Integrating untargeted metabolomics, genetically informed causal inference, and pathway enrichment to define the obesity metabolome

Running title: Combine metabolomics and genetics to study obesity

Yu-Han H. Hsu*[1-3], Christina M. Astley*[2,3], Joanne B. Cole[2-4], Sailaja Vedantam[2,3], Josep M. Mercader[3,4], Andres Metspalu[5], Krista Fischer[5,6], Kristen Fortney[7], Eric K. Morgen[7], Clicerio Gonzalez[8,9], Maria E. Gonzalez[8,9], Tonu Esko[5,3], Joel N. Hirschhorn[#1-3]

[1] Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

[2] Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, United States of America

[3] Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America

[4] Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America

[5] Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

[6] Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

[7] BioAge Labs, Richmond, CA, United States of America

[8] Instituto Nacional de Salud Publica, Cuernavaca, Morelos, Mexico

[9] Centro de Estudios en Diabetes, Mexico City, Mexico

* These authors contributed equally to this work.

Corresponding author:

Joel N. Hirschhorn

Address: Boston Children's Hospital, 3 Blackfan Circle, CLS 16028, Boston, MA 02115, USA

Phone: (617) 919-2129

Email: Joel.Hirschhorn@childrens.harvard.edu

Competing Interests:

K.Fortney and E.K.M. are affiliated with BioAge Labs, Inc.; J.N.H. serves on the Scientific Advisory Board of Camp4 Therapeutics.

ABSTRACT

Background: Obesity and its associated diseases are major health problems characterized by extensive metabolic disturbances. Understanding the causal connections between these phenotypes and variation in metabolite levels can uncover relevant biology and inform novel intervention strategies. Recent studies have combined metabolite profiling with genetic instrumental variable (IV) analyses to infer the direction of causality between metabolites and obesity, but often omitted a large portion of untargeted profiling data consisting of unknown, unidentified metabolite signals.

Methods: We expanded upon previous research by identifying body mass index (BMI)-associated metabolites in multiple untargeted metabolomics datasets, and then performing bidirectional IV analysis to classify these metabolites based on their inferred causal relationships with BMI. Meta-analysis and pathway analysis of both known and unknown metabolites across datasets were enabled by our recently developed bioinformatics suite, PAIRUP-MS.

Results: We identified 10 known metabolites that are more likely to be the causes (e.g. alpha-hydroxybutyrate) or effects (e.g. valine) of BMI, or may have more complex bidirectional cause-effect relationships with BMI (e.g. glycine). Importantly, we also identified about 5 times more unknown than known metabolites in each of these three categories. Pathway analysis incorporating both known and unknown metabolites prioritized 40 enriched ($p < 0.05$) metabolite sets for the cause versus effect groups, providing further support that these two metabolite groups are linked to obesity via distinct biological mechanisms.

Conclusions: These findings demonstrate the potential utility of our approach to uncover causal connections with obesity from untargeted metabolomics datasets. Combining genetically informed causal inference with the ability to map unknown metabolites across datasets provides

57 a path to jointly analyze many untargeted datasets with obesity or other phenotypes. This
58 approach, applied to larger datasets with genotype and untargeted metabolite data, should
59 generate sufficient power for robust discovery and replication of causal biological connections
60 between metabolites and various human diseases.

INTRODUCTION

Abnormal blood metabolite levels are important, frequent, and quantifiable feature of obesity and its associated phenotypes, which are major health problems globally[1–5]. Recently, systematic metabolite profiling (metabolomics) studies have described widespread alterations in the obesity metabolome and identified metabolite markers associated with risk of obesity-related diseases[6–9]. However, these studies broadly have two key analytic challenges limiting the biological interpretation and scope of their findings: these correlative studies have not generally been able to distinguish the cause and effect relationships between metabolites and phenotypes, and only a portion of the thousands of metabolite signals measured by untargeted profiling technology could be chemically identified and thereby routinely investigated.

Genetic instrumental variable (IV) analysis (for causal inference) and novel bioinformatics tools (for analysis of untargeted metabolite data) now provide the means to overcome these limitations and enhance our understanding of the metabolome of any phenotype. The genetic IV framework, also known as Mendelian randomization, uses genetic variants as instruments to infer causality from observational data in the presence of unmeasured confounding, provided certain methodological assumptions are met[10,11]. Bidirectional genetic IV analysis, using in turn genetic variants affecting metabolite levels and variants affecting a phenotype such as body mass index (BMI), offers a way to ascribe directionality of causal relationships and to prioritize potentially causal metabolite-phenotype associations. Previous genetic IV studies have utilized variants identified in genome-wide association studies (GWAS) to infer causality between obesity-related phenotypes and curated sets of metabolites (e.g. branched-chain and aromatic amino acids)[12–16]. However, most studies did not perform comprehensive bidirectional IV analysis and only focused on the metabolites that could be

identified and curated from profiling data, thus likely capturing only a limited slice of obesity biology and, even within that constraint, not assessing causality.

Previously, metabolites of unknown chemical identities – a large portion of untargeted profiling data – were mostly excluded from analyses (including GWAS) because inter-study comparison and biological interpretation were technically onerous or intractable[17,18]. To address these issues, we recently developed a bioinformatics suite, PAIRUP-MS[18], to match up unknown metabolites across mass spectrometry-based untargeted profiling datasets, thereby enabling meta-analysis of multiple datasets and increasing statistical power for detecting biologically interesting unknowns. In addition, PAIRUP-MS provides a framework for annotating unknown metabolites using preexisting metabolic pathways and performing pathway analysis incorporating both known and unknown metabolites.

In this study, we demonstrate how the combination of bidirectional genetic IV framework and PAIRUP-MS can be used to analyze multiple untargeted metabolomics datasets and characterize causal connections between a phenotype and the metabolome. We identified both known and unknown BMI-associated metabolites, and then performed GWAS for each metabolite and for BMI, followed by bidirectional genetic IV analysis to identify metabolites likely to be causes or effects of obesity. In addition, we highlighted distinct biological pathways enriched for the cause versus effect metabolites, confirming that the bidirectional IV approach prioritized two distinct sets of BMI-associated metabolites. This initial work illustrates an approach that can now be generalized and scaled up to much larger datasets, which will enable well-powered studies to uncover novel metabolic causes and effects of obesity or any other phenotype of interest.

MATERIALS AND METHODS

A schematic overview of our analysis plan is shown in **Figure 1** and each step is described in more detail below.

Metabolomics datasets and data processing

Study populations: The study populations have been described previously[18–20]: (1) Obesity Extremes (OE): N = 300 sampled equally from lean, obese, and the general Estonian Biobank (EB) population, (2) Mexico City Diabetes Study (MCDS): N = 865 in a prospective study, and (3) BioAge Labs Mortality Study (BioAge): N = 583 in a retrospective mortality study nested in EB. All participants provided informed consent. Individual studies were approved by their respective local ethics committees. Boston Children's Hospital Institutional Review Board approved this research.

Metabolite data processing: Untargeted liquid chromatography-mass spectrometry (LC-MS) profiling of plasma samples, quality control, and missing value imputation of the data have been described previously[18]. The processed OE dataset contained 298 samples and 13,613 metabolite signals (322 known); MCDS contained 821 samples and 7,136 signals (242 known); BioAge contained 583 samples and 14,617 signals (603 known). Within each dataset, we performed rank-based inverse normal transformation on each signal and used the resulting abundance *z*-scores in downstream analyses. For OE and MCDS data used in BMI and genetic association analyses, we performed covariate adjustment (age, sex, and fasting time for OE; age and sex for MCDS) before the transformation. In this paper, we refer to both known and unknown metabolite signals as "metabolites" for simplicity, recognizing that an unknown signal does not always represent an independent, functional circulating metabolite.

