Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation

Courtney J. Smith[1,*,†], Nasa Sinnott-Armstrong[1,2,*,†], Anna Cichońska[3]
Heli Julkunen[3], Eric Fauman[4], Peter Würtz[3], Jonathan K. Pritchard[1,5,†]

1. Department of Genetics, Stanford University School of Medicine, United States
2. Herbold Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington
3. Nightingale Health Plc, Helsinki, Finland
4. Internal Medicine Research Unit, Pfizer Worldwide Research, Development and Medical, United States
5. Department of Biology, Stanford University, United States
∗ Joint First Authors
† Corresponding Authors: Courtney J. Smith (courtrun@stanford.edu), Nasa Sinnott-Armstrong (nasa@fredhutch.org), Jonathan K. Pritchard (pritch@stanford.edu)

September 1, 2022.

Summary

Pleiotropy and genetic correlation are widespread features in GWAS, but they are often difficult to interpret at the molecular level. Here, we perform GWAS of 16 metabolites clustered at the intersection of amino acid catabolism, glycolysis, and ketone body metabolism in a subset of UK Biobank. We utilize the well-documented biochemistry jointly impacting these metabolites to analyze pleiotropic effects in the context of their pathways. Among the 213 lead GWAS hits, we find a strong enrichment for genes encoding pathway-relevant enzymes and transporters. We demonstrate that the effect directions of variants acting on biology between metabolite pairs often contrast with those of upstream or downstream variants as well as the polygenic background. Thus, we find that these outlier variants often reflect biology local to the traits. Finally, we explore the implications for interpreting disease GWAS, underscoring the potential of unifying biochemistry with dense metabolomics data to understand the molecular basis of pleiotropy in complex traits and diseases.

Introduction

A central challenge in the field of human genetics is understanding the mechanism of how genetic variants influence complex traits and diseases. Genome-wide association studies (GWAS) have begun characterizing the genetic architecture of complex traits, but the molecular mechanisms connecting genetic variants to these traits are rarely understood. This is particularly true for understanding pleiotropy, when a variant affects multiple traits [1]. It is possible to estimate the genetic correlation between traits [2, 3], but it is often unclear what contributes to this at a molecular or physiological level. A handful of *in vitro* disease-focused "post-GWAS" studies have convincingly shown the mechanisms driving pleiotropy of individual key associations [4, 5]; however, these studies are highly specific and time-consuming. Developing statistical and computational approaches to identify putative molecular mechanisms is invaluable to advancing our understanding of where and how pleiotropic GWAS variants act.

In this study, we use metabolites as model traits to understand pleiotropic features of genetic architecture. Metabolites are small molecules interconverted by a series of biochemical pathways, and are an appealing model system for studying pleiotropy because their pathways are typically well-documented and biologically simpler than those underlying other complex traits [6, 7]. Previous work in Mendelian genetics has identified inborn errors of metabolism (IEM) in many enzymes [8]. Metabolite GWAS, which have long observed pervasive pleiotropy at these IEM genes and other loci [9, 10], offer a potential opportunity to further explore the relationships between intermediate molecules and disease outcomes at scale. Here, we jointly analyzed GWAS results of 16 plasma metabolites from the Nightingale Health Nuclear Magnetic Resonance (NMR) Spectroscopy platform in nearly 100,000 individuals in the UK Biobank [11] (Figure 1; see Methods). These 16 metabolites included glucose, pyruvate, lactate, citrate, isoleucine, leucine, valine, alanine, phenylalanine, tyrosine, glutamine, histidine, glycine, acetoacetate, acetone, and 3-hydroxybutyrate. They were chosen based on their biochemical proximity to each other, their relevance to health and disease, and because the genes and enzymes involved in their metabolism are well-characterized. They play especially important roles in energy generation and energy storage pathways such as glycolysis, the citric acid cycle, amino acid metabolism and ketone body formation. They are relevant to many metabolic diseases including type 2 diabetes [12, 13, 14], cardiovascular disease [15], and non-alcoholic fatty liver disease [16].

Numerous GWAS have begun characterizing the genetic architecture of metabolites and found them to be heritable and polygenic [17, 18]. Recent metabolite studies have shown that leveraging information about the biochemical pathways relevant to a given metabolite [19, 20, 21, 7] can allow for more interpretable gene annotation of GWAS hits. This has led to the dissection of individual associations of biomarkers, such as lipids [22], glycine [23], and intermediate clinical measures [24], with cardiometabolic and other diseases. The pervasive pleiotropy at these GWAS loci with other metabolites as well as disease [24, 25] suggests the potential of utilizing these data for investigating the mechanism of pleiotropic effects as a core component of genetic architecture. While recent GWAS have begun jointly investigating multiple metabolites [26, 27, 28], they have yet to do so in the context of their biochemical pathways.

In this paper, we demonstrate that investigating the effects of pleiotropic variants on biologically-related metabolites allows for a better understanding of why these variants have their observed joint effects. Our results reveal striking heterogeneity in genetic correlation across the genome and provide a biologically intuitive basis for understanding this heterogeneity. Together, this allows us to dissect the molecular basis of metabolic disease GWAS variants and enables us to directly define the mechanism relating an example variant to its associated disease.

Results

Insights into the shared genetic architecture of biologically-related metabolites

We chose 16 metabolites from the 249 available through the Nightingale NMR platform in a subset of the UK Biobank (Figure 1; see Methods). These 16 metabolites were selected based on their biochemical proximity, relevance to health and disease, and because the genes and enzymes involved in their metabolism are well-characterized. We classified the 16 metabolites into four groups based on shared biochemistry: Glycolysis (glucose, pyruvate, lactate, citrate), Branched Chain Amino Acid (BCAA; isoleucine, leucine, valine), Other Amino Acid (alanine, phenylalanine, tyrosine, glutamine, histidine, glycine), and Ketone Body (acetoacetate, acetone, 3-hydroxybutyrate). Trait measurements were log-transformed and adjusted for relevant technical covariates. After outlier removal, we obtained a primary dataset of 94,464 genotyped European-ancestry individuals with data for all 16 metabolites.

We first sought to characterize the genetic architecture underlying these metabolites by performing GWAS for each (Supplementary Figure 2 - figure supplement 1). Hits from individual GWAS were clumped with an r^2 of 0.01 per megabase, combined across metabolites, then pruned to the SNP with the most significant P-value within 0.1 cM. This resulted in 213 lead variants with a genome-wide significant association in at least one metabolite, referred to as the metabolite GWAS hits. Glycine had the largest number of significant associations with 77 hits (Figure 2a). There were 47 variants with significant associations in more than one metabolite, including rs2939302 (near the gene *GLS2*) which was significant in 9 of the 16 metabolites, and rs1260326 (*GCKR*) which was significant in 8. Glycine also had the highest total SNP heritability of 0.284 (Supplementary File 1 and Supplementary Figure 2 - figure supplement 2).

To understand the shared genetics of these metabolites, we then investigated the extent of pleiotropy between and within biochemical groups. In order to examine this, we first calculated pairwise LDSC genetic correlation across the 16 metabolites. We found substantial genome-wide sharing for many pairs of metabolites, especially for metabolites within the same biochemical group (Figure 2b; phenotypic correlation in Supplementary Figure 2 - figure supplement 3). We then explored pleiotropic effects beyond the polygenic background by examining the structure within the metabolite GWAS hits. Pairwise Mendelian Randomization (MR) between the metabolites emphasized the intertwined nature of these traits (Figure 2c). Despite only taking into account genetic effects, MR largely clustered metabolites in a way that reflects their biochemical groups. The extensive pleiotropy across the 16 traits, with similar sharing inside biochemical groups, is also illustrated by the structure visible in the normalized effect sizes for each metabolite GWAS hit (Figure 2d). Together, these analyses support substantial, but not always consistent, genetic overlap between the traits, particularly in the polygenic components. In the remainder of this paper, we will seek a deeper understanding of the biochemical relationships between genotypes and metabolite levels.

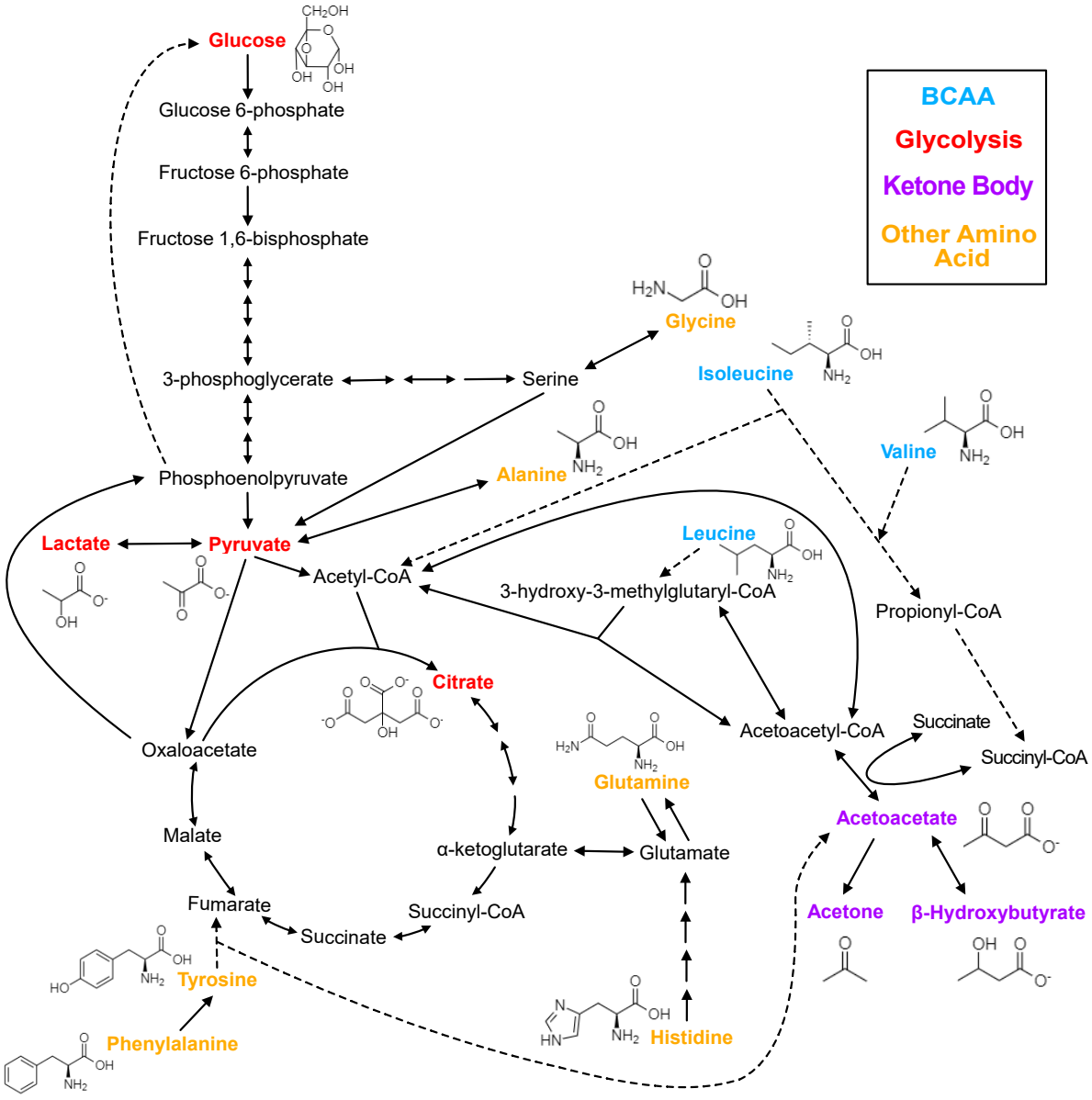

Figure 1: Biochemistry of relevant metabolites. *Pathway diagram and molecular structure of relevant metabolites, colored by their biochemical groups. The pathway diagram was curated from multiple resources (see Methods). All solid lines represent a single chemical reaction step. Dotted lines represent a simplification of multiple steps. For simplicity, only a subset of all the reactions each metabolite participates in is shown. Genes encoding the enzymes that catalyze the above chemical reactions are known and presented in Supplementary Figure 3 - figure supplement 1.*

