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference. bioRxiv. 2023.a
219. Li Y, Kellis M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Res. 2016;44(18): e144.
220. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat Genet. 2020;52(12):1355–63.
221. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the "Sum of Single Effects" model. PLoS Genet. 2022;18(7): e1010299.
222. Chen S, Nunez S, Reilly MP, Foulkes AS. Bayesian variable selection for post-analytic interrogation of susceptibility loci. Biometrics. 2017;73(2):603–14.
223. Newcombe PJ, Conti DV, Richardson S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. Genet Epidemiol. 2016;40(3):188–201.
224. Ning Z, Lee Y, Joshi PK, Wilson JF, Pawitan Y, Shen X. A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. Am J Hum Genet. 2017;101(6):903–12.

225. Fisher V, Sebastiani P, Cupples LA, Liu CT. ANNORE: genetic fine-mapping with functional annotation. Hum Mol Genet. 2021;31(1):32–40.
226. Zhang W, Li SY, Liu T, Li Y. Partitioning gene-based variance of complex traits by gene score regression. PLoS ONE. 2020;15(8): e0237657.
227. Zhu X, Stephens M. BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES. Ann Appl Stat. 2017;11(3):1561–92.
228. Deng Y, Pan W. Significance Testing for Allelic Heterogeneity. Genetics. 2018;210(1):25–32.
229. Taylor KE, Ansel KM, Marson A, Criswell LA, Farh KK. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. Bioinformatics. 2021;37(18):3004–7.
230. Schilder BM, Humphrey J, Raj T. echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline. Bioinformatics. 2022;38(2):536–9.
231. Tyler AL, Crawford DC, Pendergrass SA. The detection and characterization of pleiotropy: discovery, progress, and promise. Brief Bioinform. 2016;17(1):13–22.
232. Wu P, Wang B, Lubitz SA, Benjamin EJ, Meigs JB, Dupuis J. Approximate conditional phenotype analysis based on genome wide association summary statistics. Sci Rep. 2021;11(1):2518.
233. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. Am J Hum Genet. 2007;81(6):1158–68.
234. Taraszka K, Zaitlen N, Eskin E. Leveraging pleiotropy for joint analysis of genome-wide association studies with per trait interpretations. PLoS Genet. 2022;18(11): e1010447.
235. Deng Y, Pan W. Testing Genetic Pleiotropy with GWAS Summary Statistics for Marginal and Conditional Analyses. Genetics. 2017;207(4):1285–99.
236. Ray D, Pankow JS, Basu S. USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis. Genet Epidemiol. 2016;40(1):20–34.
237. Sitlani CM, Baldassari AR, Highland HM, Hodonsky CJ, McKnight B, Avery CL. Comparison of adaptive multiple phenotype association tests using summary statistics in genome-wide association studies. Hum Mol Genet. 2021;30(15):1371–83.
238. Guo B, Wu B. Integrate multiple traits to detect novel trait-gene association using GWAS summary data with an adaptive test approach. Bioinformatics. 2019;35(13):2251–7.
239. Turchin MC, Stephens M. Bayesian multivariate reanalysis of large genetic studies identifies many new associations. PLoS Genet. 2019;15(10): e1008431.
240. Bu D, Wang X, Li Q. Summary statistics-based association test for identifying the pleiotropic effects with set of genetic variants. Bioinformatics. 2023;39(4).
241. Deng Q, Song C, Lin S. An adaptive and robust method for multi-trait analysis of genome-wide association studies using summary statistics. Eur J Hum Genet. 2023.
242. Liu W, Xu Y, Wang A, Huang T, Liu Z. The eigen higher criticism and eigen Berk-Jones tests for multiple trait association studies based on GWAS summary statistics. Genet Epidemiol. 2022;46(2):89–104.
243. Svishcheva GR, Tiys ES, Elgaeva EE, Feoktistova SG, Timmers P, Sharapov SZ, et al. A Novel Framework for Analysis of the Shared Genetic Background of Correlated Traits. Genes (Basel). 2022;13(10).