Mapping and identifying BMI-associated metabolites (Figure 1a)

Mapping metabolites across datasets: Using the imputation-based matching algorithm in PAIRUP-MS[18], we identified 1,780 metabolite pairs (207 shared known metabolites measured in both datasets and 1,573 matched unknown or unshared known metabolites) that could be compared directly across OE and MCDS and restricted subsequent analyses to these metabolites. For pathway analyses requiring the BioAge-based metabolite set annotations (see below), we furthered mapped 1,743 (200 shared known and 1,543 matched) of these metabolite pairs to metabolites measured in BioAge.

Identifying BMI-associated metabolites: Within each cohort, we adjusted raw BMI (available for 298 OE and 818 MCDS samples) for age and sex, performed rank-based inverse normal transformation on the residuals, and used the resulting BMI *z*-scores in all further analyses. (Since the OE obese and lean samples were drawn from the BMI extremes of EB, all EB samples were used to calculate population-based *z*-scores.) To identify BMI-associated metabolites, we performed linear regression of BMI on each metabolite within each dataset, followed by inverse variance weighted meta-analysis across the two datasets, and applied a Bonferroni significance threshold ($p < 0.05/1{,}780$) in the meta-analysis.

Bidirectional instrumental variable analyses (Figure 1b)

Metabolite instrument (G_M) selection: GWAS of the BMI-associated metabolites using 294 OE and 637 MCDS samples (with available genetic data) and subsequent inverse variance weighted meta-analysis were performed as described previously[18]. To select G_M, we first identified the SNP (single nucleotide polymorphism) with the best meta-analyzed *p*-value for

each metabolite. Next, to avoid using redundant G_M, we "clumped" the best SNPs for all metabolites to select independent SNPs that have $r^2 < 0.5$ or are > 250 kb apart, and only kept the independent SNPs as G_M (along with their best-associated metabolites) in further analyses. For known metabolites in our causality groups (see below), we performed an additional sensitivity analysis using (where available) genome-wide significant ($p < 5 \times 10^{-8}$) SNPs from published metabolite GWAS[21–26] as individual G_M.

BMI instrument (G_B) selection: We used 97 BMI-associated SNPs (G_b) previously identified in GIANT[27] and their effect estimates (β_b) in our UK Biobank (UKB) GWAS to calculate a weighted genetic risk score for use as G_B (i.e. $G_B = \Sigma\ \beta_b \times G_b$). We performed BMI GWAS in UKB using 453,397 European-ancestry samples and sex-combined BMI *z*-scores, using BOLT-LMM[28] to account for relatedness and population structure (**Supplementary Text 1**). Analysis of the UKB data was approved by its governing Research Ethics Committee and the Broad Institute Institutional Review Board. The GIANT, UKB, and metabolomics cohorts have no known sample overlap. We confirmed that G_B was significantly associated with BMI in OE and MCDS and that none of the G_b are in linkage disequilibrium ($r^2 > 0.3$) with the selected G_M.

Testing for metabolite-to-BMI causal effect using G_M: The association between BMI and each G_M was extracted from the UKB GWAS summary statistics and used to calculate the Wald ratio IV effect estimate of the metabolite (shared known or matched pair) on BMI. The *p*-value for the Wald estimate was calculated using an asymptotic standard error estimate as described previously[29]. This *p*-value – a test of the null hypothesis of no causal effect of the metabolite – was used to rank metabolites as more or less likely to be causal for BMI.

Testing for BMI-to-metabolite causal effect using G_B: We performed linear regression of each BMI-associated metabolite on G_B in OE and MCDS separately, followed by inverse

variance weighted meta-analysis. The Wald ratio IV effect estimate of BMI on each metabolite was calculated using the meta-analyzed statistics, and the corresponding *p*-value was used to rank metabolites as more or less likely to be effects of BMI. As a sensitivity analysis, we performed the MR-PRESSO global test[30] to assess overall horizontal pleiotropy among the individual SNPs (G_b) contained within G_B, using metabolite-G_b association in the OE-MCDS meta-analysis and BMI-G_b association in UKB for 96 of 97 BMI SNPs (rs2033529 was excluded due to absence in our metabolite GWAS).

Defining cause, effect, and bidirectional metabolite groups (Figure 1c)

To rank BMI-associated metabolites as more or less likely to be the causes or effects of obesity, we used the -$\log_{10}$ *p*-value of the IV effect estimate for either the metabolite (G_M) or BMI (G_B) instrument, reasoning that the statistical significance of these estimates is informative. Metabolites in the top and bottom quartiles of these two *p*-value-based rankings were assigned to three distinct groups corresponding to different types of causal connections with BMI: (1) "cause": metabolites that were ranked in the top quartile using G_M and the bottom using G_B, and thus are likely to be upstream causes for BMI ; (2) "effect": metabolites that were ranked in the bottom quartile using G_M and the top using G_B, and thus are likely to be downstream effects of BMI; (3) "bidirectional": metabolites that were in the top quartiles of both rankings, suggesting complex bidirectional cause-effect relationships with BMI.

Pathway analyses of the defined metabolite groups (Figure 1d)

Metabolite set annotations: BioAge data and PAIRUP-MS were used to generate metabolite set annotations as described previously[18]. Briefly, pathway annotations from

198 ConsensusPathDB[31] were consolidated into 690 metabolite sets with unique metabolite
199 combinations (i.e. one metabolite set may correspond to multiple pathways containing identical
200 sets of metabolites). We then used metabolite correlations in BioAge to expand the metabolite
201 sets to include both known and unknown metabolites, calculating a membership score for each
202 metabolite in each set.

203 *Pathway analyses:* We applied the pathway analysis framework in PAIRUP-MS to
204 identify enriched metabolite sets for the cause, effect, and bidirectional metabolite groups we
205 defined. We compared each of the three groups individually versus all other BMI-associated
206 metabolites and, in a fourth analysis, compared the cause versus effect groups. First, for each
207 metabolite set in each comparison analysis, a two-tailed Wilcoxon rank-sum test was performed
208 to compare the membership scores of the two groups of metabolites. Next, to account for
209 correlation structure in our data, iterations of this procedure were performed using "null"
210 metabolite groups to calculate a permutation-based enrichment *p*-value for each metabolite set
211 (**Supplementary Figure 1**).
212
213 **Performing *m*/*z* query for unknown metabolites**
214 To assess if the unknown metabolites captured information redundant to the known
215 metabolites in our dataset (and to look up potential identities of unknowns classified in the three
216 causality groups), we performed *m*/*z* query as described previously[18], using the "LC-MS Search"
217 tool in the Human Metabolome Database (HMDB)[32]. The unknowns were annotated as an *m*/*z*-
218 matched adduct of a known metabolite in our data, an *m*/*z*-matched adduct of an HMDB
219 metabolite not identified in our data, or as a metabolite without a match in HMDB.