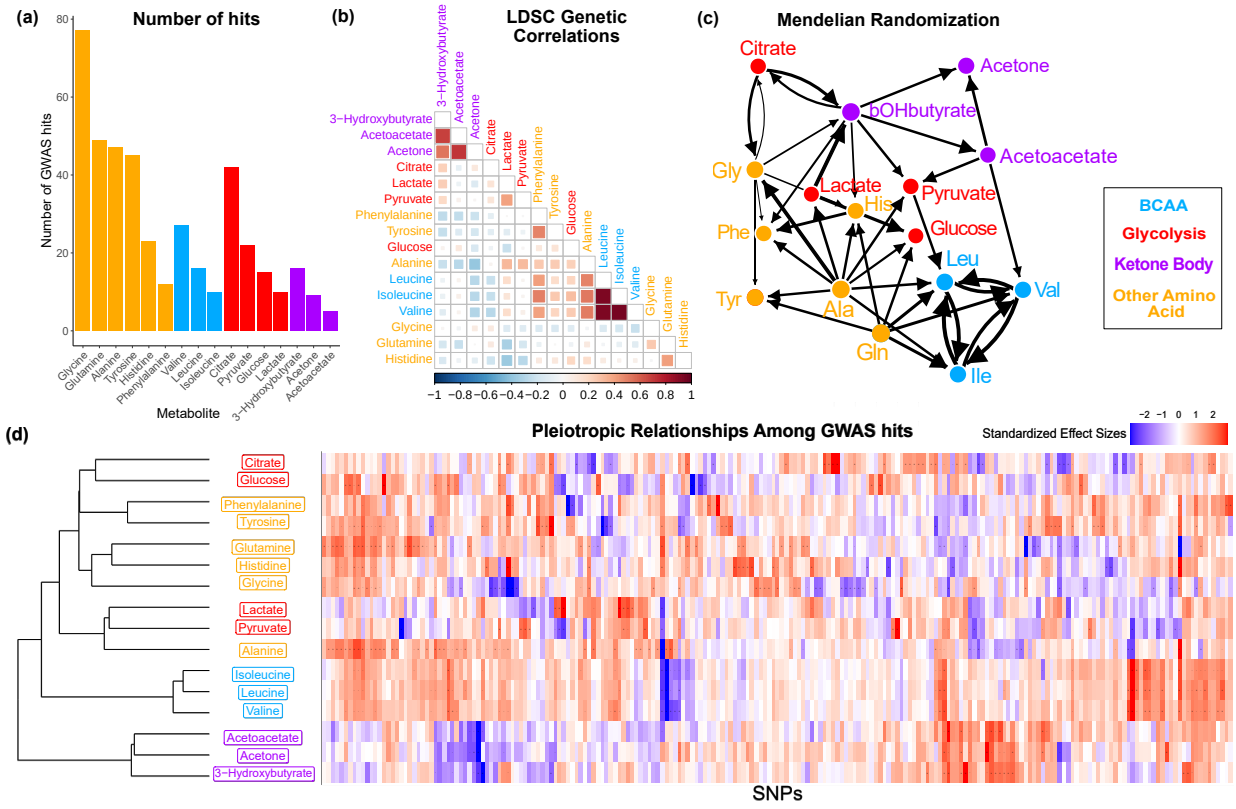

Figure 2: Overview genetic architecture of metabolites. **a.** *Number of GWAS hits per metabolite.* **b.** *Pairwise LDSC genetic correlations between the metabolites, clustered by genetic correlation.* **c.** *Mendelian Randomization weighted results between the metabolites.* **d.** *Biclustered standardized effect size in each metabolite for the 213 metabolite GWAS hits. For visualization, effect sizes were divided by the standard error then inverse normal transformed and standardized. Each variant was aligned to have a positive median score across metabolites.*

Characterizing the biological functions of candidate genes

An important step in understanding the pathway level mechanisms of variants is knowing which gene a variant is affecting and how that gene relates to the biology of the pathway. Different types of genes influence trait biology through distinct mechanisms. Metabolite biology is documented in genetic and biochemical databases based on the extensive history of biochemical research (Supplementary File 2). Thus, we developed a pipeline for annotating the 213 metabolite GWAS hits with a single most likely gene using gene proximity and manual curation of these databases (Supplementary File 3 and Supplementary Figure 2 - figure supplement 4; see Methods). We annotated 68 variants with genes encoding pathway-relevant enzymes (25-fold enrichment, 95% CI [20-fold, 33-fold], Poisson rate test P $<$ 2e-16), 46 with genes encoding transporters (5.2-fold enrichment, 95% CI [3.7-fold, 7.2-fold], P = 9e-16), and 30 with genes encoding transcription factors (7-fold enrichment among liver marker TFs, 95% CI [3.0-fold, 14-fold], P = 3e-5; Figure 3). Overall, 69% of variants were assigned to the closest gene and 49% of variants assigned to a pathway-relevant enzyme gene were assigned known inborn errors of metabolism (IEM) genes [8]. The substantial enrichment for biologically interpretable variants suggests that examining the genetic basis of these traits will allow for the development of hypotheses around relevant molecular mechanisms underlying pleiotropy.

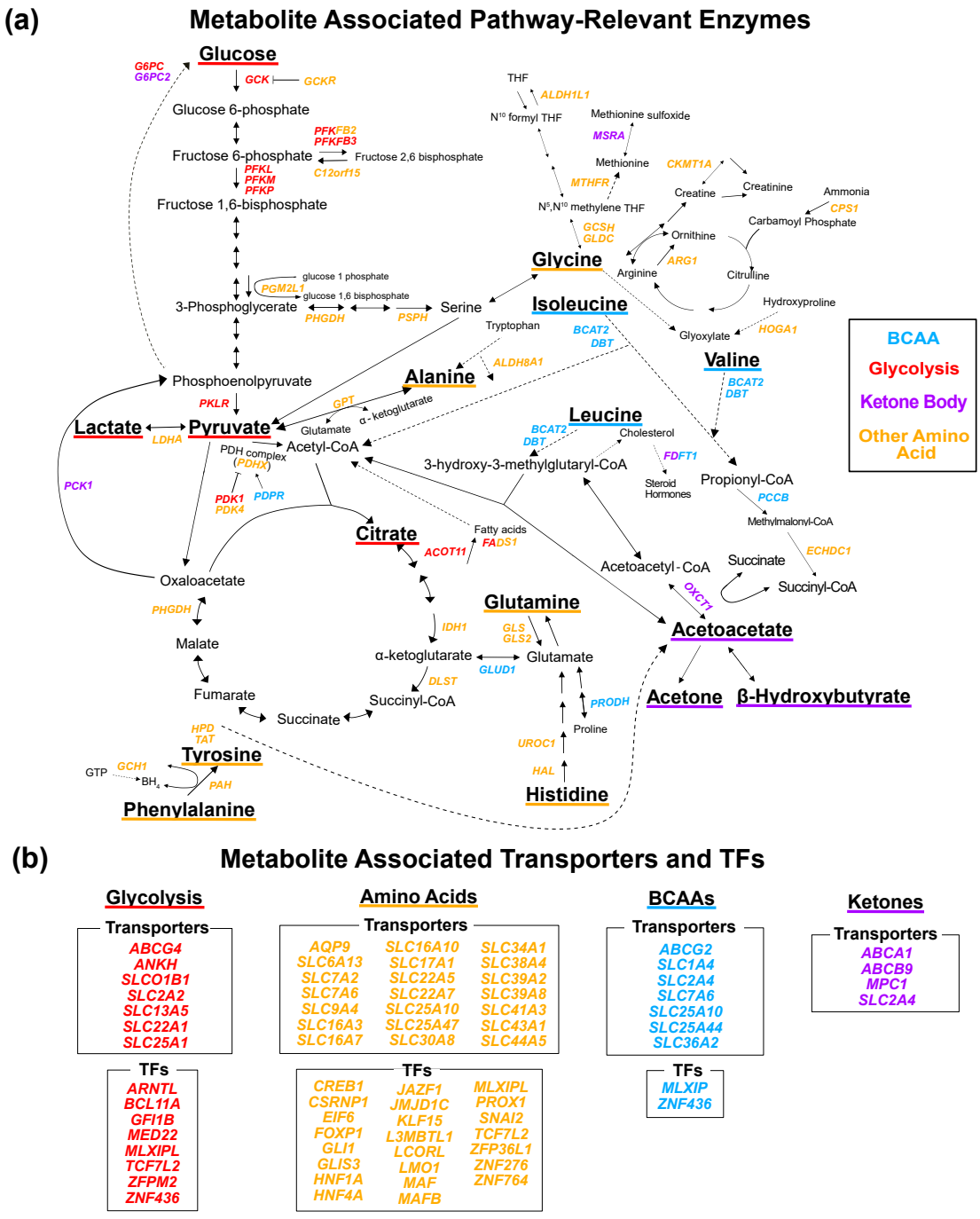

Figure 3: Gene annotation of metabolite GWAS hits. *Each gene is colored based on the biochemical group with the most associated metabolites ($P < 1e\text{-}4$). If multiple biochemical groups are tied for the most associations for a given gene, they are all shown.* **a.** *Expanded pathway diagram with all genes (italicized) that encode pathway-relevant enzymes and were a metabolite GWAS hits.* **b.** *List of all genes of the metabolite GWAS hits that encode transporters and TFs. There were 69 metabolite GWAS hits that are not shown. Of these, 60 were annotated with genes assigned to the gene type general cell function (14 of these 60 were related to lipid function), and 9 were assigned to a gene of unknown function or that did not have any genes nearby (see Methods).*

Next we sought to understand which genes and subpathways were most relevant to each biochemical group. We assigned each gene to the biochemical group with the most associated metabolites (Supplementary File 4; see Methods). Genes were largely assigned to the group whose relevant biology was nearest the protein encoded by the gene. For example, *BCAT2* encodes an enzyme responsible for the first step in the breakdown of all three BCAAs and was assigned to the BCAA group. *OXCT1* encodes an enzyme responsible for the conversion of acetoacetyl-CoA to the ketone body acetoacetate and was assigned to the Ketone Body group. Similarly, *SLC7A9* encodes a protein that transports amino acids and was assigned to the Other Amino Acid group, while *TCF7L2* is a TF assigned to the Glycolysis group and involved in blood glucose homeostasis. These results confirm that these variants are affecting known trait-relevant biology and reflecting the local structure of these pathways.

Interestingly, a large fraction of the genes involved in trait-relevant biology were genome-wide significant hits for at least one of the 16 metabolites. Specifically, of the 139 total genes encoding enzymes in the pathway diagram for these metabolites (Supplementary Figure 3 - figure supplement 1), 51 genes had at least one GWAS hit. Additionally, we performed an ancestry-inclusive GWAS of all 98,189 individuals with complete metabolite data for followup analysis. In this ancestry-inclusive analysis, we identified 41 additional hits not found in the European-only GWAS, including associations at 7 additional pathway-relevant genes (Supplementary File 5; Supplementary Figure 3 - figure supplement 2). This highlights the potential for large-scale, ancestry-inclusive GWAS to discover more biochemically-relevant associations among these traits. Together, these findings suggest that GWAS reflect, and have the potential to illuminate, the complex biochemical pathways interconverting these metabolites.

Investigating the mechanisms of pleiotropy in trait pairs

Given the overlap between the biology of these metabolites and their hits, we next sought to understand the molecular causes of pleiotropy in trait pairs. We found 26 genetically correlated metabolite pairs at a local false sign rate < 0.005. For example, alanine and its strongest genetic correlation partner, isoleucine, share a genetic correlation of $r_g = 0.52$ (SE = 0.05, P = 9e-23). Similarly, plotting the effects of the 213 GWAS variants on these two traits indicates a strong positive correlation (Figure 4a). Nonetheless, we noted several outlier loci, including rs370014171 (*PDPR*) and rs77010315 (*SLC36A2*), which have strong discordant effects. We were intrigued to understand why these two variants had discordant effects on alanine and isoleucine relative to their overall positive genetic correlation, while the majority of other variants had concordant effects.

Outlier variants are appealing case studies for understanding the molecular basis of pleiotropy because they affect traits in an exceptional way. Thus, we reasoned that understanding large-effect variants inconsistent with the global genetic correlation would reflect interesting biology relevant to the traits. For example, the proteins encoded by *PDPR* and *SLC36A2* are both located between alanine and isoleucine in the biochemical pathway (Figure 4b). This suggests that where variants act in the pathway may influence the direction of effect they have on metabolites. To better understand how these two variants affect alanine and isoleucine and explain their outlier behavior, we examined their effect size and direction in the context of their location in the pathway. We then used the variants' metabolite associations to develop candidate mechanisms for how each variant could be jointly influencing the levels of these metabolites.

As an illustration, we first consider variant rs370014171. This variant was assigned to gene *PDPR* because it was the second closest gene, the closest pathway-relevant enzyme, and within 100

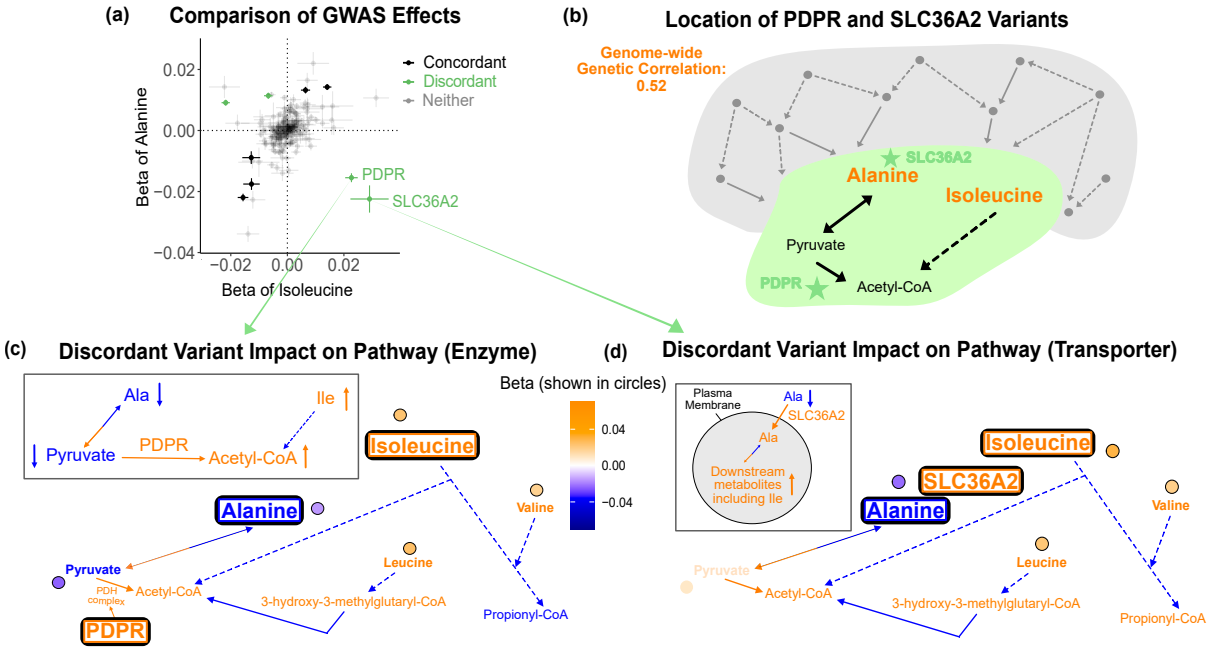

Figure 4: Discordant variant analysis. a. *Comparison between the effects on alanine levels versus the effects on isoleucine levels for the 213 metabolite GWAS hits. This highlights two discordant variants: rs370014171 (PDPR) and rs77010315 (SLC36A2). Variants without an association $P < 1e\text{-}4$ in both metabolites are labeled "Neither".* **b.** *Graphical representation of where these two discordant variants act in the pathway, represented by green stars, relative to other upstream variants driving the positive genetic correlation. Below are the hypothesized mechanisms explaining the GWAS results in relevant metabolites for each of these discordant variants. Data are shown in circles with the coloring corresponding to the effect (beta) of that variant on that metabolite. A black outline represents an association with $P < 1e\text{-}4$. Orange text and arrows represents a hypothesized increase (direction, not magnitude) in flux and blue corresponds to a decrease.* **c.** *Results for rs370014171 near the gene PDPR which encodes a protein that activates the conversion of pyruvate to acetyl-CoA. All solid lines represent a single chemical reaction step. Dotted lines represent a simplification of multiple steps.* **d.** *Results for rs77010315 in the gene SLC36A2 which encodes a small amino acid transporter.*