244. Qi G, Chatterjee N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. PLoS Genet. 2018;14(10): e1007549.
245. Jordan DM, Verbanck M, Do R. HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. Genome Biol. 2019;20(1):222.
246. Ballard JL, O'Connor LJ. Shared components of heritability across genetically correlated traits. Am J Hum Genet. 2022;109(6):989–1006.
247. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50(2):229–37.
248. Lee CH, Shi H, Pasaniuc B, Eskin E, Han B. PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. Am J Hum Genet. 2021;108(1):36–48.
249. Guo B, Wu B. Powerful and efficient SNP-set association tests across multiple phenotypes using GWAS summary data. Bioinformatics. 2019;35(8):1366–72.
250. Dutta D, Scott L, Boehnke M, Lee S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. Genet Epidemiol. 2019;43(1):4–23.
251. Van der Sluis S, Dolan CV, Li J, Song Y, Sham P, Posthuma D, et al. MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. Bioinformatics. 2015;31(7):1007–15.
252. Wang T, Lu H, Zeng P. Identifying pleiotropic genes for complex phenotypes with summary statistics from a perspective of composite null hypothesis testing. Brief Bioinform. 2022;23(1).
253. Luo L, Shen J, Zhang H, Chhibber A, Mehrotra DV, Tang ZZ. Multi-trait analysis of rare-variant association summary statistics using MTAR. Nat Commun. 2020;11(1):2850.
254. Zeng P, Hao X, Zhou X. Pleiotropic mapping and annotation selection in genome-wide association studies with penalized Gaussian mixture models. Bioinformatics. 2018;34(16):2797–807.
255. Deng Q, Gupta A, Jeon H, Nam JH, Yilmaz AS, Chang W, et al. graph-GPA 2.0: improving multi-disease genetic analysis with integration of functional annotation data. Front Genet. 2023;14:1079198.
256. von Berg J, Ten Dam M, van der Laan SW, de Ridder J. PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics. Bioinformatics. 2022;38(Suppl 1):i212–9.
257. Julienne H, Laville V, McCaw ZR, He Z, Guillemot V, Lasry C, et al. Multitrait GWAS to connect disease variants and biological mechanisms. PLoS Genet. 2021;17(8): e1009713.
258. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709–17.
259. Zhang Z, Jung J, Kim A, Suboc N, Gazal S, Mancuso N. A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics. Am J Hum Genet. 2023;110(11):1863–74.

260. Zilinskas R, Li C, Shen X, Pan W, Yang T. Inferring a directed acyclic graph of phenotypes from GWAS summary statistics. bioRxiv. 2023.
261. Yin L, Chau CK, Lin YP, Rao S, Xiang Y, Sham PC, et al. A framework to decipher the genetic architecture of combinations of complex diseases: applications in cardiovascular medicine. Bioinformatics. 2021;37(22):4137–47.
262. Asgari Y, Sugier PE, Baghfalaki T, Lucotte E, Karimi M, Sedki M, et al. GCPBayes pipeline: a tool for exploring pleiotropy at the gene level. NAR Genom Bioinform. 2023;5(3):lqad065.
263. Liu J, Wan X, Ma S, Yang C. EPS: an empirical Bayes approach to integrating pleiotropy and tissue-specific information for prioritizing risk genes. Bioinformatics. 2016;32(12):1856–64.
264. Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. 2014;10(11): e1004787.
265. Weissbrod O, Flint J, Rosset S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. Am J Hum Genet. 2018;103(1):89–99.
266. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. Am J Hum Genet. 2017;101(6):939–64.
267. Zhang Y, Lu Q, Ye Y, Huang K, Liu W, Wu Y, et al. SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol. 2021;22(1):262.
268. Werme J, van der Sluis S, Posthuma D, de Leeuw CA. An integrated framework for local genetic correlation analysis. Nat Genet. 2022;54(3):274–82.
269. Ning Z, Pawitan Y, Shen X. High-definition likelihood inference of genetic correlations across human complex traits. Nat Genet. 2020;52(8):859–64.