RESULTS

Identifying known and unknown metabolites associated with BMI

We used untargeted metabolomics data from OE and MCDS to identify metabolites associated with BMI. First, we identified 207 pairs of shared known metabolites measured in both cohorts, and used PAIRUP-MS to match 1,573 additional pairs of unknown or unshared known metabolites likely to represent identical or highly correlated metabolites. Then, by performing meta-analysis of both the shared known and matched pairs across the cohorts, we identified 577 BMI-associated metabolites at Bonferroni significance ($p < 0.05/1,780$), the majority of which were unknown metabolites: 418 (72.4%) consisted of two paired unknown metabolites, 59 (10.2%) consisted of a known metabolite matched to an unknown metabolite, and only 100 (17.3%) consisted of shared known metabolites. When we clustered these metabolites, we observed metabolite clusters that consisted mostly or entirely of matched pairs of unknown chemical identities (**Supplementary Figure 2**). Therefore, including these unknown metabolites in downstream analyses increased the number of candidate metabolites by nearly five-fold, and allowed us to investigate aspects of obesity biology not represented by the curated, known metabolites.

Identifying metabolites more likely to be causal for BMI

Before we could determine whether the BMI-associated metabolites are likely to be causal for BMI, we first needed to identify the SNP best-associated with each metabolite to use as genetic instrument (G_M in **Figure 1**). We therefore performed GWAS of metabolite levels in both OE and MCDS, followed by meta-analysis. We identified genome-wide significant ($p < 5 \times 10^{-8}$) SNPs for 204 (35 shared known and 169 matched) of the BMI-associated metabolites

(**Figure 2**); 66 (14 shared known and 52 matched) of these were also significant after correction for multiple hypothesis testing ($p < 5 \times 10^{-8}/577$). Overall, the matched, unknown metabolites showed comparable degree of genetic associations as the shared known metabolites, even in loci not associated with any of the knowns. Analyzing the unknowns thus greatly improved our ability to obtain significant and novel genetic instruments for metabolite signals, despite a relatively small GWAS sample size.

We observed that all 577 BMI-associated metabolites had best-associated SNPs with at least suggestive significance (maximum $p = 2.5 \times 10^{-6}$) and therefore considered the best-associated SNP for each metabolite as potential instrument. To avoid analyzing metabolites sharing the same instruments, we included only genetically independent G_M ($r^2 < 0.5$ or > 250kb apart) and the 324 (40 shared known and 284 matched) metabolites best-associated with these instruments in subsequent IV analyses (**Supplementary Table 1**). For each metabolite, we estimated the association between G_M and BMI using a large independent cohort, UKB, in a two-sample design to calculate the metabolite-to-BMI IV effect estimate. We identified 50 (11 shared known and 39 matched) metabolites with nominally significant ($p < 0.05$) metabolite-to-BMI IV *p*-values, which indicates that they are more likely to be upstream causes for BMI (**Supplementary Table 1**).

Identifying metabolites more likely to be effects of BMI

Next, to determine if the BMI-associated metabolites are likely to be effects of BMI, we combined 97 BMI SNPs previously identified in GIANT into a weighted genetic risk score using UKB effect estimates as weights. As expected, the score is a valid genetic instrument for BMI (G_B in **Figure 1**) in OE and MCDS (meta-analyzed BMI-G_B association $p = 5.9 \times 10^{-7}$). For each

of the 324 BMI-associated metabolites, we estimated the association between G_B and the metabolite using OE and MCDS data to calculate the BMI-to-metabolite IV effect estimate. A total of 56 (8 shared known and 48 matched) metabolites had nominally significant ($p < 0.05$) BMI-to-metabolite IV *p*-values and thus are more likely to be downstream effects of BMI (**Supplementary Table 1**).

Defining cause, effect, and bidirectional metabolite groups

In order to further characterize the causal relationships between BMI and its associated metabolites, we ranked the metabolites based on the significance of their G_M and G_B IV effect estimate *p*-values (i.e. metabolite-to-BMI or BMI-to-metabolite IV *p*-values, respectively), and classified a subset of them into "cause", "effect", or "bidirectional" groups using quartile cutoffs of the rankings (**Figure 3**). We defined 25 metabolites as more likely to be cause (5 shared known and 20 matched), 26 as more likely to be effect (3 shared known and 23 matched), and 19 as more likely to be bidirectional (2 shared known and 17 matched) with respect to BMI. The shared known metabolites in each group are listed in **Table 1**; the top cause, effect, and bidirectional metabolites are alpha-hydroxybutyrate, valine, and glycine, respectively. Details for all metabolites in each group are shown in **Supplementary Table 1**. We also performed *m/z* query in HMDB to obtain potential identities for the unknowns in the matched metabolite pairs (**Supplementary Table 2**) and found only 6 out of the 60 matched pairs to be potentially redundant with the known metabolites curated in our data. Hence, we identified about 5 times more matched, unknown metabolites in the three causality categories compared to only analyzing the known metabolites. In addition, we performed sensitivity analyses to assess how our genetic IV and classification scheme would be influenced by weak instrument or pleiotropy

bias (**Supplementary Text 2** and **Supplementary Table 3**); we obtained results that generally support the robustness of our approach.

Prioritizing enriched pathways for cause, effect, and bidirectional metabolites

We identified many more matched, unknown metabolite pairs in the cause, effect, and bidirectional groups compared to the shared known metabolites, but it is difficult to hypothesize on their roles in obesity biology without knowing their chemical identities. Therefore, to extract useful information from the unknowns and to gain clues about the biology broadly captured by the three causality groups, we performed PAIRUP-MS pathway analyses encompassing both known and unknown metabolites, using metabolite set annotations generated from a separate cohort, BioAge. First, we carried out three separate analyses to identify pathways with nominally significant ($p < 0.05$) enrichment for metabolites in the cause, effect, or bidirectional groups, respectively, when compared against all other BMI-associated metabolites (**Supplementary Table 4**). While the most enriched metabolite sets in each analysis are associated with different pathways, several metabolite sets were enriched in multiple analyses (e.g. "NAD *de novo* biosynthesis" was enriched for both cause and effect metabolites).

Hence, in order to identify pathways that are the most distinct between the defined metabolite groups, we next performed a pathway analysis directly comparing the cause versus effect metabolites, prioritizing 40 metabolite sets at nominal significance ($p < 0.05$; **Supplementary Table 4**). The 13 cause metabolite sets (in which cause metabolites have higher membership scores than effect metabolites) are associated with various pathways, such as those connected to inflammation (e.g. nitric oxide signaling), redox metabolism (e.g. cysteine/methionine metabolism), and appetite regulation (e.g. endocannabinoid signaling). The

27 effect metabolite sets also contain varied pathways including those related to lysine catabolism, neurobiology (e.g. addiction and catecholamine biosynthesis), and stress response (e.g. FoxO signaling). While the known metabolites in our analysis have been linked to some of the enriched metabolite sets in literature, the unknown metabolites contributed most of the data used to prioritize these sets.

Finally, to better visualize the distinguishing features between the cause versus effect metabolites in terms of their roles in biological pathways, we constructed a heat map of the metabolites' membership scores in the enriched metabolite sets using unsupervised clustering (**Figure 4**). The metabolites formed two major clusters consisting of metabolites that are mostly in the cause or effect groups, with a handful of metabolites clustering with the contrasting group (i.e. cause metabolite "misclassified" in the effect cluster or vice versa). Even more strikingly, the cause and effect metabolite sets formed two pure clusters consisting of all cause or all effect sets. This clustering pattern provides further evidence that the cause and effect metabolites we defined are involved in distinct biological processes and thus may be associated with BMI through different mechanisms.