kb (12.3 kb to its gene boundaries). PDPR activates the enzyme that catalyzes the conversion of pyruvate to acetyl-CoA (Figure 4c; Supplementary Figure 4 - figure supplement 1). A candidate mechanism for this variant, supported by the effect size and direction for the 16 metabolites where relevant, is that it increases PDPR activity. There was colocalization of association signals across the 5 significant metabolites using both conditional SNP-level analyses (Figure 4 - figure supplement 2) and running `coloc` once adjusting for secondary signals at alanine (Figure 4 - figure supplement 3 and 4 - figure supplement 4). This would lead to increased conversion of pyruvate to acetyl-CoA and thus decreased pyruvate (β = -0.023 SDs, SE = 0.003, P = 3e-20). To compensate for the subsequent decreased pyruvate levels, there would be increased conversion of alanine to pyruvate causing a decrease in alanine. In response to the increased acetyl-CoA, there would be decreased breakdown of metabolites normally catabolized for its production, including isoleucine, resulting in an increase in isoleucine levels. Thus, this variant has an opposite effect on alanine and isoleucine, despite their overall positive genetic correlation, likely because it affects the activity of an enzyme that acts in the pathway *between* the pair of metabolites. As expected due to the high correlation between the levels of the three BCAAs, this variant is also a discordant variant for alanine with valine (r_g = 0.51, SE = 0.05, P = 2e-21), and alanine with leucine (r_g = 0.49, SE = 0.06, P = 1e-16).

As a second example, variant rs77010315 is a missense variant in *SLC36A2*. *SLC36A2* encodes a transporter for small amino acids such as alanine (Figure 4d; Supplementary Figure 4 - figure supplement 5). There was colocalization of association signals across the four significant metabolites using both conditional SNP-level analyses (Figure 4 - figure supplement 6) and running `coloc` (Figure 4 - figure supplement 7). A candidate mechanism explaining the observed metabolite associations in our data and outlier behavior for this variant is that it increases transport of alanine into cells by SLC36A2. This would result in a decrease in levels of alanine in the blood, but an increase of alanine in cells. This additional intracellular alanine would then allow for increased conversion of alanine to pyruvate, thereby increasing levels of downstream metabolites in the blood, including isoleucine. Thus, this variant has an opposite effect on alanine and isoleucine, despite their overall positive genetic correlation, but in this case because it affects biology between the metabolites at the transporter level.

Quantifying global properties of molecular pleiotropy

Based on these results, we hypothesized that the two variants described above, and others like them, exhibit outlier behavior because they affect biology *between* the two metabolites (Figure 5a). We consider biology "between" a given pair of metabolites as the shortest biochemical path connecting them, which can include a path converting one metabolite to the other, as well as other scenarios such as those involving colliders. This is because all biochemical reactions, including one-way reactions, can have bi-directional causal relationships due to Le Châtelier's Principle (see Methods for details; Supplementary Figure 5 - figure supplement 1). Genetic correlation reflects the direction of effect that most associated variants have on two traits. However, when two metabolites are biologically near each other, the region containing “between” biology is relatively small, such that only a minority of variants directly affect the “between” region.

Thus, we hypothesized that the genetic correlation of two biologically-related metabolites mostly reflects the effects of variants upstream or downstream of the metabolites, masking the effects of those between. We developed an analogous hypothesis that variants affecting biology upstream or downstream of the two metabolites have concordant effects (Figure 5b). While less common, the overall genetic correlation for two biologically-related metabolites can also be negative due to

factors such as feedback loops. In this case, variants acting between the two metabolites would have the same direction of effect on both metabolites, making them discordant with the negative overall genetic correlation (Supplementary Figure 5 - figure supplement 2).

To evaluate these models, we defined outliers based on the consistency of their effects with the overall LDSC genetic correlation. If a variant had an effect direction opposite the overall LDSC genetic correlation in at least one significant metabolite pair (P $<$ 5e-8 in one, P $<$ 1e-4 in the other), it was classified as "discordant". For example, a discordant variant for a metabolite pair with a positive genetic correlation would have a negative association in one of the metabolites and a positive association in the other. If a variant had an effect direction consistent with the overall genetic correlation for its significant metabolite pairs, it was classified as "concordant". Variants without multiple associations, or where associated traits were not significantly genetically correlated, were classified as "neither". In total, of the 62 metabolite GWAS hits that had at least one significant metabolite pair, we found 26 total discordant variant-metabolite pairs across 14 variants (Supplementary File 6).

We then investigated overall properties of discordant variants relative to concordant ones. We discovered that discordant variants are more likely to affect genes encoding enzymes and transporters than all other genes types, including TFs, general cell function genes, and those of unknown function (Odds Ratio = 4.09, 95% CI [1.08, 23.1], P = 0.034; Figure 5c). This is in contrast to concordant variants, which do not show an enrichment for enzymes and transporters relative to other gene types (Odds Ratio = 0.75, 95% CI [0.38, 1.48], P = 0.42). These observations are consistent with our model that discordant variants tend to affect biology between relevant pairs of metabolites since TFs and general cell function genes generally act outside these metabolic pathways. Thus, they are more likely to affect biology upstream or downstream of both metabolites. In addition, for variants affecting pathway-relevant enzymes, where the location in the pathway that the variant is acting relative to the metabolites is clear, we were able to directly test our hypothesis. We found that discordant variants affecting pathway-relevant enzymes are much more likely to act between, rather than upstream or downstream, the metabolites for which they are discordant (Odds Ratio = 23.0, 95% CI [1.58, 1510.45], P = 0.0072; Figure 5d).

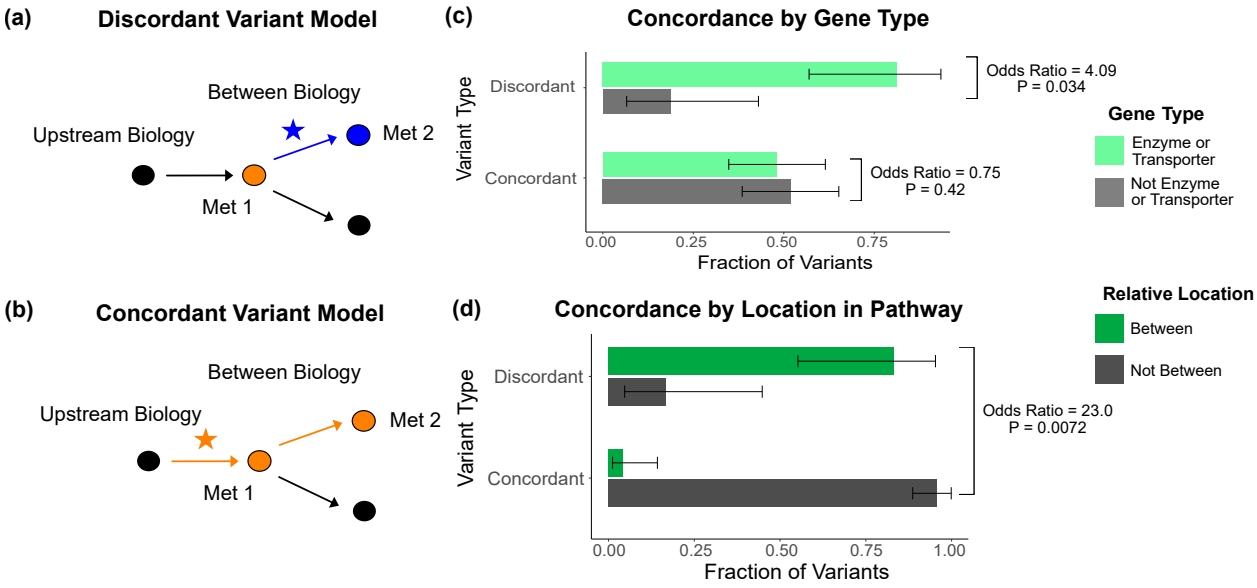

Figure 5: Characterization of discordant and concordant variants. a. *Proposed model for the mechanism of a discordant variant. This example is for a discordant variant that has opposite effect directions on a pair of metabolites with a positive overall genetic correlation because it affects biology between them.* **b.** *Proposed model for the mechanism of a concordant variant. This example is for a concordant variant that has the same effect direction on a pair of metabolites with a positive overall genetic correlation because it affects biology upstream both metabolites.* **c.** *Fraction of the discordant and concordant variants that have a pathway-relevant enzyme or transporter gene type annotation versus those with a different gene type annotation. Discordant variants are enriched for the gene types of pathway-relevant enzyme or transporter, as would be expected in the model of discordant variants generally affecting biology between metabolites.* **d.** *Fraction of the discordant and concordant variants annotated with a pathway-relevant enzyme that affect biology between versus not between their significant metabolite pairs. Significance tests were performed using Fisher's Exact method and the plotted SEs are from 95% CI calculated by Wilson Score Interval.*

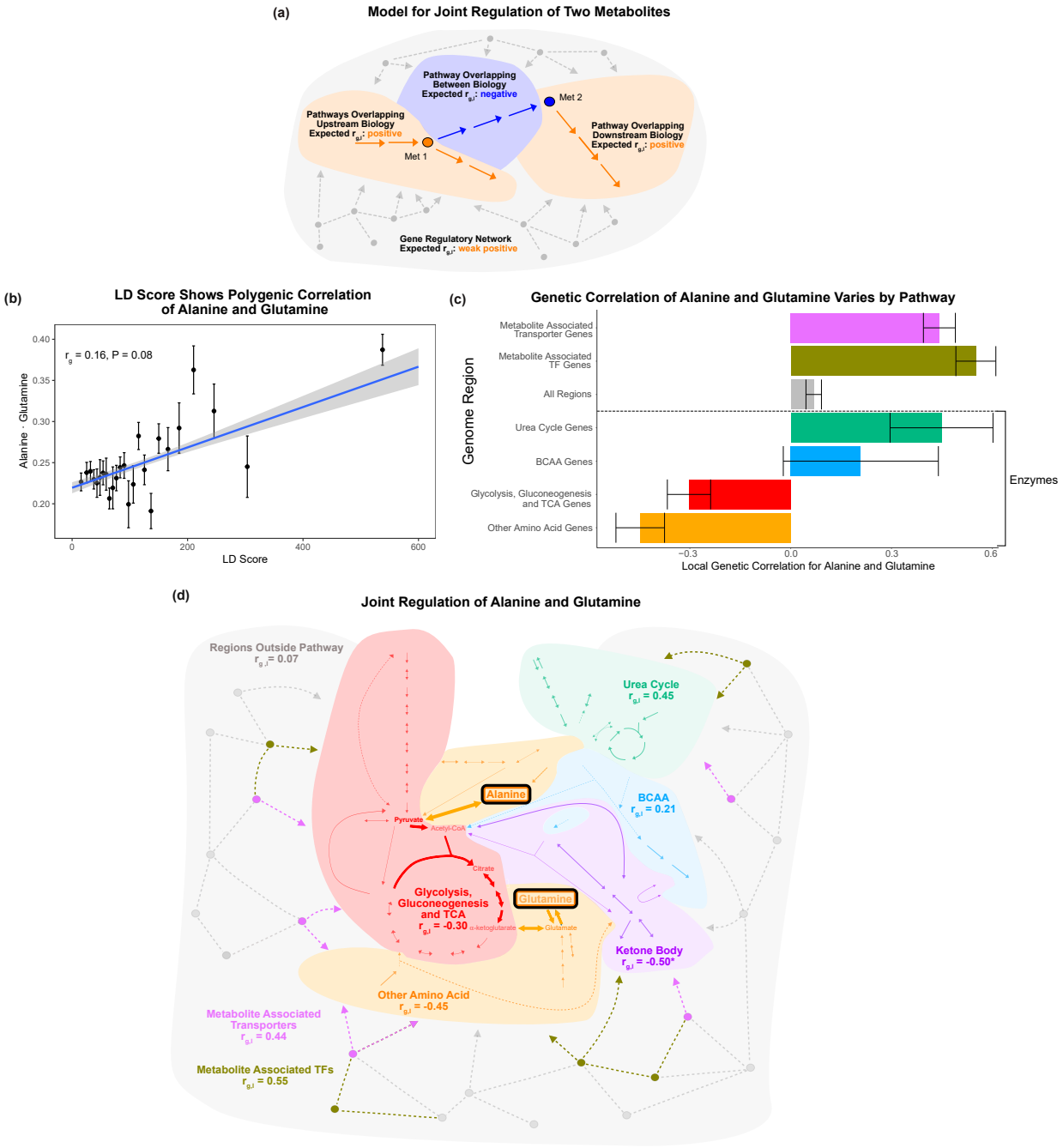

Figure 6: Local genetic correlation. **a.** *Model of expected local genetic correlation direction, with contrasting effects of variants affecting "between" versus outside biology at a pathway and genome-wide level.* **b.** *LD Score showing the polygenic correlation of alanine and glutamine. For the x-axis, LD Scores were binned into 25 bins. The y-axis shows the mean and SE within each bin.* **c.** *Results for the local genetic correlation of alanine and glutamine for variants within 100 kb of genes in each pathway. Standard errors are shown. Genesets listed below the dotted line include only enzymes and are considered pathway-relevant enzymes for these metabolites. Summary statistics for BOLT-REML and other methods can be found in Supplementary File 7.* **d.** *Pathway diagram showing the pathways included in the local genetic correlation analysis and the positioning of their genes relative to alanine and glutamine. *Ketone Body Genes were omitted from panel c because the limited number of genes meant they failed to robustly converge. All arrows and nodes in the gray section are hypothetical and shown for illustration purposes.*

We then sought to extend this finding by developing a model contrasting the effects of all variants affecting between versus outside biology at a pathway and genome-wide level (Figure 6a). In aggregate, this model predicts that pathways overlapping biology between two metabolites will have a local genetic correlation opposite that of nearby adjacent pathways and that the magnitude of both will exceed that of the global polygenic background. As a case study, we focused on alanine and glutamine, which have a weak positive overall genetic correlation (r_g = 0.16, SE = 0.09, P = 0.08; Figure 6b; Supplementary Figure 6 - figure supplement 1). We then ran BOLT-REML [29] on variants within 100 kb of genes in each pathway and estimated the corresponding local genetic correlations (see Methods).