270. Brown BC, Ye CJ, Price AL, Zaitlen N. Transethnic Genetic-Correlation Estimates from Summary Statistics. Am J Hum Genet. 2016;99(1):76–88.
271. Gao B, Yang C, Liu J, Zhou X. Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies. PLoS Genet. 2021;17(1): e1009293.
272. Zheng J, Richardson TG, Millard LAC, Hemani G, Elsworth BL, Raistrick CA, et al. PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics. Gigascience. 2018;7(8).
273. Ming J, Wang T, Yang C. LPM: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations. Bioinformatics. 2020;36(8):2506–14.
274. Peyrot WJ, Price AL. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. Nat Genet. 2021;53(4):445–54.
275. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am J Hum Genet. 2017;100(3):473–87.
276. Guo H, Li JJ, Lu Q, Hou L. Detecting local genetic correlations with scan statistics. Nat Commun. 2021;12(1):2033.
277. Wu Y, Zhong X, Lin Y, Zhao Z, Chen J, Zheng B, et al. Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. Proc Natl Acad Sci U S A. 2021;118(25).
278. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol. 2004;33(1):30–42.
279. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res. 2007;16(4):309–30.
280. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomization–a multivariate approach. Stat Med. 2005;24(14):2241–54.
281. Bowden J, Holmes MV. Meta-analysis and Mendelian randomization: A review. Res Synth Methods. 2019;10(4):486–96.
282. Kraft P, Chen H, Lindström S. The Use Of Genetic Correlation And Mendelian Randomization Studies To Increase Our Understanding of Relationships Between Complex Traits. Curr Epidemiol Rep. 2020;7(2):104–12.
283. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304–14.
284. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7.
285. Burgess S, Foley CN, Allara E, Staley JR, Howson JMM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nat Commun. 2020;11(1):376.
286. Zhao J, Ming J, Hu X, Chen G, Liu J, Yang C. Bayesian weighted Mendelian randomization for causal inference based on summary statistics. Bioinformatics. 2020;36(5):1501–8.
287. Xu S, Wang P, Fung WK, Liu Z. A novel penalized inverse-variance weighted estimator for Mendelian randomization with applications to COVID-19 outcomes. Biometrics. 2023;79(3):2184–95.
288. Qi G, Chatterjee N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. Nat Commun. 2019;10(1):1941.
289. Xue H, Shen X, Pan W. Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. Am J Hum Genet. 2021;108(7):1251–69.
290. Cheng Q, Yang Y, Shi X, Yeung KF, Yang C, Peng H, et al. MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting for linkage disequilibrium and horizontal pleiotropy. NAR Genom Bioinform. 2020;2(2):lqaa028.
291. Cheng Q, Qiu T, Chai X, Sun B, Xia Y, Shi X, et al. MR-Corr2: a two-sample Mendelian randomization method that accounts for correlated horizontal pleiotropy using correlated instrumental variants. Bioinformatics. 2022;38(2):303–10.
292. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50(5):693–8.
293. Zhu X, Li X, Xu R, Wang T. An iterative approach to detect pleiotropy and perform Mendelian Randomization analysis using GWAS summary statistics. Bioinformatics. 2021;37(10):1390–400.

294. Hu X, Zhao J, Lin Z, Wang Y, Peng H, Zhao H, et al. Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. Proc Natl Acad Sci U S A. 2022;119(28): e2106858119.
295. Mounier N, Kutalik Z. Bias correction for inverse variance weighting Mendelian randomization. Genet Epidemiol. 2023;47(4):314–31.
296. Cheng Q, Zhang X, Chen LS, Liu J. Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. Nat Commun. 2022;13(1):6490.
297. Ding M. A Two-stage Linear Mixed Model (TS-LMM) for Summary-data-based Multivariable Mendelian Randomization. medRxiv. 2023.
298. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nat Genet. 2018;50(12):1728–34.
299. Wang L, Gao B, Fan Y, Xue F, Zhou X. Mendelian randomization under the omnigenic architecture. Brief Bioinform. 2021;22(6).