DISCUSSION

The study of comprehensive metabolite profiles defines an exciting frontier in human pathophysiology. However, metabolite-phenotype associations discovered in metabolomics studies are often correlative in nature and additional causal inference approaches, such as genetic IV analysis, are required to help assess causality between metabolites and phenotypes. Furthermore, unknown metabolite signals are often filtered out prior to analysis of untargeted metabolomics data, greatly limiting investigation to *a priori* candidate metabolites, reducing the search space, and hindering downstream analyses such as pathway enrichment. Here we present a paradigm for combining untargeted metabolomics, genomics, and our recently described bioinformatics suite, PAIRUP-MS, to overcome these challenges. Using obesity as an exemplar state of metabolic dysregulation, we illustrate the potential utility of this approach to advance our understanding of causal connections in metabolic diseases.

In this study, we meta-analyzed hundreds of unknown metabolites from two cohorts using PAIRUP-MS, identifying novel associations between the unknowns, BMI, genetic variants, and biological pathways. Indeed, using bidirectional genetic IV analysis, we discovered about 5 times as many unknown than known metabolites with potential causal connections to BMI. While these unknowns are likely not all fully independent and functional circulating molecules, their associations with genetic variants and BMI, distinct from those with known metabolites, suggest that a sizable number of unknown metabolites reflect aspects of BMI biology not captured by known metabolites. Furthermore, the much larger number of candidate metabolites allowed us to perform PAIRUP-MS pathway analyses that account for potential redundancy, prioritizing biological pathways specific to the metabolites with cause or effect relationships to BMI. Because of the relatively small sample sizes of our cohorts, some of our

results did not meet stringent multiple hypothesis testing significance thresholds; nevertheless, they demonstrate a useful and generalizable analytic framework to probe the metabolome of obesity and other diseases as larger datasets become available.

We identified novel metabolites that may be causes of obesity, as well as replicating two known metabolites, valine and tyrosine, that may be the effects of BMI[14]. The strongest causal evidence among known metabolites was for alpha-hydroxybutyrate, which has been linked to insulin resistance, oxidative stress, glutathione biosynthesis, and mitochondrial dysfunction[6,33,34]. The oxidative stress and glutathione connections are especially intriguing since "glutathione-mediated detoxification" emerged as a significant causal pathway when we compared the cause and effect metabolite groups in pathway analysis. It is also notable that the IV effect estimate of alpha-hydroxybutyrate on BMI is protective while the observational association suggests this metabolite is obesogenic. We postulate that a mitochondrial dysfunction/altered redox state linked to high alpha-hydroxybutyrate level could lead to decreased weight gain, while shared common causes, such as an obesogenic diet, may lead to increases in both alpha-hydroxybutyrate level and BMI. This example highlights the advantage of genetic IV analyses over observational studies alone to explore the potential impact of a theoretical intervention targeted to obesity-associated metabolites that have yet to be fully characterized[35,36].

The validity of genetic IV analysis rests upon several key assumptions. Specifically, the genetic instrument must explain variation in the exposure variable and the instrument must not be associated with the outcome variable except through its relationship with the exposure (no genetic pleiotropy). Weak instrument bias towards the null and pleiotropy bias away from the null may lead to misclassification of metabolites in our three causality groups. To address weak instrument bias for our known metabolite instruments, we performed sensitivity analysis using

stronger instruments from published metabolite GWAS, showing that our results are generally robust against weak instrument bias, although some misclassification is possible due to limited power of our internal instruments. However, we could not conduct similar analysis for the unknown metabolite instruments since there is not yet a straightforward way to obtain external instruments for comparison. To address pleiotropy bias for our BMI instrument, we used a recently developed method, MR-PRESSO, to show that our BMI IV estimates are likely robust against extreme cases of pleiotropy bias. We could not examine pleiotropy in the metabolite instruments due to the lack of multiple instruments for each metabolite (especially for the unknowns where additional instruments could not be obtained from published GWAS).

Larger GWAS of both known and unknown metabolites, conducted across multiple datasets, will make it possible to extend our paradigm to understand causal biological mechanisms for various metabolic diseases and alleviate the limitations described above. With more candidate metabolites and genetic instruments emerging from better-powered studies, our approach can be expanded to mediation analyses[37], to pathway Mendelian randomization[38], or to metabolite IV subsetting according to predicted biological pathway memberships[39]. In conclusion, this study showcases the benefit of combining untargeted metabolomics with a bidirectional genetic IV approach to define the metabolome of a major human disease state, obesity. We therefore advocate for broader sharing of untargeted metabolomics and genetic datasets, similar to the approach taken by international efforts to optimize GWAS of many other phenotypes. Broader sharing would improve power and reliability of methodological frameworks such as the one presented here, and would enable a fuller realization of the potential of metabolomics to generate important insights into human diseases.

ACKNOWLEDGEMENTS

We thank the Broad Metabolomics Platform and SIGMA T2D Consortium for sharing data resources. This research has been conducted using the UK Biobank Resource under Application Number 11898. This work was supported by the National Heart, Lung, and Blood Institute grant F31HL126581 (Y.H.H.), National Institute of Diabetes and Digestive and Kidney Diseases grants T32DK110919 (Y.H.H.), K12DK094721 (C.M.A), and R01DK075787 (J.N.H.), Endocrine Scholars Award (C.M.A.), Doris Duke Charitable Foundation grant 215205 (J.N.H.), Estonian Research Council grants IUT20-60 (A.M.), PUT1665 (K.Fischer), and PUT1660 (T.E.), European Union through Horizon 2020 grant 692145 (A.M.), and European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 (A.M.).

COMPETING INTERESTS

K.Fortney and E.K.M. are affiliated with BioAge Labs, Inc.; J.N.H. serves on the Scientific Advisory Board of Camp4 Therapeutics.

REFERENCES

1 Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2014; **384**: 766–781.

2 Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* 2006; **444**: 840–846.

3 Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX *et al.* Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. *Arter Thromb Vasc Biol* 2006; **26**: 968–976.

4 Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* 2008; **371**: 569–578.

5 Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *JAMA* 2013; **309**: 71–82.

6 Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF *et al.* A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cell Metab* 2009; **9**: 311–326.

7 Ho JE, Larson MG, Ghorbani A, Cheng S, Chen MH, Keyes M *et al.* Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. *PLoS One* 2016; **11**: e0148361.

8 Cheng S, Rhee EP, Larson MG, Lewis GD, McCabe EL, Shen D *et al.* Metabolite

profiling identifies pathways associated with metabolic risk in humans. *Circulation* 2012; **125**: 2222–2231.

9 Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E *et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* 2011; **17**: 448–453.

10 Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004; **33**: 30–42.

11 Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014; **23**: R89-98.

12 Fall T, Hagg S, Magi R, Ploner A, Fischer K, Horikoshi M *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med* 2013; **10**: e1001474.

13 Holmes M V, Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B *et al.* Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *Am J Hum Genet* 2014; **94**: 198–208.

14 Wurtz P, Wang Q, Kangas AJ, Richmond RC, Skarp J, Tiainen M *et al.* Metabolic Signatures of Adiposity in Young Adults: Mendelian Randomization Analysis and Effects of Weight Change. *PLoS Med* 2014; **11**: e1001765.

15 Liu J, van Klinken JB, Semiz S, van Dijk KW, Verhoeven A, Hankemeier T *et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* 2017; **66**: 2915–2926.