We found that the local genetic correlations around genes in the Glycolysis, Gluconeogenesis and Citric Acid Cycle Pathway and around genes in the Other Amino Acid Pathway were negative (Figure 6c). Both of these pathways encompass genes affecting biology between alanine and glutamine (Figure 6d). In striking contrast, nearby pathways, such as the Urea Cycle, had a positive local genetic correlation for these metabolites ($r_{g,l}$ = 0.45; SE = 0.15, P = 0.003). Similarly, we found that regions overlapping genes encoding metabolite associated transporters and TFs had strong positive genetic correlations consistent with their shared role in the upstream regulation of these two traits ($r_{g,l}$ = 0.44, SE = 0.05, P = 1e-20; $r_{g,l}$ = 0.55, SE = 0.06, P = 2e-20). All genes outside the core pathways had a weak positive genetic correlation, perhaps reflecting that they are embedded in the global gene regulatory network ($r_{g,l}$ = 0.068; SE = 0.02, P = 0.003). Our findings were broadly consistent using individual level data with Haseman-Elston regression [30], and summary statistics with ρ-HESS [3], stratified LD score regression [31] and a non-parametric Fligner-Killeen variance test (see Methods; Supplementary File 7). These results support the model that variants affecting biology between the metabolites frequently contrast with the contributions of upstream and downstream pathways. This emphasizes that the heterogeneity in genetic effects reflecting local biology shared by the traits can be masked in the global genetic correlation. In addition, these results offer biological intuition for interpreting genetic correlation of molecular traits at a pathway and genome-wide level.

Using metabolites to understand the mechanism of a disease-associated variant

Motivated by the interpretability of these results, we applied this logic to develop an example model for a variant associated with increased risk for a disease (Figure 7a). In this model, we hypothesized one mechanism for how a variant could be associated with increased risk for a disease is that it could impact metabolites in a way that is consistent with disease etiology. For example, the variant could increase metabolites associated with increased risk for the disease and/or decrease metabolites associated with decreased risk.

To apply this model to our data, we considered metabolite GWAS hits that were annotated with pathway-relevant enzymes and associated with increased risk for coronary artery disease (CAD) [32, 33]. The variant that best fit these criteria were rs61791721. This variant was assigned the nearest pathway-relevant enzyme gene, *PCCB* which encodes a protein that catalyzes the conversion of propionyl-CoA to succinyl-CoA at the intersection of BCAA and fatty acid oxidation (Supplementary Figure 7 - figure supplement 1).

We combined results from the literature and incident analysis to understand the association of relevant metabolites with CAD (Supplementary Figure 7 - figure supplement 2). We then compared these to the effects of this variant on these metabolites (Figure 7b). In this analysis we included high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), total fatty

acids, and total triglycerides, due to the extensive evidence implicating their association with CAD, and because they are directly adjacent to the biology of the other 16 metabolites. Consistent with the metabolites' corresponding direction of risk for CAD, this *PCCB* variant was negatively associated with glycine and HDL-C, and positively associated with isoleucine, leucine, valine, tyrosine, total fatty acids, total triglycerides and LDL-C (P < 1e-5; Supplementary Files 8 and 9).

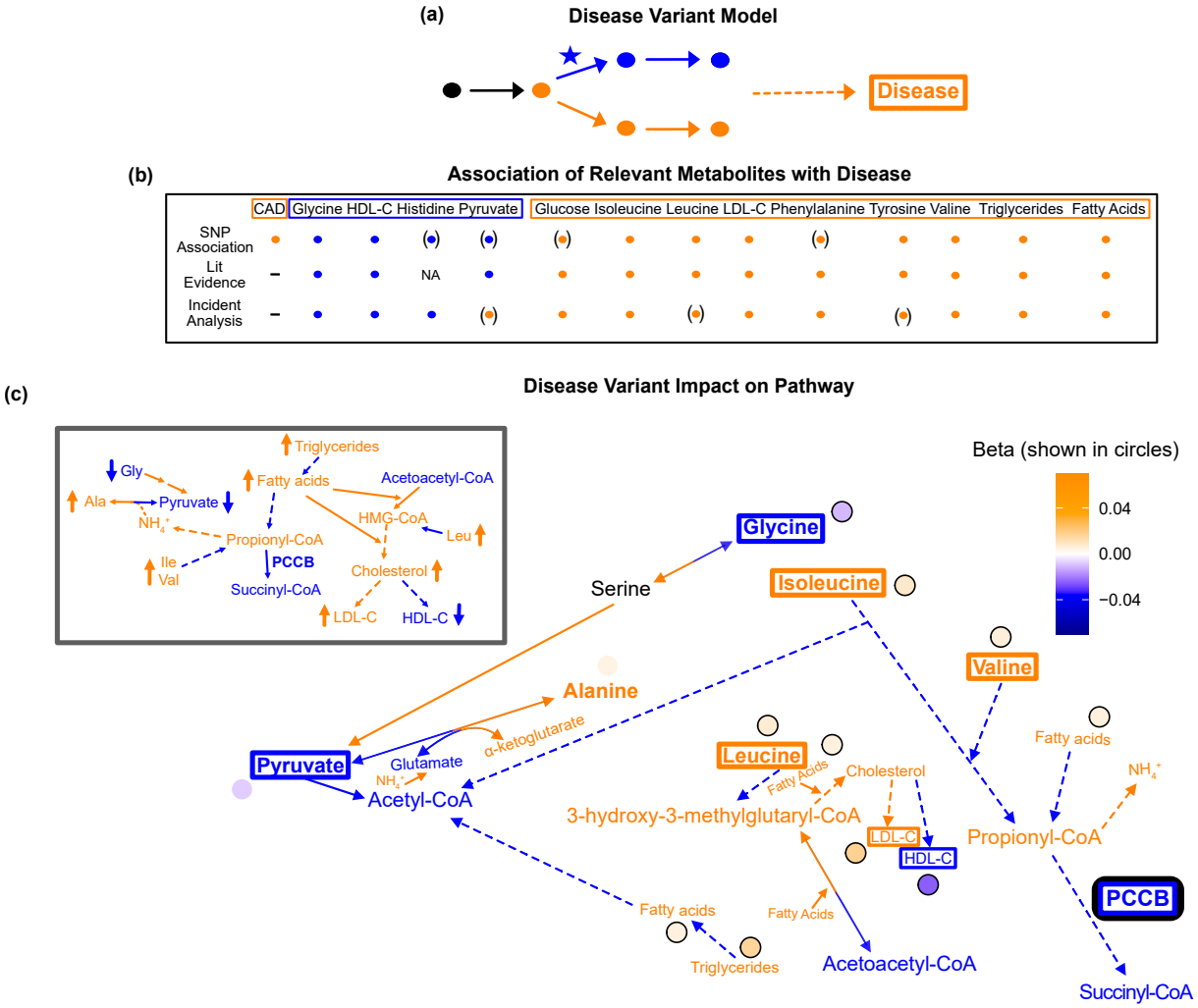

Figure 7: Pathway impact and pathology of example disease GWAS hit. a. *Proposed model for the impact of a disease hit on a relevant pathway, contributing to an increased risk in the disease.* **b.** *This variant is associated with an increase in levels of metabolites that have been implicated with increased risk of CAD, and a decrease in the levels of metabolites that have been implicated with decreased risk. Parentheses indicate nonsignificant associations, "NA" indicates no evidence was found, and "-" indicates a placeholder because CAD is being compared with itself.* **c.** *Results for rs61791721 with gene assignment PCCB which encodes a protein that catalyzes the conversion of propionyl-CoA to succinyl-CoA. The hypothesized mechanism is that the variant is decreasing the activity of PCCB, resulting in the above metabolite associations. Ammonium is represented by its chemical formula (NH_4^+). Data are shown in circles with the coloring corresponding to the effect (beta) of that variant on that metabolite. All solid lines represent a single chemical reaction step. Dotted lines represent a simplification of multiple steps.*

This PCCB variant has been associated with CAD in multiple prior GWAS [32, 33], yet neither the gene this variant affects nor the biological mechanism explaining its association with CAD are known. However, this variant affects many metabolites associated with CAD in a direction consistent with increased risk. Thus, we hypothesized that we could begin to understand why this variant is associated with CAD by understanding the pleiotropic effects of this variant on the metabolites.

A potential mechanism that we hypothesized could result in this pathogenic constellation of metabolite effects is that the variant could decrease PCCB activity, resulting in lower levels of succinyl-CoA and increased propionyl-CoA (Figure 7c). Consistent with this model, there was colocalization of association signals across all associated traits other than lipids using both conditional SNP-level analyses (Figure 7 - figure supplement 3) and running `coloc` (Figure 7 - figure supplement 4). The increased propionyl-CoA could result in excess ammonium being produced, and because alanine is a reservoir for nitrogen waste, this would increase conversion of pyruvate to alanine to capture the toxic ammonium [34, 35]. More glycine may then be broken down in response to the decrease in pyruvate levels, decreasing glycine levels. Conversely, the increased levels of propionyl-CoA would likely mean less valine, isoleucine, fatty acids, and thus triglycerides, would need to be broken down, resulting in an increase in their levels. This increase in fatty acids may stimulate the activity of 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase and synthase, resulting in an increase in HMG-CoA and cholesterol [36, 37]. Increased HMG-CoA could lead to increased leucine because less leucine would need to be broken down to produce HMG-CoA, while increased cholesterol would lead to an increase in LDL-C and a decrease in HDL-C. Therefore, this variant is potentially associated with CAD because it is decreasing PCCB activity, resulting in myriad deleterious downstream metabolic consequences.

While *in vivo* functional validation would be needed to draw causal conclusions about the effect of this variant on these metabolites and of these metabolites on CAD, this example demonstrates that we can begin to generate informed hypotheses and dissect the molecular basis underlying disease GWAS hits by understanding the mechanism of relevant pleiotropic effects on metabolites. In addition, the pathways implicated by this analysis can also be independently prioritized as potentially playing an important role in cardiometabolic disease by leveraging the molecular basis of genetic correlation discussed in Figure 6. For example, alanine and glutamine have opposite associations with CAD and type 2 diabetes despite having an overall positive phenotypic correlation [38, 39]. This suggests that the pathways described above with a negative local genetic correlation for alanine and glutamine are likely relevant to the molecular basis of these diseases. Thus, understanding the molecular basis of pleiotropy and genetic correlation of metabolites can improve our understanding of the variants and pathways contributing to complex disease biology.

Discussion

In this work, we investigate the joint effects of pleiotropic variants on 16 biologically-related metabolites in the context of their biochemical pathways. We build on prior studies examining the genetic architecture of metabolites by characterizing the genes and mechanisms through which variants affect these metabolites, and find a strong enrichment for genes encoding pathway-relevant enzymes and transporters. Our results offer biological intuition for the interpreting genetic correlation of molecular traits at a pathway and genome-wide level.

We demonstrate the effects of variants acting on biology between metabolites often contrast substantially with the contributions of upstream and downstream pathways, as well as the polygenic

background. Perhaps paradoxically, while the overall genetic correlation between two traits provides a global view of shared effects, the genes that are directly involved in the traits' core biology are most likely to have divergent effects. We show that one explanation of this is the substantial outlier contributions from variants acting directly between metabolites of interest. We anticipate that further mechanisms, such as context-specific variant effects and differential regulation by peripheral genes, will be discovered in future studies.

In addition, we show specific examples of candidate molecular mechanisms explaining the association of variants with multiple biologically-related metabolites. These include associations at *PDPR*, *SLC36A2*, and *PCCB*, where we show that the direction and magnitude of their effects is consistent with metabolite biochemistry and disease etiology. These proposed molecular mechanisms enable biological prioritization of interesting candidates for future post-GWAS *in vitro* studies. Overall, these results suggest specific genetic and molecular underpinnings of complex disease variants, and provide a roadmap for further discovery through the interpretation of pleiotropic variant effects on disease-relevant metabolites.

In this work, we focus on metabolites clustered at the intersection of amino acid catabolism, glycolysis, and ketone body metabolism. However, the approaches and results from this paper have the potential to reveal novel insights into genetic effects on many biochemical pathways and molecular traits. In addition, integrating this work with proteomic and intermediate metabolomic data will offer additional evidence to develop and support these hypothesized mechanisms. These data may also clarify the relevance of additional mechanisms – such as buffering, feedback, and kinetics – in controlling the plasma levels of these metabolites. Finally, expanding the sample size and diversity of ancestries included in future GWAS, measurements of which are currently underway on the Nightingale platform and others, will increase power to detect novel findings such as important associations for variants with low allele frequency.