300. Gkatzionis A, Burgess S, Conti DV, Newcombe PJ. Bayesian variable selection with a pleiotropic loss function in Mendelian randomization. Stat Med. 2021;40(23):5025–45.
301. Xue H, Pan W. Inferring causal direction between two traits in the presence of horizontal pleiotropy with GWAS summary data. PLoS Genet. 2020;16(11): e1009105.
302. Xue H, Pan W. Robust inference of bi-directional causal relationships in presence of correlated pleiotropy with GWAS summary data. PLoS Genet. 2022;18(5): e1010205.
303. Liu Z, Qin Y, Wu T, Tubbs JD, Baum L, Mak TSH, et al. Reciprocal causation mixture model for robust Mendelian randomization analysis using genome-scale summary data. Nat Commun. 2023;14(1):1131.
304. Darrous L, Mounier N, Kutalik Z. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. Nat Commun. 2021;12(1):7274.
305. Zuber V, Lewin A, Levin MG, Haglund A, Ben-Aicha S, Emanueli C, et al. Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. Am J Hum Genet. 2023;110(7):1177–99.
306. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2019;48(3):713–27.
307. Lorincz-Comi N, Yang Y, Li G, Zhu X. MRBEE: A novel bias-corrected multivariable Mendelian Randomization method. bioRxiv. 2023.
308. Lin Z, Xue H, Pan W. Robust multivariable Mendelian randomization based on constrained maximum likelihood. Am J Hum Genet. 2023;110(4):592–605.
309. Jin C, Lee B, Shen L, Long Q. Integrating multi-omics summary data using a Mendelian randomization framework. Brief Bioinform. 2022;23(6).
310. Zuber V, Colijn JM, Klaver C, Burgess S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat Commun. 2020;11(1):29.
311. Jiang L, Xu S, Mancuso N, Newcombe PJ, Conti DV. A Hierarchical Approach Using Marginal Summary Statistics for Multiple Intermediates in a Mendelian Randomization or Transcriptome Analysis. Am J Epidemiol. 2021;190(6):1148–58.
312. Zhao Q, Chen Y, Wang J, Small DS. Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. Int J Epidemiol. 2019;48(5):1478–92.
313. Fan Q, Zhang F, Wang W, Xu J, Hao J, He A, et al. GWAS summary-based pathway analysis correcting for the genetic confounding impact of environmental exposures. Brief Bioinform. 2018;19(5):725–30.
314. Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6(1):899.
315. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
316. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nat Commun. 2020;11(1):3861.
317. Xue H, Shen X, Pan W. Causal Inference in Transcriptome-Wide Association Studies with Invalid Instruments and GWAS Summary Data. J Am Stat Assoc. 2023;118(543):1525–37.
318. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
319. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9(1):1825.
320. Xu Z, Wu C, Wei P, Pan W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. Genetics. 2017;207(3):893–902.
321. Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, et al. Transcriptome-wide association studies accounting for colocalization using Egger regression. Genet Epidemiol. 2018;42(5):418–33.
322. Rojo C, Zhang Q, Keleş S. iFunMed: Integrative functional mediation analysis of GWAS and eQTL studies. Genet Epidemiol. 2019;43(7):742–60.
323. Dong X, Su YR, Barfield R, Bien SA, He Q, Harrison TA, et al. A general framework for functionally informed set-based analysis: Application to a large-scale colorectal cancer study. PLoS Genet. 2020;16(8): e1008947.
324. Zhang Y, Quick C, Yu K, Barbeira A, Luca F, Pique-Regi R, et al. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome Biol. 2020;21(1):232.
325. Yang Y, Yeung KF, Liu J. CoMM-S(4): A Collaborative Mixed Model Using Summary-Level eQTL and GWAS Datasets in Transcriptome-Wide Association Studies. Front Genet. 2021;12: 704538.
326. Shi X, Chai X, Yang Y, Cheng Q, Jiao Y, Huang J, et al. A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. bioRxiv. 2019:789396.