16 Haase CL, Tybjaerg-Hansen A, Qayyum AA, Schou J, Nordestgaard BG, Frikke-Schmidt R. LCAT, HDL cholesterol and ischemic cardiovascular disease: a Mendelian randomization study of HDL cholesterol in 54,500 individuals. *J Clin Endocrinol Metab*

2012; **97**: E248-56.

17 Patti GJ, Tautenhahn R, Siuzdak G. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nat Protoc* 2012; **7**: 508–516.

18 Hsu Y-HH, Churchhouse C, Pers TH, Mercader JM, Metspalu A, Fischer K *et al.* PAIRUP-MS: Pathway analysis and imputation to relate unknowns in profiles from mass spectrometry-based metabolite data. *PLoS Comput Biol* 2019; **15**: 1–26.

19 Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015; **44**: 1137–1147.

20 Williams Amy AL, Jacobs Suzanne SBR, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Márquez-Luna C *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 2014; **506**: 97–101.

21 Rhee EP, Ho JE, Chen MH, Shen D, Cheng S, Larson MG *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab* 2013; **18**: 130–143.

22 Draisma HHM, Pool R, Kobl M, Jansen R, Petersen AK, Vaarhorst AAM *et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun* 2015; **6**: 7208.

23 Shin S-YY, Fauman EB, Petersen A-KK, Krumsiek J, Santos R, Huang J *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014; **46**: 543–50.

24 Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet* 2017; **49**: 568–578.

25 Burkhardt R, Kirsten H, Beutner F, Holdt LM, Gross A, Teren A *et al.* Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genet* 2015; **11**: e1005510.

26 Kettunen J, Demirkan A, Wurtz P, Draisma HH, Haller T, Rawal R *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* 2016; **7**: 11122.

27 Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**: 197–206.

28 Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015; **47**: 284–290.

29 Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res* 2017; **26**: 2333–2355.

30 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**: 693–698.

31 Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 2013; **41**: D793-800.

32 Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018; **46**: D608–D617.

33 Gall WE, Beebe K, Lawton KA, Adam KP, Mitchell MW, Nakhle PJ *et al.* alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a

nondiabetic population. *PLoS One* 2010; **5**: e10883.

34 Thompson Legault J, Strittmatter L, Tardif J, Sharma R, Tremblay-Vaillancourt V, Aubut C *et al.* A Metabolic Signature of Mitochondrial Dysfunction Revealed through a Monogenic Form of Leigh Syndrome. *Cell Rep* 2015; **13**: 981–989.

35 Burgess S, Harshfield E. Mendelian randomization to assess causal effects of blood lipids on coronary heart disease: Lessons from the past and applications to the future. Curr. Opin. Endocrinol. Diabetes Obes. 2016. doi:10.1097/MED.0000000000000230.

36 Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK *et al.* Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *Lancet* 2012. doi:10.1016/S0140-6736(12)60312-2.

37 Van Der weele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Method* 2013. doi:10.1515/em-2012-0010.

38 Burgess S, Thompson SG. Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* 2015; **181**: 251–260.

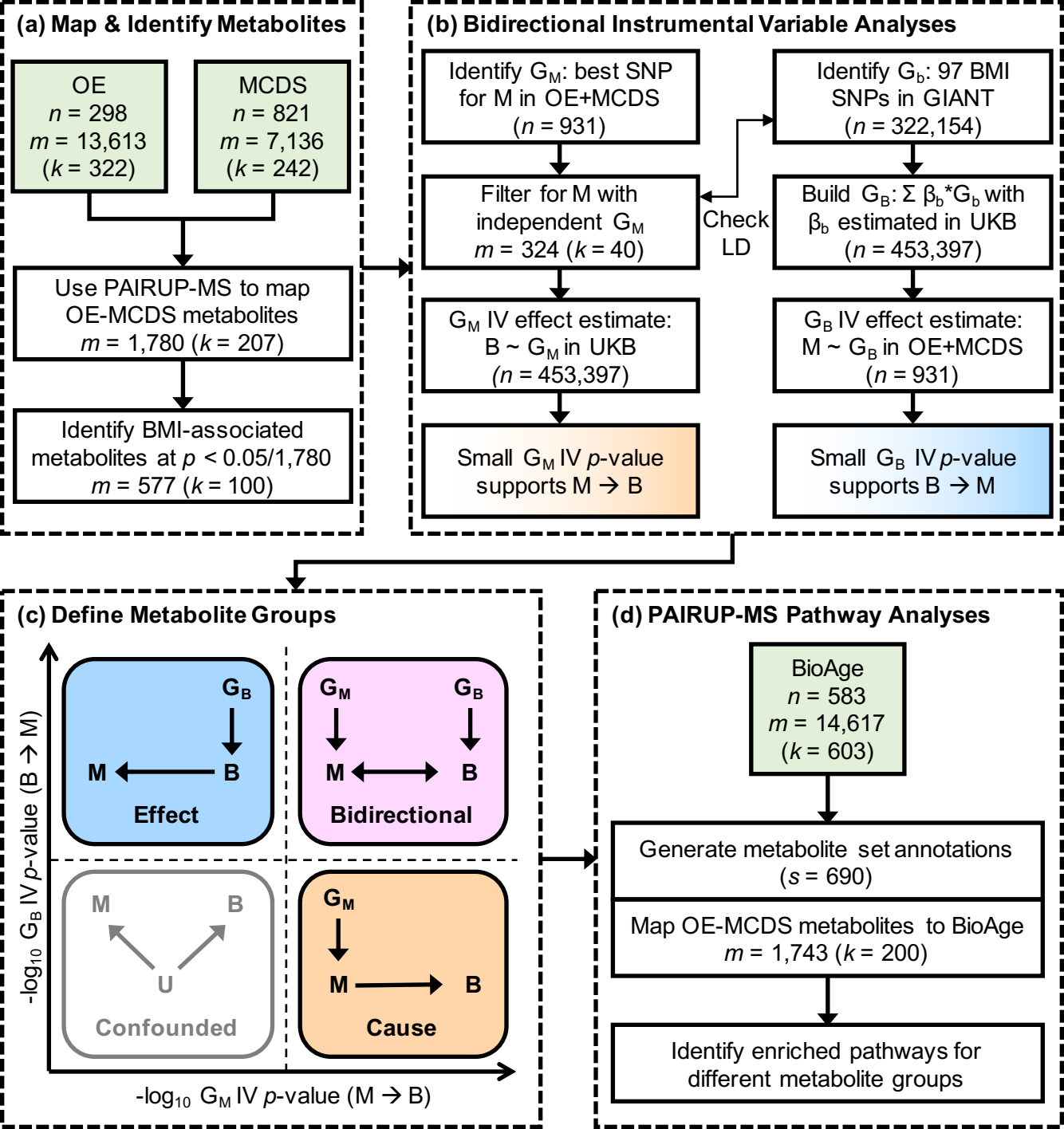
39 Wittemans LBL, Lotta LA, Oliver-Williams C, Stewart ID, Surendran P, Karthikeyan S *et al.* Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nat Commun* 2019; **10**: 1–13.