While we largely focused on the molecular trait space here, many of these concepts may be useful in the endophenotype and disease space as well. For instance, this approach may help identify variants and pathways most relevant to the core shared biology of a given pair of diseases, potentially revealing more about the molecular bases of the diseases and prioritizing additional drug target candidates. For example, discordant variants have already been identified for some disease pairs, such as for BMI and type 2 diabetes [40], and non-alcoholic fatty liver disease and type 2 diabetes [41].

One limitation of this study is that the metabolites were measured in the blood, while most of the relevant biology and pathways occurs within cells in various tissues throughout the body. Thus, we anticipate extensions of this work to include biomarker measurements from additional cell types and tissues, such as urine, saliva, biopsy samples, and *in vitro*-differentiated cells. Further, longitudinal analysis of relevant disease cohorts will allow insights into disease progression and subtyping.

In conclusion, this work underscores the potential of unifying biochemistry with genetic data to understand the molecular basis of complex traits and diseases and the mechanism through which variants impact these traits.

Online Methods

Population Definition

We defined our GWAS population as a subset of the UK Biobank [42]. For our cohort, we use the individuals for which Nightingale plasma metabolite data was available after filtering based on trait QC characteristics (see "Trait QC and Covariate Adjustment"). We then filtered individuals by the following QC metrics:

1. Not marked as outliers for heterozygosity and missing rates (`het_missing_outliers` column)
2. Do not show putative sex chromosome aneuploidy (`putative_sex_chromosome_aneuploidy` column)
3. Have at most 10 putative third-degree relatives (`excess_relatives` column).
4. No closer than second degree relatives.

From these, we defined 3 cohorts: White British, non-British White, and everyone. We identified White British individuals using the `in_white_British_ancestry_subset` column in the sample QC file. We identified non-British White individuals through self-identification as White, excluding individuals marked as `in_white_British_ancestry_subset` (n = 30,116 who passed QC metrics 1-4 above). As was done for the White British in the initial UK Biobank study design [42], we identified global principal components of the genotype data, and then defined ancestry clusters using `aberrant` with the strictness parameter $\lambda = 20$. Non-British White individuals who were outliers for any of projected principal component pairs PC1/PC2, PC3/PC4, and PC5/PC6 were excluded (n = 25,137 remaining). We performed our first GWAS on the set of individuals in these White British and non-British White cohorts.

The combination of the two sources of European and White British ancestry individuals resulted in a total of 433,390 European ancestry individuals in UK Biobank, of whom 94,464 had available quality controlled Nightingale data. Our main goal for this study was to understand general principles of genetic architecture, which are not expected to vary among human populations, and thus in the main analysis we excluded non-European individuals on the basis of power and concerns about structure confounding. However, this analysis is significantly limited by the allele frequency differences between populations, and we sought to develop an alternative, inclusive strategy that did not rely on self-identity.

Metabolomics Data Generation

The metabolomics data was generated by Nightingale Health using a high-throughput NMR-based platform developed by Nightingale Health Ltd. Randomly selected EDTA non-fasting (average 4h since last meal) plasma samples (aliquot 3) from approximately 120,000 UK Biobank participants were measured in molar concentration units. No power calculation was performed, but we anticipated numerous discoveries on the basis of prior GWAS with similar or smaller sample sizes [22, 43, 24]. The measurements took place between June 2019 and April 2020 using six spectrometers at Nightingale Health, based in Finland. The Nightingale NMR biomarker profile contains 249 metabolic measures from each plasma sample in a single experimental assay, including 168 measures in absolute levels and 81 ratio measures. The biomarker coverage is based on feasibility for accurate quantification in a high-throughput manner and therefore mostly reflects molecules with high

concentration in circulation, rather than selected based on prior biological knowledge. Additional details about the data generation can be found at:

`https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/nmrm\_companion\_doc.pdf`.

Trait Selection and Grouping

Sixteen metabolites were chosen from the available Nightingale metabolites based on their biochemical proximity, relevance to health and disease, and because the genes and enzymes involved in their metabolism are well-characterized. Specifically, we first filtered to the 168 metabolites that were not metabolite ratio measurements (n=81) because we wanted to focus on absolute metabolites levels. We then filtered out the lipids and lipoprotein measures, including cholesterol and fatty acids, because the complexity of their biochemistry make it difficult to map out the chemical reactions directly interconverting one to another, and because many of these metabolites have already been extensively studied in large GWAS [43, 20]. However, many of these are important metabolites in the discussion of cardiometabolic disease so we additionally ran GWAS for total triglycerides, total fatty acids, HDL cholesterol, and LDL cholesterol as part of the interpretation of the *PCCB* variant using the same pipeline as for the sixteen metabolites below.

Finally we removed remaining derived measures (such as total combined concentration of BCAA) and those primarily reflecting physiological conditions such as fluid balance (creatinine and albumin) and GlycA (inflammation). One exception to this filtering was the three ketone bodies (3-hydroxybutyrate, Acetone and Acetoacetate) which were included due to their proximity and clear direct interconversions connecting them to the metabolic pathways of the remaining amino acid and glycolysis-related metabolites. Metabolites were classified into four biochemical groups based on biochemical similarity. The three branched chain amino acids: valine, leucine and isoleucine were classified as "BCAA", the remaining amino acids in the dataset: glycine, alanine, glutamine, tyrosine, phenylalanine and histidine were classified as "Other Amino Acid", the three ketone bodies: 3-hydroxybutyrate, acetone and acetoacetate were classified as "Ketone Body" and the four metabolites in or immediately adjacent to glycolysis: glucose, pyruvate, lactate and citrate were classified as "Glycolysis".

Trait QC and Covariate Adjustment

Trait measurements were filtered to only include baseline samples then log-transformed. Outlier removal was performed by dropping any sample that had a metabolite level greater than 20 fold the interquartile range or greater than 10 fold below the median across all samples for that metabolite. PCA was run for the remaining samples and outliers were dropped using aberrant (lambda = 20) on the top two PCs [44]. Remaining log transformed measurements were adjusted for spectrometer, week, and weekday. A total of 106,175 individuals had QC'd metabolomics data, ranging from 46-80 years old (mean 65.5; median 67), and of whom 54% were female, 90% were genotyped on the UK Biobank array (10% on BiLEVE), 94% identified as white (field 21000 code 1 or 1001-1003), 0.6% identified as multiracial (field 21000 code 2 or 2001-2004), 1.9% identified as Asian or Asian British (field 21000 code 3 or 3001-3004), 1.5% of whom identified as Black or Black British (field 21000 code 4 or 4001-4004), and 1.5% identified with a different label or declined to provide a label (field 21000 code -1, -3, 5, or 6). Samples were subset to the GWAS population defined above resulting in 94,464 individuals for the European ancestry GWAS.

GWAS

We performed GWAS in BOLT-LMM v2.3.2 [29] adjusting for sex, array, age, and genotype principal components 1-10 using the following command (data loading arguments removed for brevity):

```
bolt --phenoCol= [Metabolite] \
    --covarCol=sex \
    --covarCol=Array \
    --qCovarCol=age \
    --qCovarCol=PC{1:10} \
    --lmmForceNonInf \
    --numThreads=24 \
    --bgenMinMAF=1e-3 \
    --bgenMinINFO=0.3
```

The resulting GWAS summary statistics were then filtered to minor allele frequency greater than 0.01 and INFO score greater than 0.7 for further analyses (referred to as the Filtered Metabolite Sumstats). The LDSC munge_sumstats.py script was then use to munge the data (referred to as the Munged Metabolite Sumstats) [2].

GWAS Hit Processing

To evaluate GWAS hits, we took the Filtered Metabolite Sumstats and ran the following command using plink version 1.9 [45]:

```
plink --bfile [] --clump [GWAS input file] --clump-p1 1e-4 --clump-p2 1e-4
      --clump-r2 0.01 --clump-kb 1000 --clump-field P_BOLT_LMM --clump-snp-field SNP
```

We greedily merged GWAS hits across the 16 metabolites located within 0.1 cM of each other and took the SNP with the minimum p-value across all merged lead SNPs. In this way, we avoided potential overlapping variants that were driven by the same, extremely large, gene effects. This resulted in 213 lead GWAS variants, referred to as the metabolite GWAS hits.

Gene and Gene Type Annotation

We defined all genes in any GO [46, 47], KEGG [48], or Reactome MSigDB [49, 50] pathway as our full list of putative genes (in order to avoid pseudogenes and genes of unknown function). We initially extended genes by 100 kb (truncating at the chromosome ends) and used the corresponding regions, overlapped with SNP positions, to define SNPs within range of a given gene. Gene positions were defined based on Ensembl 87 gene annotations on the GRCh37 genome build. We then performed manual curation using GeneCards [51] to validate gene assignments and prioritize a single gene per SNP. Gene boundaries for genes encoding pathway-relevant enzymes in KEGG were extended up to 500 kb and assigned to a variant if the gene was biologically relevant to the metabolites the variant was significant in. If there were multiple genes within 100 kb of the variant, then gene assignments were made based on the following priority order: any genes encoding a pathway-relevant enzyme, genes encoding transporters, genes involved in translation/transcription regulation (referred to as TF for brevity), any genes whose function is known. If there were multiple genes of the same gene

type, then the assignment was made based on the relevance of the gene to the metabolites the variant was significant in, proximity of the gene to the variant, and, if applicable, any additional evidence in the literature (Oxford BIG [52] and Open Target Genetics [53, 54]). However, even for these cases where there was not high confidence in the exact gene assignment, for instance because there were multiple genes from the same gene family nearby, the top gene candidates all had the same gene type. Thus, because the major downstream analyses were designed in a way that only the gene type assigned to each variant mattered, the accuracy of the exact gene assignment should not affect the findings. If no genes with known function were within 100 kb of the variant then the window was extended up to 200 kb. The distance of a variant to a given gene was defined as the number of base pairs from the variant to the closer of the start or end of the gene boundaries or was set to 0 if the variant was within the gene boundaries.

We classified each gene using GeneCards [51] into one of five gene types: pathway-relevant enzyme, transporter, TF, general cell function and unknown. Genes encoding enzymes that catalyze a reaction in or adjacent to the direct synthesis or degradation of one of the 16 metabolites were defined as pathway-relevant enzymes using manual curation from GO, KEGG, REACTOME and Stanford's Human Metabolism Map [55], in addition to GeneCards. Genes encoding known transporters were classified as transporters. Genes involved in translation/transcription regulation were classified as TF. Genes whose function is known but not already classified as a pathway-relevant enzyme, transporter, or TF, were classified as "general cell function". Genes with unknown function or if there were no genes within 200 kb of the SNP were classified as "unknown". See 3 for each metabolite GWAS hit's gene and gene type annotations.

Gene type enrichments were calculated with a Poisson rate test. The baseline was the total of the GWAS hits among the 1.95 Gb of the genome within 100 kb of a gene in any pathway, and the test was performed with the number of GWAS hits within 100 kb of each pathway of interebst. There were 2.8 GWAS hits per megabase within 100 kb of a pathway-relevant enzyme versus 0.1 GWAS hits per megabase among all genes (25-fold enrichment, Poisson rate test $P < 2e\text{-}16$). There were 0.58 GWAS hits per megabase within 100 kb of a transporter versus 0.1 GWAS hits per megabase for all genes (5.2-fold enrichment, Poisson rate test $P = 9e\text{-}16$). We also repeated this analysis using closest genes rather than assigned genes, which allowed us to use a Fisher exact test (as each variant has a single closest gene). This resulted in a 27-fold ($P < 2e\text{-}16$) enrichment for pathway-relevant enzymes and an 11-fold ($P < 2e\text{-}15$) enrichment for transporters, respectively.

For TF enrichments, we used TF-Marker [56] to annotate tissue-specific marker gene TFs. We considered "TF" ($n = 1316$) and "TFMarker" ($n = 18$) genes as relevant genes, and TF Pmarker ($n = 1424$) genes as putatively relevant. We considered enrichment among the 628 genes not associated with cancer or stem cell biology (of which 267 are putative) as our set of tissue-specific TFs for downstream analysis. We consider our background in all cases to be our total GWAS hits number ($n = 213$) compared to the effective genome size (2.86 Gb). In specifically this TF set, we observed a 5.96-fold enrichment over the genome wide background (0.44 GWAS hits per megabase, $p = 4e\text{-}6$) among relevant gene bodies and 2.50-fold enrichment (0.19 hits/Mb, $P = 0.0007$) within 100 kb of a relevant gene, which was comparable for putatively relevant genes (4.85-fold and 2.86-fold, respectively). This was substantially higher than that of all genes in the genome (1.69-fold within gene bodies and 1.36-fold within 100kb of genes in any pathway) and comparable to that of all TFs regardless of their function in cancer or stem cells (4.82-fold within gene bodies and 2.18-fold within 100kb).

We next filtered tissue-specific TFs to those acting in Liver (28 relevant and 30 putative), Kidney (25 relevant and 20 putative), or Pancreas (14 relevant and 3 putative). Kidney and Pancreas TFs

had no more than 1 GWAS hit each and were excluded for these analyses. For Liver TFs, we observed a 18-fold enrichment (1.3 GWAS hits per megabase, P = 0.0057) within gene bodies and a 7.6-fold enrichment (0.56 hits/Mb, P = 0.002) within 100kb of genes. Results were similar when removing the cancer and stem cell filter (1.23 hits/Mb and 0.51 hits/Mb respectively) and dropped slightly when further including putatively relevant TFs (0.71 hits/Mb and 0.48 hits/Mb). Together, this suggests that liver marker TFs are specifically enriched for variants affecting our metabolite levels.

Ancestry-Inclusive GWAS

For the ancestry-inclusive analysis, we performed the same method as the European-ancestry only analysis except we omitted the step filtering individuals on the basis of self-identified race/ethnicity and ancestry PC outlier status. For the ancestry-inclusive analysis, we again used the European ancestry LD matrix as European-ancestry individuals were the overwhelming majority in the study. This resulted in a total of 98,189 individuals for the GWAS, which identified 238 lead GWAS variants across the 16 metabolites. This was inspired by recent "mega-analysis" studies [57]. We examined these associations and identified novel genes by comparing the list of pathway genes in the European-only analysis to those discovered in the ancestry-inclusive analysis.