327. Park Y, Sarkar A, Bhutani K, Kellis M. Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv. 2017:107623.
328. Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, et al. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. PLoS Genet. 2021;17(4): e1008973.
329. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019;51(3):568–76.
330. Gleason KJ, Yang F, Pierce BL, He X, Chen LS. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. Genome Biol. 2020;21(1):236.
331. Wu Y, Qi T, Wray NR, Visscher PM, Zeng J, Yang J. Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. Cell Genom. 2023;3(8): 100344.
332. Zhang Z, Bae YE, Bradley JR, Wu L, Wu C. SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification. Nat Commun. 2022;13(1):6336.
333. Zeng P, Dai J, Jin S, Zhou X. Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies. Hum Mol Genet. 2021;30(10):939–51.
334. Luningham JM, Chen J, Tang S, De Jager PL, Bennett DA, Buchman AS, et al. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. Am J Hum Genet. 2020;107(4):714–26.
335. Dutta D, He Y, Saha A, Arvanitis M, Battle A, Chatterjee N. Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. Nat Commun. 2022;13(1):4323.
336. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, et al. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am J Hum Genet. 2019;105(2):258–66.
337. Chatzinakos C, Georgiadis F, Lee D, Cai N, Vladimirov VI, Docherty A, et al. TWAS pathway method greatly enhances the number of leads for uncovering the molecular underpinnings of psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2020;183(8):454–63.
338. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019;51(4):675–82.
339. Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. Quant Biol. 2021;9(2):107–21.
340. Zhu A, Matoba N, Wilson EP, Tapia AL, Li Y, Ibrahim JG, et al. MRLocus: Identifying causal genes mediating a trait through Bayesian estimation of allelic heterogeneity. PLoS Genet. 2021;17(4): e1009455.
341. Porcu E, Rüeger S, Lepik K, Santoni FA, Reymond A, Kutalik Z. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun. 2019;10(1):3300.
342. Gleason KJ, Yang F, Chen LS. A robust two-sample transcriptome-wide Mendelian randomization method integrating GWAS with multi-tissue eQTL summary statistics. Genet Epidemiol. 2021;45(4):353–71.
343. Al-Barghouthi BM, Rosenow WT, Du KP, Heo J, Maynard R, Mesner L, et al. Transcriptome-wide association study and eQTL colocalization identify potentially causal genes responsible for human bone mineral density GWAS associations. Elife. 2022;11.
344. Plagnol V, Smyth DJ, Todd JA, Clayton DG. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics. 2009;10(2):327–34.
345. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021;17(9): e1009440.
346. Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics. 2018;34(15):2538–45.
347. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nat Commun. 2021;12(1):764.
348. Wang F, Panjwani N, Wang C, Sun L, Strug LJ. A flexible summary statistics-based colocalization method with application to the mucin cystic fibrosis lung disease modifier locus. Am J Hum Genet. 2022;109(2):253–69.
349. Liu J, Wan X, Wang C, Yang C, Zhou X, Yang C. LLR: a latent low-rank approach to colocalizing genetic risk variants in multiple GWAS. Bioinformatics. 2017;33(24):3878–86.
350. King EA, Dunbar F, Davis JW, Degner JF. Estimating colocalization probability from limited summary statistics. BMC Bioinformatics. 2021;22(1):254.
351. Kuksa PP, Lee CY, Amlie-Wolf A, Gangadharan P, Mlynarski EE, Chou YF, et al. SparkINFERNO: a scalable high-throughput pipeline for inferring molecular mechanisms of non-coding genetic variants. Bioinformatics. 2020;36(12):3879–81.
352. Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nat Genet. 2020;52(10):1122–31.
353. Chen BY, Bone WP, Lorenz K, Levin M, Ritchie MD, Voight BF. ColocQuiaL: a QTL-GWAS colocalization pipeline. Bioinformatics. 2022;38(18):4409–11.
354. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99(6):1245–60.
355. Ji Y, Wei Q, Chen R, Wang Q, Tao R, Li B. Integration of multidimensional splicing data and GWAS summary statistics for risk gene discovery. PLoS Genet. 2022;18(6): e1009814.