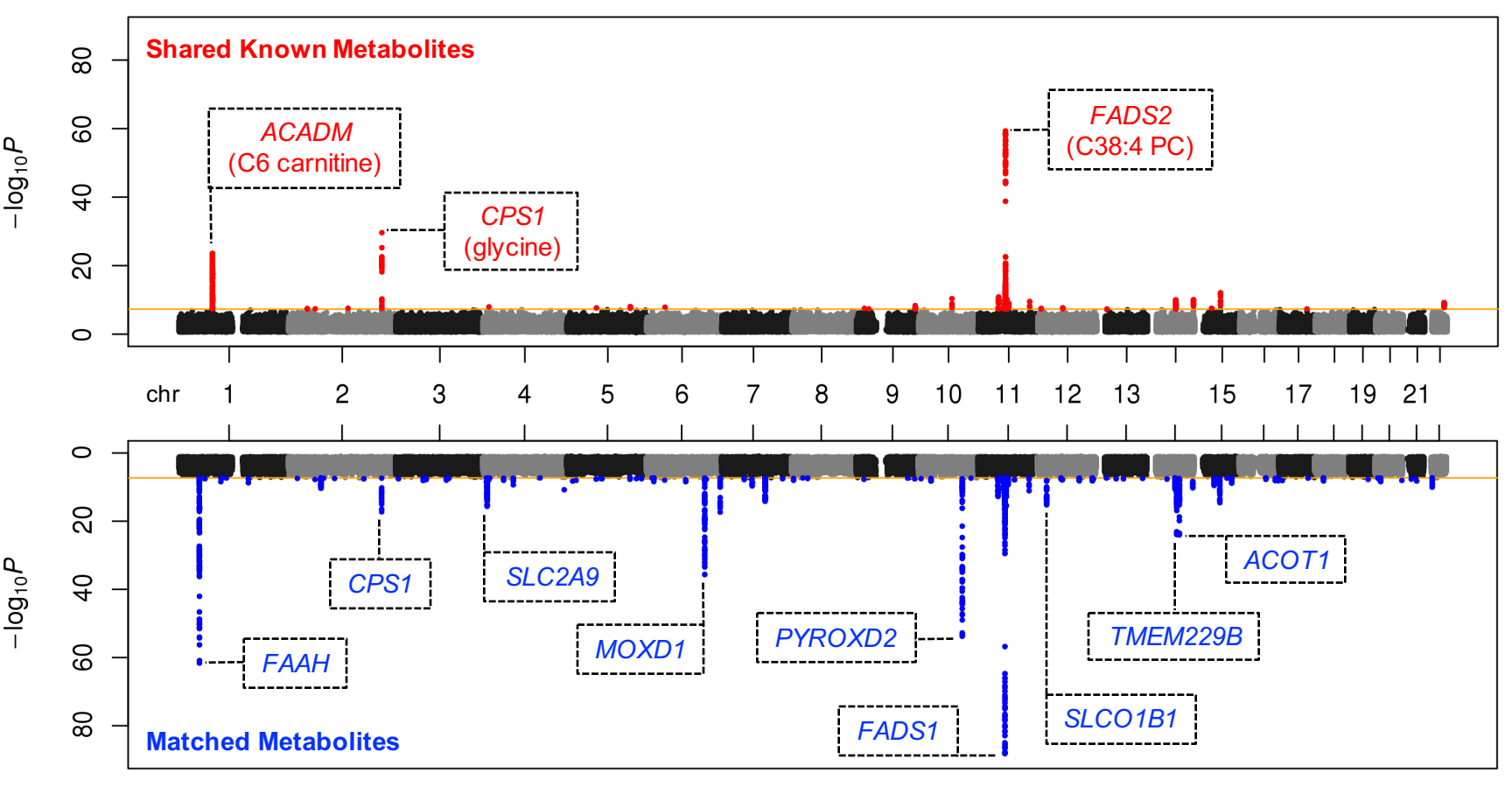
FIGURE LEGENDS

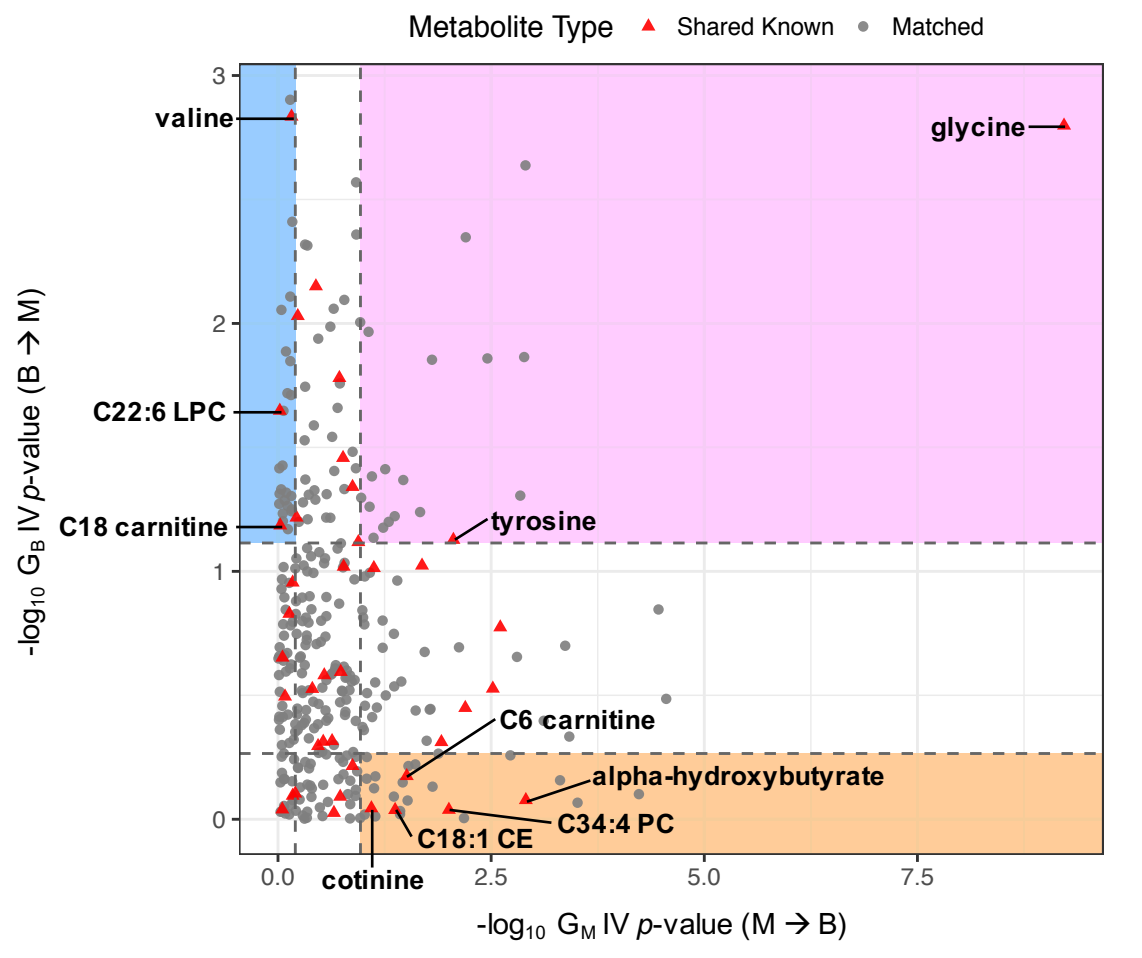
Figure 1. Overview for identifying and characterizing causal connections in the obesity metabolome. (a) OE and MCDS metabolomics datasets, matched using PAIRUP-MS, were used to identify known and unknown metabolites associated with BMI. **(b)** Independent genetic instruments (G_M) for the BMI-associated metabolites were selected using OE and MCDS data, and then used to test for a metabolite-to-BMI (M → B) causal effect in UKB; in parallel, BMI genetic instrument (G_B), a polygenic risk score built using GIANT BMI-associated SNPs (G_b) and UKB effect estimate weights (β_b), was used to test for a BMI-to-metabolite (B → M) causal effect in OE and MCDS. **(c)** A subset of metabolites was categorized into "cause", "effect", and "bidirectional" groups based on the significance of the G_M and G_B IV effect estimate *p*-values, reflecting different types of causal connections between the metabolites and BMI. **(d)** Pathway analyses of the three metabolite groups were performed using metabolite set annotations generated using PAIRUP-MS and BioAge data. Y ~ X, regression of Y on X; U, unmeasured confounder; *n*, number of samples; *m*, number of metabolites; *k*, number of known (or shared known) metabolites; *s*, number of metabolite sets.