Coloc based colocalization

Full summary statistics for all QCd SNPs within 1Mb ofthe target gene were considered at each locus, and these were loaded for all the evaluated traits. Standard deviations of the technical covariate adjusted trait measurements were used for input standard deviation calculations, and dichotomous traits were set to "cc" type datasets while continuous trait were set to "quant." `coloc.abf` was run using the default arguments [58, 59]. Output was aggregated and `PP.H4.abf` (the posterior probability of both traits having an association, and that this association is shared) was plotted for all pairs of traits run individually.

HESS Trait Heritability and Pathway Enrichments

We ran HESS [60] using the following commands:

```
hess.py --local-hsqg {\filtsumstats} --chrom {chrom} \
        --bfile 1kg_eur_1pct_chr{chrom} --partition EUR/fourier_ls-chr{chrom}.bed \
        --out {Metabolite}_step1
hess.py --prefix {Metabolite}_step1 --out {Metabolite}_step2
```

Where `1kg_eur_1pct_chr{chrom}` were downloaded from:

`https://ucla.box.com/shared/static/l8cjbl5jsnghhicn0gdej026x017aj9u.gz`

and `EUR/fourier_ls-chr{chrom}.bed` were downloaded from:

`https://bitbucket.org/nygcresearch/ldetect-data/src/ac125e47bf7f/?at=master`

We intersected the resulting heritability estimates per LD block with gene lists from each pathway (see Local ρ-HESS; within 100 kb of the gene boundary was used as the tested window) and calculated the total heritability within the pathway as the sum of the heritabilities across LD blocks and the variance of the heritability within the pathway as the sum of the variances within each

LD block. Overall, this gave a per-pathway estimate. We generated genome-wide estimates of heritability as well as heritability estimates for the subset of the genome nearby any coding gene in MSigDB as background controls from which to estimate the heritability enrichments, and used the coding gene numbers for reporting as they are more conservative.

LDSC Genetic Correlation

LD Score regression [2] was used to generate genetic correlation estimates. The following command was used:

```
ldsc.py --rg {\mungsumstats} --ref-ld-chr eur_ref_ld_chr
        --w-ld-chr eur_w_ld_chr
```

`eur_*_ld_chr` were downloaded from `https://data.broadinstitute.org/alkesgroup/LDSCORE/`.

Mendelian Randomization

The Rücker model selection framework was applied. Briefly, MR was run with inverse-variance weighted (IVW) and MR-Egger with fixed and random effects, and selection between different methods for results to present was based on the goodness-of-fit and heterogeneity parameters for the individual MR regressions as previously described [61, 62].

Discordant Variant Analysis

All pairwise combinations of LDSC Genetic Correlation (as described above) were performed for the 16 metabolites. Pairs were filtered to those that had a genetic correlation significantly different than 0 using ashR [63] with a local false sign rate of 0.005. We then annotated all metabolite GWAS hits with pairs of metabolites for which the variant had a P $<$ 1e-4 association with both metabolites and a P $<$ 5e-8 association with at least one, defined as significant metabolite pairs. A variant was classified as "discordant" if it had the same effect direction in both metabolites of at least one significant metabolite pair that had a negative global genetic correlation, or if it had opposite effect directions in the two metabolites of at least one significant metabolite pair that had a positive global genetic correlation. 14 variants of the 62 that had at least one significant metabolite pair were classified as discordant. Variants that had the same set of effect directions as the sign of the global LDSC genetic correlation for all of its significant metabolite pairs were classified as "concordant". Variants that had no significant metabolite pairs were classified as "neither".

The "between" region for a given pair of metabolites was defined as the shortest realistic biochemical path connecting them, and any alternative paths of reasonably similar distance and likelihood. This region can include a path converting one metabolite to the other, as well as other scenarios such as those involving colliders. This is because all biochemical reactions, including one-way reactions, can have bi-directional causal relationships due to Le Châtelier's Principle. In other words, even in a one-way (irreversible) reaction, changes in product levels can induce changes in reactant levels in order to re-establish equilibrium. Genes were defined as acting between a given metabolite pair either if they encoded an enzyme catalyzing a reaction in the "between" region defined above or if they encoded a transporter that primarily transports either of the two metabolites themselves or an intermediate metabolite in the "between" region. Variants were defined as acting between a given metabolite pair if the gene they affect was defined as between. Pathways were defined as

between a given metabolite pair if many of the genes defined as between the metabolites were part of the pathway or if many of the genes in the pathway were defined as between. Note that even if a pathway is defined as "between", not all genes in the pathway will always be between and vice versa; however, this is likely to only make the resulting differences in genetic correlation for "between" vs not "between" pathways more conservative.

Local ρ-HESS

We ran HESS [3] using the following commands:

```
hess.py --local-rhog {Met1_sumstats}
        {Met2_sumstats} --chrom {chrom} --bfile 1kg_eur_1pct_chr{chrom} \
        --partition EUR/fourier_ls-chr{chrom}.bed --out {Met1_Met2}_step1 \
hess.py --prefix {Met1_Met2}_step1_trait1 \
        --out {Met1_Met2}_step2_trait1
hess.py --prefix {Met1_Met2}_step1_trait2 \
        --out {Met1_Met2}_step2_trait2
hess.py --prefix {Met1_Met2}_step1 \
          --local-hsqg-est {Met1_Met2}_step2_trait1 {Met1_Met2}_step2_trait2 \
          --num-shared 94464 \
          --pheno-cor {gcov_int from LDSC genetic correlation for Met1_Met2} \
          --out {Met1_Met2}_step3
```

Where `1kg_eur_1pct_chr{chrom}` were downloaded from:

`https://ucla.box.com/shared/static/l8cjbl5jsnghhicn0gdej026x017aj9u.gz`

and `EUR/fourier_ls-chr{chrom}.bed` were downloaded from:

`https://bitbucket.org/nygcresearch/ldetect-data/src/ac125e47bf7f/?at=master`

We then used the local rho HESS results and estimated the local genetic covariance and correlation across all LD blocks overlapping pathway regions.

We defined the pathway regions based on gene boundaries of relevant genes in Supplementary Figure 3 - figure supplement 1 as follows: "Other Amino Acid Genes" includes all genes colored orange, "Ketone Body Genes" includes all genes colored purple, "Glycolysis, Gluconeogenesis and TCA Genes" includes all genes colored red, and "Urea Cycle Genes" includes all genes colored green. "BCAA Genes" included all genes in KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION except *OXCT2*, *HMGCL*, *HMGCS1*, *HMGCS2*, *ACAT1*, *ACAT2*, *OXCT1*, *DLD*, *AGXT2*, *ABAT*, and *AACS* and also included *ECHDC1*. "All Regions Outside Pathway Genes" was defined as all LD blocks not overlapping any of the regions defined above. "Metabolite Associated TF Genes" and "Metabolite Associated Transporters Genes" were defined as all LD blocks overlapping any of TFs or Transporters respectively annotating the Metabolite GWAS Hits.

Fligner-Killeen Variance Test

Rather than aggregating variant effects and estimating total genetic covariance and heritability per pathway, which is not robust to outlier effects, we additionally tried a non-parametric approach. Individual r_g and h^2 estimates for LD blocks were compared between the baseline (all coding genes) and the pathway of interest by listing all per-block genetic covariance scores and computing a

Fligner-Killeen Variance Test within each pathway in R. This enables direct evaluation of genetic covariances between the pathways, at the cost of simultaneously capturing the enrichment of heritability and genetic covariance therein.

BOLT-REML

Genotyped variants within 100 kb of genes in each pathway were aggregated, and the resulting matrices were tested using the following command in BOLT-LMM:

```
bolt
        --remove {non-European ancestry individuals}
        --phenoFile={Technical-adjusted metabolites} \
        --phenoCol=Ala \
        --phenoCol=Gln \
        --covarCol=sex --covarCol=Array --qCovarCol=age --qCovarCol=PC{1:10} \
        --geneticMapFile=genetic_map_hg19_withX.txt.gz  ‘# downloaded with bolt‘ \
        --numThreads=24 --verboseStats \
        --modelSnps {pathway SNPs} \
        --reml \
        --noMapCheck

```

Standard errors were as reported by BOLT-REML.

Haseman-Elston Regression

Genotyped variants were pruned to MAF $> 1\%$ and approximate linkage equilibrium among individuals included in the GWAS using `--indep 50 5 0.5`. The resulting variants were used to construct genetic relatedness matrices (GRMs) that included the genotyped SNPs within 100 kb of genes in each pathway, and the resulting matrices were tested using the following command in GCTA:

```
{GCTA} --HEreg-bivar {trait1} {trait2} --thread-num 16 --grm {GRM}
```

Results using multiple GRMs (`--mgrm`) to jointly test all pathways were qualitatively similar outside of the genome-wide GRM, which no longer captured the within-pathway component.

Stratified LD Score Regression

Analyses were performed as described in LDSC Genetic Correlation, except that rather than `eur_ref_ld_chr` as the reference LD Scores, instead LD Scores computed on variants within 100 kb of genes in each pathway were utilized.

Disease Variant Analysis

The metabolite GWAS hits annotated with pathway-relevant enzymes were overlapped with significant hits for CAD, identifying the variant rs61791721 as the most significant variant [32, 33].

Incident coronary artery disease cases were defined among UK Biobank participants as those individuals who received a first diagnosis of myocardial infarction (MI) using the analytical MI model (field 42000) after the date of baseline assessment. Prevalent cases (individuals with a first diagnosis before date of assessment) were excluded. A Cox proportional hazard model was run with the technical-covariate-adjusted, log-transformed metabolite levels predicting incident MI status, adjusted for age, age^2, age * sex, age^2 * sex, and statin usage (defined based on a list of individual drug codes as previously described [7]). Effect sizes presented are based on the estimates from these models run independently for each metabolite.

Colocalization analysis

We wanted to evaluate the extent to which our associations might represent single causal variants across multiple traits and used conditional association at the locus to evaluate this. For each variant within 500 kb of our lead SNPs in at least one metabolite, we ran a conditional analysis for the variants within 1 Mb of the gene body of our putative target gene. Then we ran the following association test in plink2:

```
plink2 --glm cols=chrom,pos,ref,alt,a1freq,firth,test,
nobs,orbeta,se,ci,tz,p hide-covar omit-ref
     --pfile <imputed genotypes>
     --covar <age/sex/PCs>
     --keep <94464 European-ancestry individuals in the BOLT-LMM GWAS>
     --out conditional/$gene/$snp
     --pheno <technical-residualized traits>
     --extract <(variants within 1Mb of gene body)
     --condition <conditional SNP>
```

For single SNP conditioning tests and `--condition-list` for conditioning on multiple variants. Associations were visually inspected to detect highly linked variants and conditioning tests were repeated with top associations in any of the key traits until there were no significant variants remaining.

For the *PCCB* vignette, additional traits were included in the analysis, including fatty acids and lipids in the Nightingale-assayed individuals and clinical biomarkers in the full cohort of European-ancestry UK Biobank participants, where traits were residualized as previously described [7]. We further included a GWAS for "hard" CAD as previously defined [64], for which results were qualitatively similar when evaluating "soft" CAD (including angina cases) and employing only EHR-based diagnoses (rather than additionally including self reported case status). Results for "hard" CAD are shown in the supplement.

Pathway diagrams

Diagrams were drawn using Affinity Design, and molecular structures were made using ChemDraw. Pathway information was curated from GO [46, 47], KEGG [48], or Reactome MSigDB [49, 50], and Stanford's Human Metabolism Map [55], along with manual curation from public domain biochemistry knowledge (Supplementary File 2).

Acknowledgements

We thank Alyssa Lyn Fortier, Hanna M. Ollila, Shoa Clarke and other members of the Pritchard lab and Assimes lab for helpful discussions. The authors are grateful to UK Biobank and its participants for access to data to undertake this study (Project #30418 and #24983). Nightingale Health Plc is acknowledged for early access to the UK Biobank NMR biomarker data. C.J.S. is supported by a National Science Foundation Graduate Research Fellowship and Stanford's Knight-Hennessy Scholars Program. This work was supported by NIH grants 5R01HG011432 and 5R01AG066490 (to J.K.P.). We would like to thank the reviewers for their careful reading of our manuscript and their thoughtful comments and suggestions that improved our manuscript.

Competing interests

Anna Cichońska is a former employee and holds stock options with Nightingale Health Plc. Heli Julkunen is an employee and holds stock options with Nightingale Health Plc. Eric Fauman is affiliated with Pfizer Worldwide Research, has no financial interests to declare, contributed as an individual and the work was not part of a Pfizer collaboration nor was it funded by Pfizer. Peter Würtz is an employee and shareholder of Nightingale Health Plc. The other author declares that no competing interests exist.

Data availability

The source data and analyzed data have been deposited in Dryad under the accession code `https://datadryad.org/stash/share/gs99jEeFTcFu4ATHddLKXDPyVtMsV_PQZHQ1dGHCMRo`. Data analysis and figure generation scripts are available on Github at: `https://github.com/courtrun/Pleiotropy-of-UKB-Metabolites`.