356. Zhang W, Lu T, Sladek R, Li Y, Najafabadi HS, Dupuis J. SharePro: an accurate and efficient genetic colocalization method accounting for multiple causal signals. bioRxiv. 2023:2023.07.24.550431.
357. Shi H, Burch KS, Johnson R, Freund MK, Kichaev G, Mancuso N, et al. Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. Am J Hum Genet. 2020;106(6):805–17.
358. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet. 2013;92(5):667–80.

359. Panjwani N, Wang F, Mastromatteo S, Bao A, Wang C, He G, et al. LocusFocus: Web-based colocalization for the annotation and functional follow-up of GWAS. PLoS Comput Biol. 2020;16(10): e1008336.
360. Zhang T, Klein A, Sang J, Choi J, Brown KM. ezQTL: A Web Platform for Interactive Visualization and Colocalization of QTLs and GWAS Loci. Genomics Proteomics Bioinformatics. 2022;20(3):541–8.
361. Lamparter D, Bhatnagar R, Hebestreit K, Belgard TG, Zhang A, Hanson-Smith V. A framework for integrating directed and undirected annotations to build explanatory models of cis-eQTL data. PLoS Comput Biol. 2020;16(6): e1007770.
362. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5–22.
363. Schultheiss SJ, Münch MC, Andreeva GD, Rätsch G. Persistence and availability of Web services in computational biology. PLoS ONE. 2011;6(9): e24914.
364. Veretnik S, Fink JL, Bourne PE. Computational biology resources lack persistence and usability. PLoS Comput Biol. 2008;4(7): e1000136.
365. Wren JD. 404 not found: the stability and persistence of URLs published in MEDLINE. Bioinformatics. 2004;20(5):668–72.
366. Kern F, Fehlmann T, Keller A. On the lifetime of bioinformatics web services. Nucleic Acids Res. 2020;48(22):12523–33.
367. Taschuk M, Wilson G. Ten simple rules for making research software more robust. PLoS Comput Biol. 2017;13(4): e1005412.
368. Brazas MD, Yim D, Yeung W, Ouellette BF. A decade of Web Server updates at the Bioinformatics Links Directory: 2003–2012. Nucleic Acids Res. 2012;40(Web Server issue):W3-w12.
369. Chakiachvili M, Milanesi S, Arigon Chifolleau AM, Lefort V. WAVES: a web application for versatile enhanced bioinformatic services. Bioinformatics. 2019;35(1):140–2.
370. Daniluk P, Wilczyński B, Lesyng B. WeBIAS: a web server for publishing bioinformatics applications. BMC Res Notes. 2015;8:628.
371. Jia L, Yao W, Jiang Y, Li Y, Wang Z, Li H, et al. Development of interactive biological web applications with R/Shiny. Brief Bioinform. 2022;23(1).
372. Joppich M, Zimmer R. From command-line bioinformatics to bioGUI PeerJ. 2019;7: e8111.
373. Kadri S, Sboner A, Sigaras A, Roy S. Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology. J Mol Diagn. 2022;24(5):442–54.
374. Williams CL, Sica JC, Killen RT, Balis UG. The growing need for microservices in bioinformatics. J Pathol Inform. 2016;7:45.
375. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review. 2015;49(1):71–9.
376. Gomes J, Bagnaschi E, Campos I, David M, Alves L, Martins J, et al. Enabling rootless Linux Containers in multi-user environments: the udocker tool. Comput Phys Commun. 2018;232:84–97.
377. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A. 2007;104(21):8685–90.
378. Kontou PI, Pavlopoulou A, Dimou NL, Pavlopoulos GA, Bagos PG. Network analysis of genes and their association with diseases. Gene. 2016;590(1):68–78.
379. Corrigendum to: Bipartite graphs in systems biology and medicine: a survey of methods and applications. Gigascience. 2020;9(1).
380. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. BioData Min. 2011;4:10.

Publisher's Note