Figure 2. Joint Manhattan plots summarizing GWAS of BMI-associated metabolites in OE and MCDS. Genetic associations for 100 shared known (top) or 477 matched (bottom) BMI-associated metabolites were consolidated to plot the best *p*-value for each SNP (i.e. only the *p*-value for the best associated metabolite was plotted for each SNP). Genome-wide significance threshold ($p < 5 \times 10^{-8}$) is marked by the orange lines. Genome-wide significant SNPs are plotted in red or blue, for shared known or matched metabolites, respectively. Lead SNPs of the most

541 significant loci ($p < 1 \times 10^{-15}$) are annotated with nearest genes (within 5kb), along with the best
542 associated known metabolites if applicable.
543
544 **Figure 3. Classifying BMI-associated metabolites using IV effect estimate *p*-values for G_M**
545 **(metabolite-to-BMI direction, x-axis) and G_B (BMI-to-metabolite direction, y-axis).** Top and
546 bottom quartile cutoffs along each axis are shown as dashed lines. Shared known metabolites in
547 "cause" (orange), "effect" (blue), and "bidirectional" (pink) regions are labeled with their names.
548
549 **Figure 4. Clustered heat map of cause and effect metabolites' memberships in metabolite**
550 **sets prioritized by pathway analysis.** Hierarchical clustering was performed using membership
551 scores in the BioAge-based metabolite set annotations. Each column is a shared known (red
552 label) or matched (black label) metabolite from the cause (yellow bar) or effect (blue bar)
553 metabolite group. Each row is a significant ($p < 0.05$) metabolite set in pathway analysis in either
554 the cause (yellow bar) or effect (blue bar) direction (with representative pathway name shown in
555 label; see **Supplementary Table 4** for full pathway list). Larger number in membership rank
556 (darker red) indicates higher membership score. Dashed light blue boxes highlight the two major
557 cause and effect clusters according to the clustering dendrograms.


(a) Map & Identify Metabolites
OE
n = 298
m = 13,613
(k = 322)
MCDS
n = 821
m = 7,136
(k = 242)
Use PAIRUP-MS to map OE-MCDS metabolites
m = 1,780 (k = 207)
Identify BMI-associated metabolites at p < 0.05/1,780
m = 577 (k = 100)
(b) Bidirectional Instrumental Variable Analyses
Identify G_M: best SNP for M in OE+MCDS
(n = 931)
Filter for M with independent G_M
m = 324 (k = 40)
G_M IV effect estimate: B ~ G_M in UKB
(n = 453,397)
Small G_M IV p-value supports M → B
Check LD
Identify G_b: 97 BMI SNPs in GIANT
(n = 322,154)
Build G_B: Σ β_b*G_b with β_b estimated in UKB
(n = 453,397)
G_B IV effect estimate: M ~ G_B in OE+MCDS
(n = 931)
Small G_B IV p-value supports B → M
(c) Define Metabolite Groups
Effect
Bidirectional
Confounded
Cause
-log10 G_B IV p-value (B → M)
-log10 G_M IV p-value (M → B)
(d) PAIRUP-MS Pathway Analyses
BioAge
n = 583
m = 14,617
(k = 603)
Generate metabolite set annotations
(s = 690)
Map OE-MCDS metabolites to BioAge
m = 1,743 (k = 200)
Identify enriched pathways for different metabolite groups


Shared Known Metabolites
ACADM
(C6 carnitine)
CPS1
(glycine)
FADS2
(C38:4 PC)
−log10P
0
20
40
60
80
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
CPS1
SLC2A9
FAAH
MOXD1
PYROXD2
FADS1
SLCO1B1
TMEM229B
ACOT1
Matched Metabolites


Metabolite Type
Shared Known
Matched
valine
glycine
C22:6 LPC
C18 carnitine
tyrosine
C6 carnitine
alpha-hydroxybutyrate
C34:4 PC
C18:1 CE
cotinine
3
2
1
0
0.0
2.5
5.0
7.5
-log10 GB IV p-value (B → M)
-log10 GM IV p-value (M → B)