Supplementary Figures and Tables

Supplementary Figures

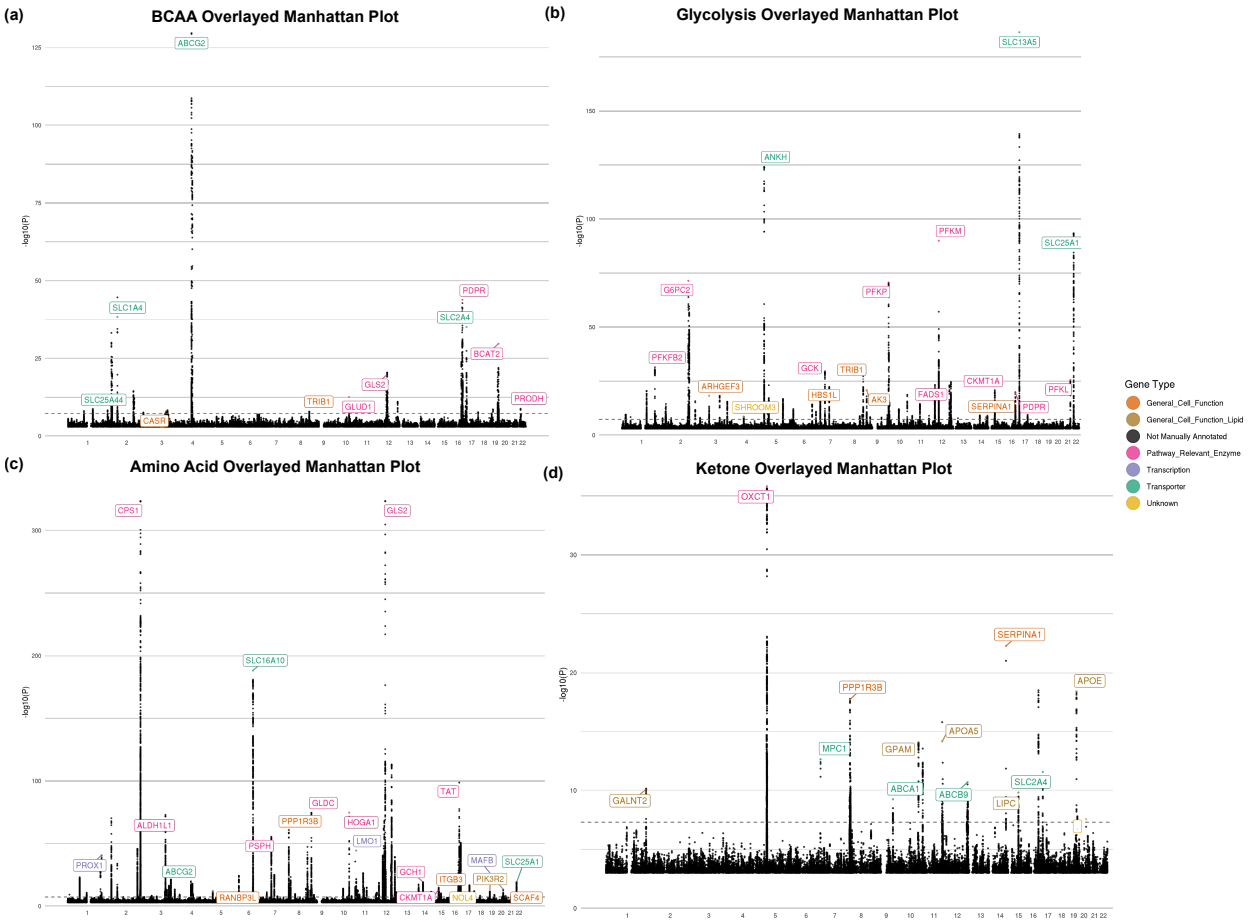

Figure 2 - figure supplement 1: Manhattan plots. *Manhattan plots for each metabolite overlayed by biochemical group. Coloring reflects gene type assignment for where relevant.*

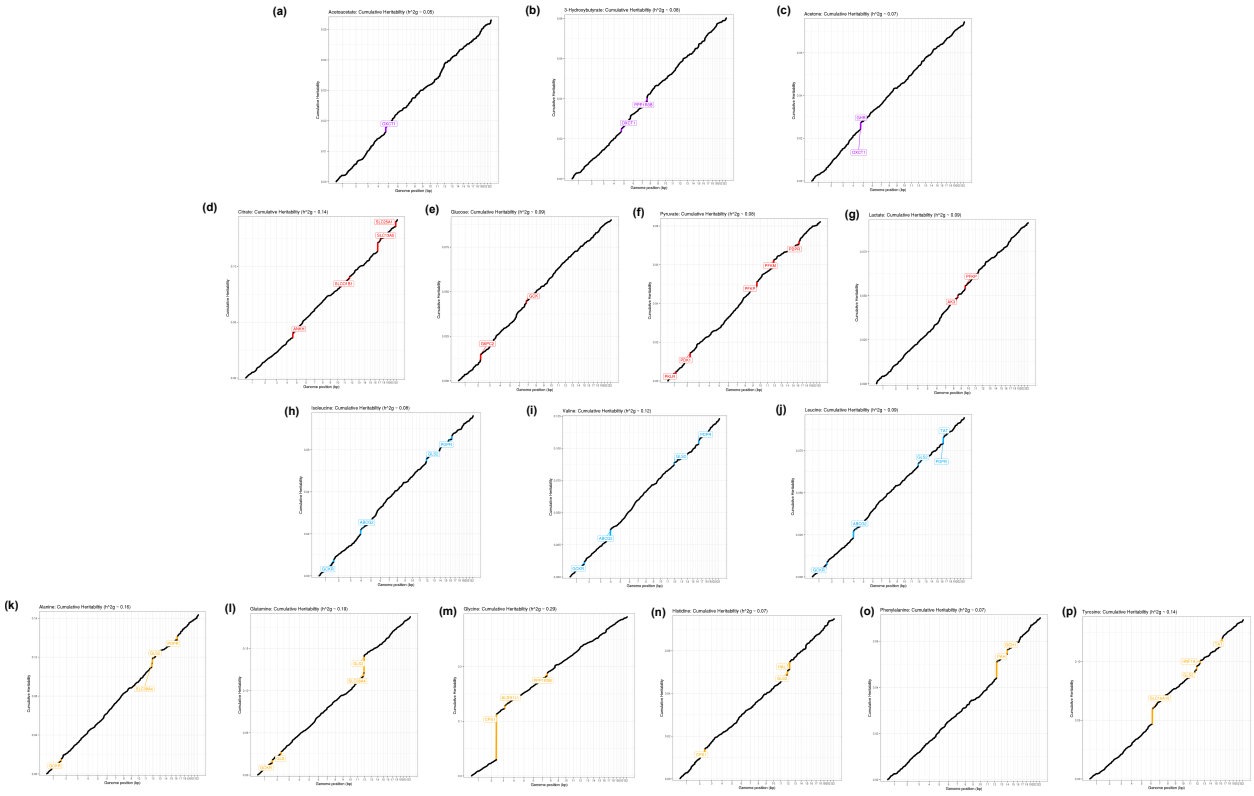

Figure 2 - figure supplement 2: HESS heritability plots. *Cumulative HESS heritability plots for the 16 metabolites, colored by biochemical group.*

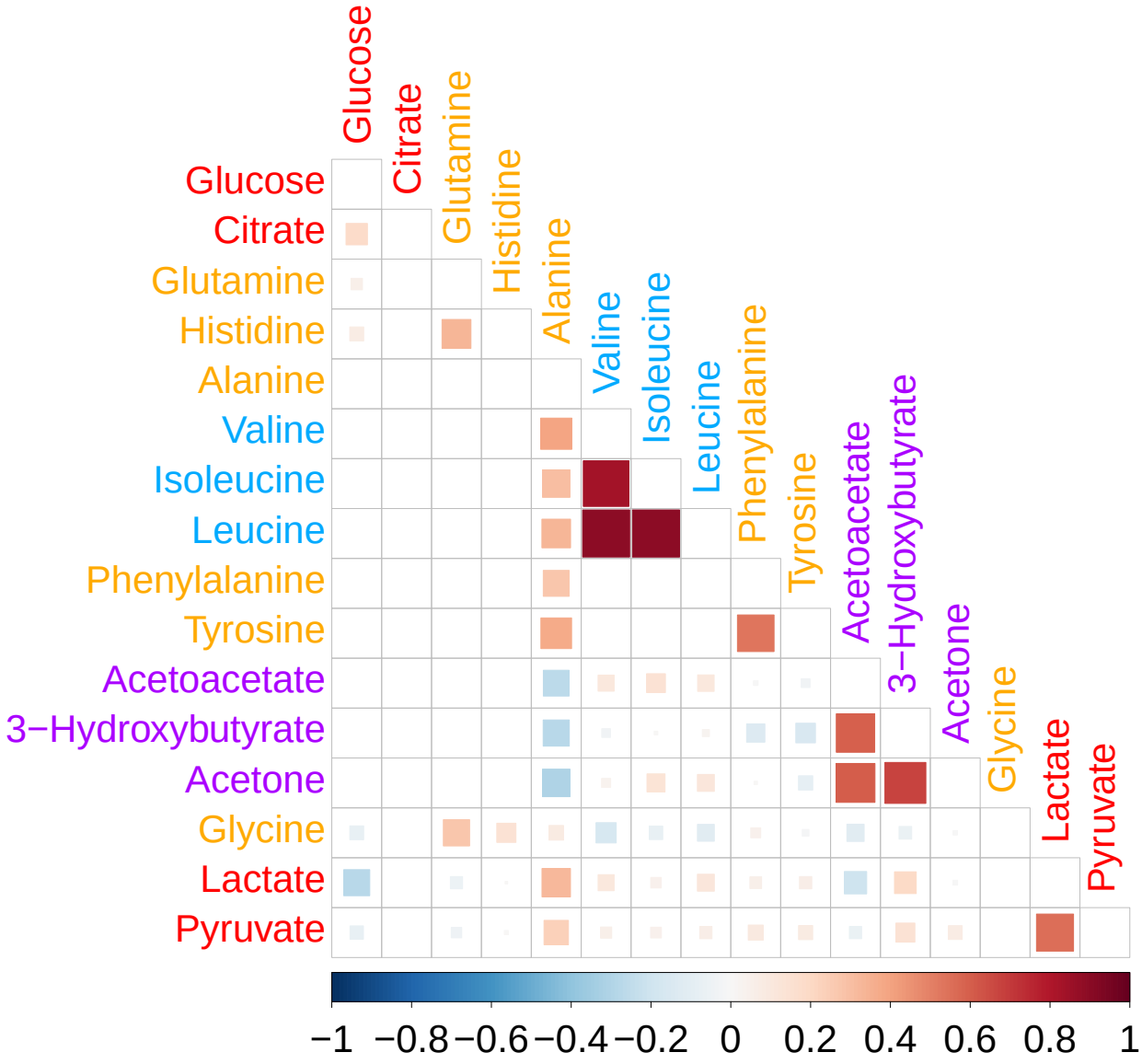

Figure 2 - figure supplement 3: Phenotypic correlation. *Correlation matrix of the residualized phenotype levels for the 16 metabolites. Metabolites are colored by biochemical group.*

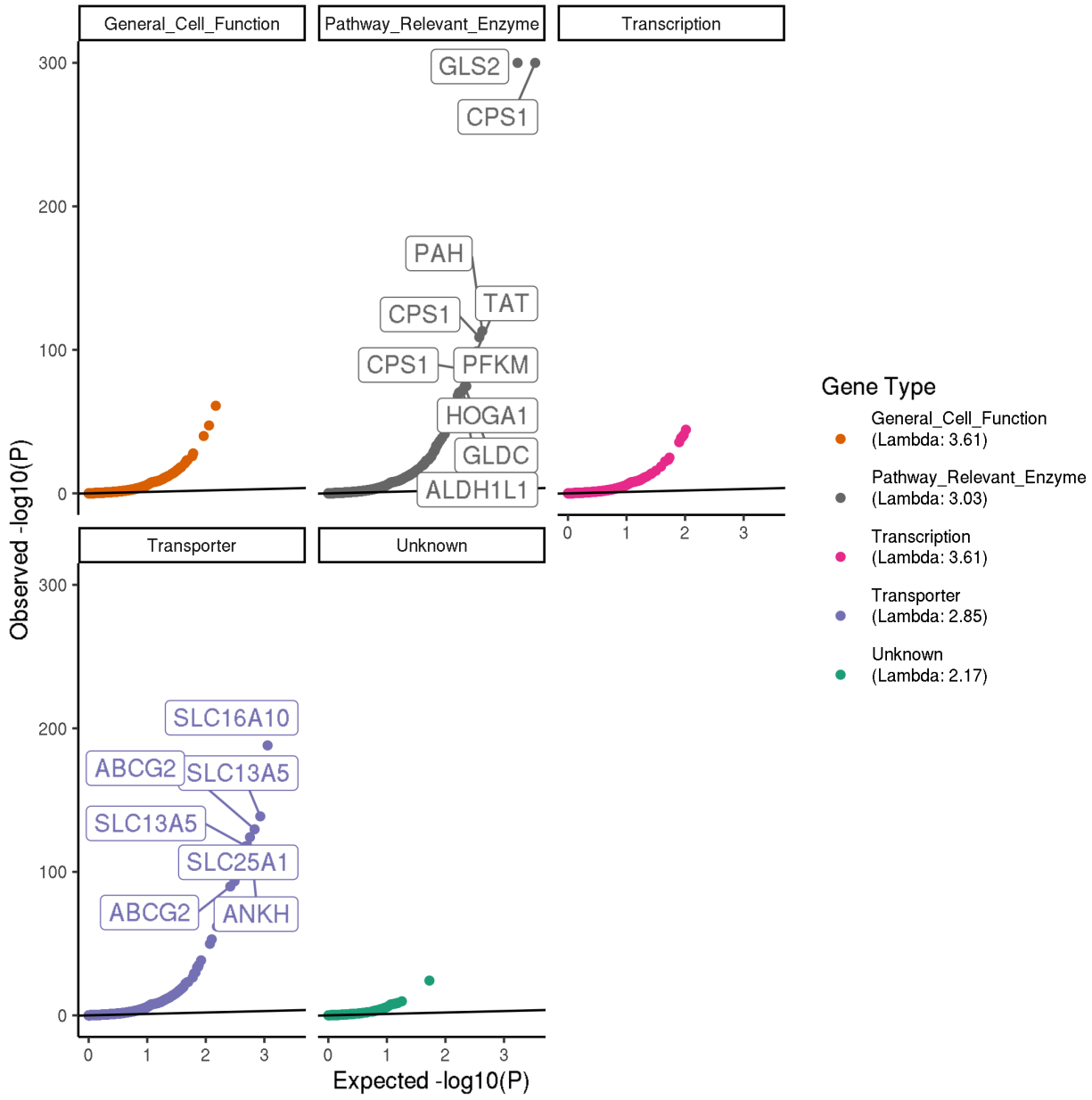

Figure 2 - figure supplement 4: Quantile-quantile plots. *Quantile-quantile plots showing the observed -log10(P-value) from the 16 GWAS in each metabolite for the 213 Metabolite GWAS Hits. The most significant trait for each SNP is excluded, and the plots are faceted by gene type. Note that for plotting purposes, P-values were set to a minimum of 1e-300. Variants with quantile < 0.005 were labeled with their gene annotation.*