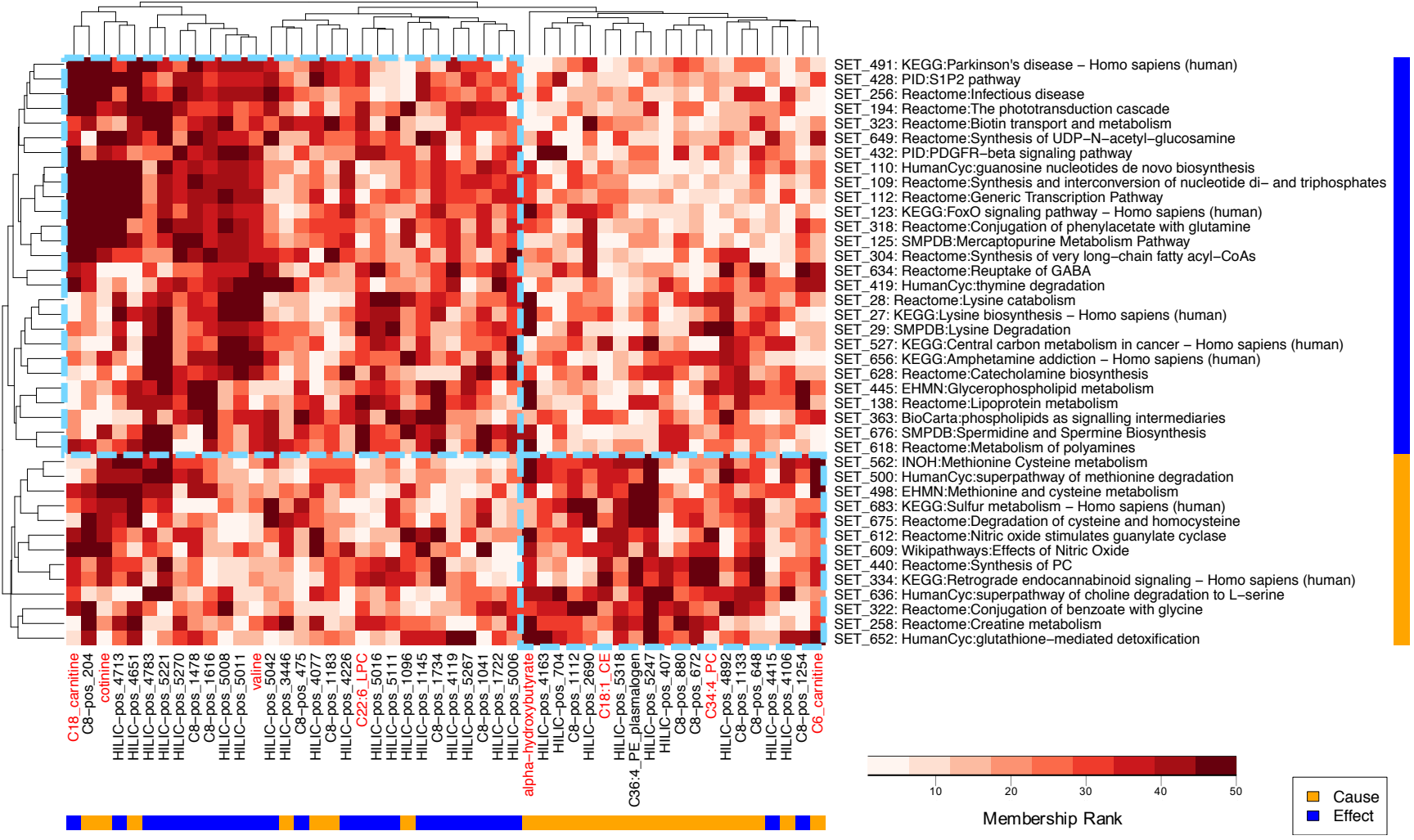

SET_491: KEGG:Parkinson's disease – Homo sapiens (human)
SET_428: PID:S1P2 pathway
SET_256: Reactome:Infectious disease
SET_194: Reactome:The phototransduction cascade
SET_323: Reactome:Biotin transport and metabolism
SET_649: Reactome:Synthesis of UDP–N–acetyl–glucosamine
SET_432: PID:PDGFR–beta signaling pathway
SET_110: HumanCyc:guanosine nucleotides de novo biosynthesis
SET_109: Reactome:Synthesis and interconversion of nucleotide di– and triphosphates
SET_112: Reactome:Generic Transcription Pathway
SET_123: KEGG:FoxO signaling pathway – Homo sapiens (human)
SET_318: Reactome:Conjugation of phenylacetate with glutamine
SET_125: SMPDB:Mercaptopurine Metabolism Pathway
SET_304: Reactome:Synthesis of very long–chain fatty acyl–CoAs
SET_634: Reactome:Reuptake of GABA
SET_419: HumanCyc:thymine degradation
SET_28: Reactome:Lysine catabolism
SET_27: KEGG:Lysine biosynthesis – Homo sapiens (human)
SET_29: SMPDB:Lysine Degradation
SET_527: KEGG:Central carbon metabolism in cancer – Homo sapiens (human)
SET_656: KEGG:Amphetamine addiction – Homo sapiens (human)
SET_628: Reactome:Catecholamine biosynthesis
SET_445: EHMN:Glycerophospholipid metabolism
SET_138: Reactome:Lipoprotein metabolism
SET_363: BioCarta:phospholipids as signalling intermediaries
SET_676: SMPDB:Spermidine and Spermine Biosynthesis
SET_618: Reactome:Metabolism of polyamines
SET_562: INOH:Methionine Cysteine metabolism
SET_500: HumanCyc:superpathway of methionine degradation
SET_498: EHMN:Methionine and cysteine metabolism
SET_683: KEGG:Sulfur metabolism – Homo sapiens (human)
SET_675: Reactome:Degradation of cysteine and homocysteine
SET_612: Reactome:Nitric oxide stimulates guanylate cyclase
SET_609: Wikipathways:Effects of Nitric Oxide
SET_440: Reactome:Synthesis of PC
SET_334: KEGG:Retrograde endocannabinoid signaling – Homo sapiens (human)
SET_636: HumanCyc:superpathway of choline degradation to L–serine
SET_322: Reactome:Conjugation of benzoate with glycine
SET_258: Reactome:Creatine metabolism
SET_652: HumanCyc:glutathione–mediated detoxification
C18_carnitine
C8–pos_204
cotinine
HILIC–pos_4713
HILIC–pos_4651
HILIC–pos_4783
HILIC–pos_5221
HILIC–pos_5270
C8–pos_1478
C8–pos_1616
HILIC–pos_5008
HILIC–pos_5011
valine
HILIC–pos_5042
HILIC–pos_3446
C8–pos_475
HILIC–pos_4077
C8–pos_1183
HILIC–pos_4226
C22:6_LPC
HILIC–pos_5016
HILIC–pos_5111
HILIC–pos_1096
HILIC–pos_1145
C8–pos_1734
HILIC–pos_4119
HILIC–pos_5267
C8–pos_1041
HILIC–pos_1722
HILIC–pos_5006
alpha–hydroxybutyrate
HILIC–pos_4163
HILIC–pos_704
HILIC–pos_1112
C8–pos_2690
C18:1_CE
HILIC–pos_5318
C36:4_PE_plasmalogen
HILIC–pos_5247
HILIC–pos_407
C8–pos_880
C8–pos_672
C34:4_PC
HILIC–pos_4892
C8–pos_1133
C8–pos_648
HILIC–pos_4415
HILIC–pos_4106
C8–pos_1254
C6_carnitine
10
20
30
40
50
Membership Rank
Cause
Effect

Table 1. BMI-associated known metabolites classified into cause, effect, or bidirectional groups using the bidirectional IV effect estimate *p*-values. Y ~ X, regression of Y on X; B, BMI; M, metabolite; covariate adjustment for B and M as described in Methods; SNP, hg19 chromsome:position is shown; β, effect size estimate; EA, effect allele (i.e. metabolite level-increasing allele). Nominally significant *p*-values (< 0.05) for IV effect estimates are in bold italic.

Group	Metabolite	Observational Association		Metabolite Instrument			G_M IV Estimate (M → B)		G_B IV Estimate (B → M)	
		β B ~ M	*P* B ~ M	SNP	EA	*P* M ~ G_M	β	*P*	β	*P*
Cause	alpha-hydroxybutyrate	0.235	7.09E-13	11:119745598	T	4.82E-07	-0.040	***1.24E-03***	-0.036	8.36E-01
	C34:4 PC	0.153	3.45E-06	3:182171263	A	1.42E-07	-0.027	***9.90E-03***	0.019	9.15E-01
	C6 carnitine	0.207	2.66E-10	1:76224010	C	2.88E-24	-0.010	***3.09E-02***	-0.076	6.69E-01
	C18:1 CE	-0.215	5.54E-11	20:38984849	T	3.44E-07	-0.016	***4.23E-02***	-0.018	9.17E-01
	cotinine	-0.169	3.10E-07	10:123918365	C	1.07E-06	0.012	8.03E-02	-0.022	9.02E-01
Effect	valine	0.445	1.20E-47	14:32724292	A	7.01E-08	-0.003	6.94E-01	0.708	***1.46E-03***
	C22:6 LPC	-0.180	4.38E-08	18:71068347	A	3.26E-07	-0.001	9.52E-01	-0.450	***2.25E-02***
	C18 carnitine	-0.193	4.65E-09	6:110760008	A	1.27E-07	0.001	9.41E-01	-0.351	***6.53E-02***
Bidirectional	glycine	-0.308	7.55E-22	2:211540507	A	2.35E-30	0.030	***6.03E-10***	-0.712	***1.59E-03***
	tyrosine	0.376	6.22E-33	6:111477887	C	1.12E-07	-0.022	***8.77E-03***	0.334	7.46E-02