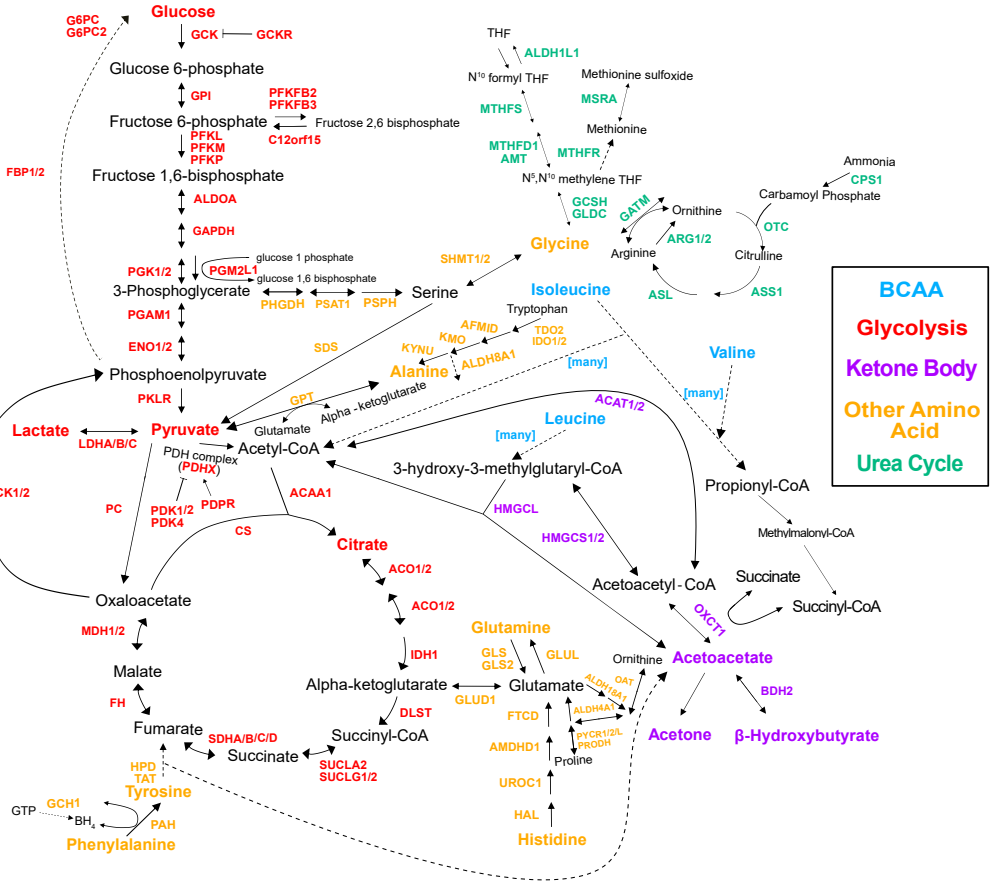

Figure 3 - figure supplement 1: Genes in the pathway. *Pathway diagram showing the genes in the pathway, regardless of whether there was a variant annotated with them or not.*

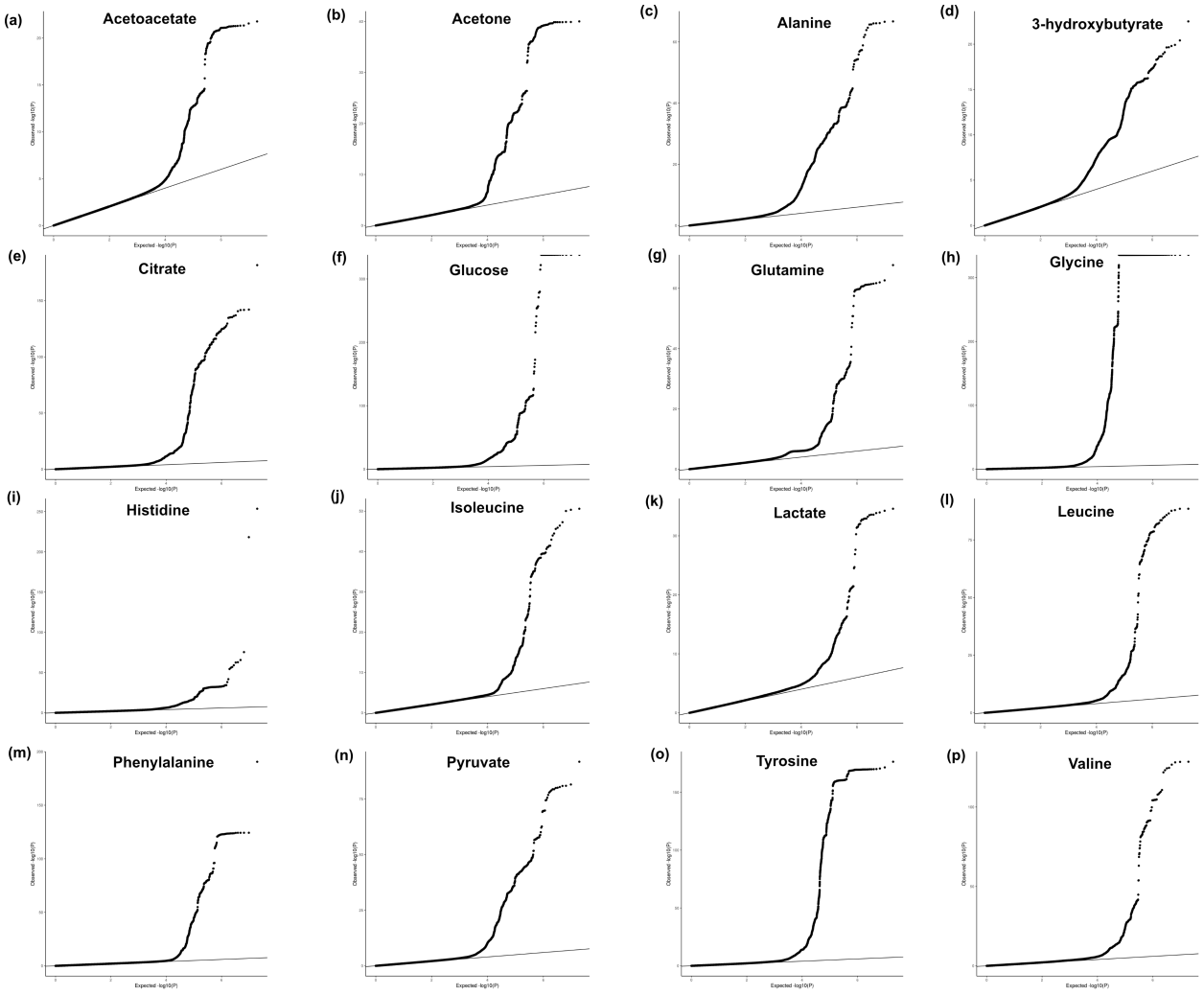

Figure 3 - figure supplement 2: Quantile-quantile plots. *Quantile-quantile plots showing the observed -log10(P-value) from the 16 GWAS in each metabolite for the full summary statistics of the ancestry-inclusive GWAS. Note that for plotting purposes, P-values were set to a minimum of 1e-324.*

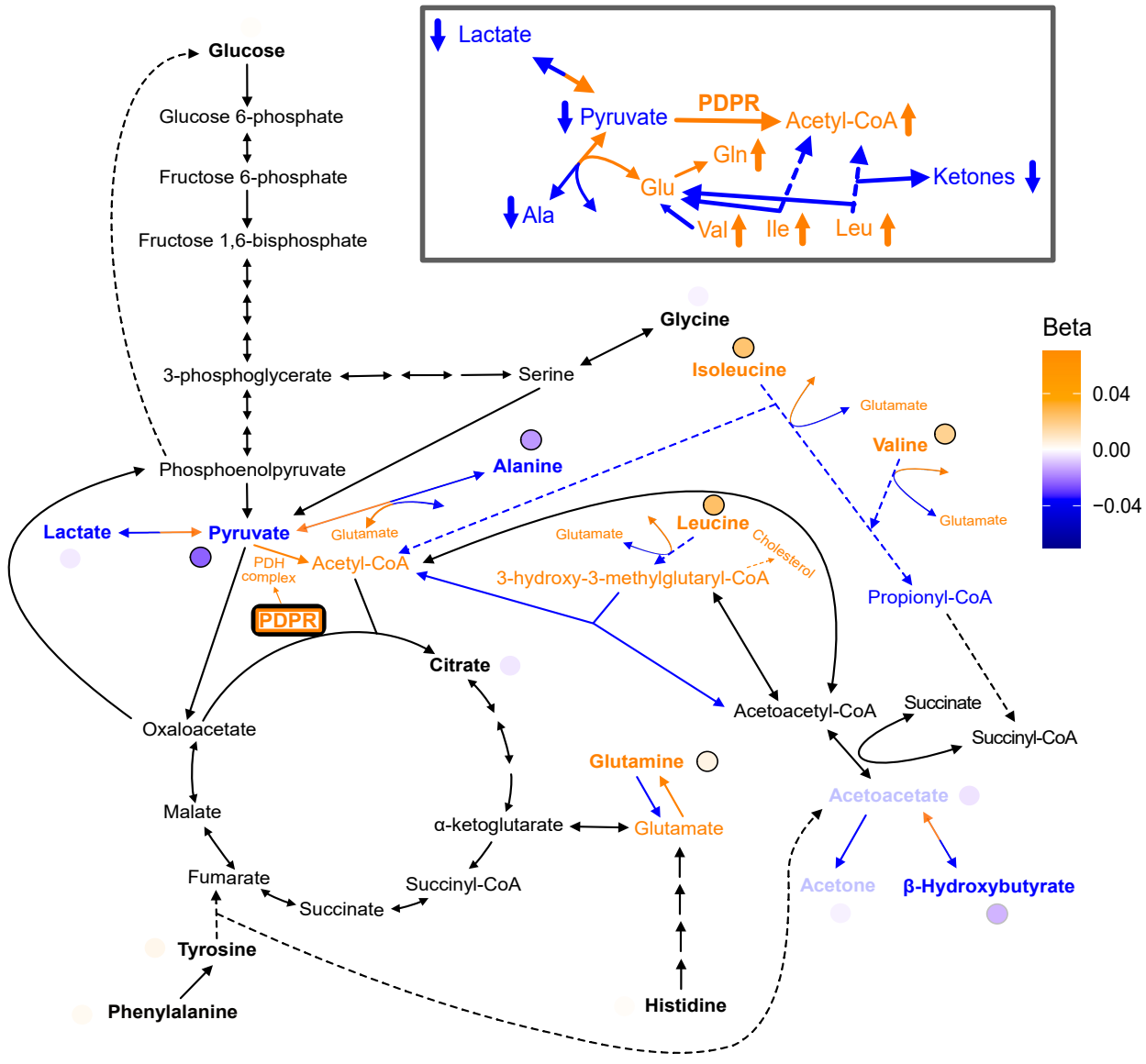

Figure 4 - figure supplement 1: Full Pathway Results for PDPR. *Full pathway results and possible mechanism for discordant variant rs3700141711 with gene annotation PDPR.*

rs370014171 (PDPR): The discordant variant rs370014171 had gene annotation *PDPR*, which encodes the regulatory unit of the protein pyruvate dehydrogenase phosphatase which is responsible for the activation of the pyruvate dehydrogenase (PDH) complex that catalyzes the conversion of pyruvate to acetyl-CoA (Supplementary Expanded Pathway Figure 4 - figure supplement 1). One possible mechanism for this variant is that it is increasing PDPR activity leading to increased conversion of pyruvate to acetyl-CoA, resulting in decreased pyruvate levels and, by compensation, decreased lactate levels. In addition, to compensate for the decreased pyruvate levels, there could be an increased conversion of alanine to pyruvate and glutamate. This would cause a decrease in levels of alanine and an increase in levels of glutamate. In response to the increased acetyl-CoA, there could be decreased breakdown of metabolites that are normally broken down for its production, including isoleucine and leucine, resulting in an increase in their levels. With the increase in acetyl-CoA levels, less HMG-CoA is broken down to acetyl-CoA. The reaction of HMG-CoA to acetyl-CoA also produces acetoacetate, so this decrease results in a decrease in acetoacetate, as well as the other ketone bodies (acetone and β-hydroxybutyrate), which are directly downstream of acetoacetate. Some of the increase in HMG-CoA (from its decreased breakdown) leads to an increase in cholesterol. Meanwhile, alanine to glutamate is an important regulator of BCAA levels since the first step in BCAA breakdown is a reversible conversion to glutamate. An increase in glutamate's levels means again less BCAA will need to be broken down, which is another potential reason for increased levels of the BCAAs (isoleucine, leucine, and valine). This variant was also found to be significantly associated with an increase in isoleucine, leucine, and valine and a decrease in alanine in another recent metabolomics study [24]. In this dataset, after these four amino acids, the metabolite with the next most significant association with this variant lysine (Z = 3.8, P = 1e-4) [24]. Lysine is another amino acid that can be catabolized to produce acetyl-CoA and thus like isoleucine, leucine and valine, based on the hypothesized mechanism for this variant it makes sense lysine would have a positive association.

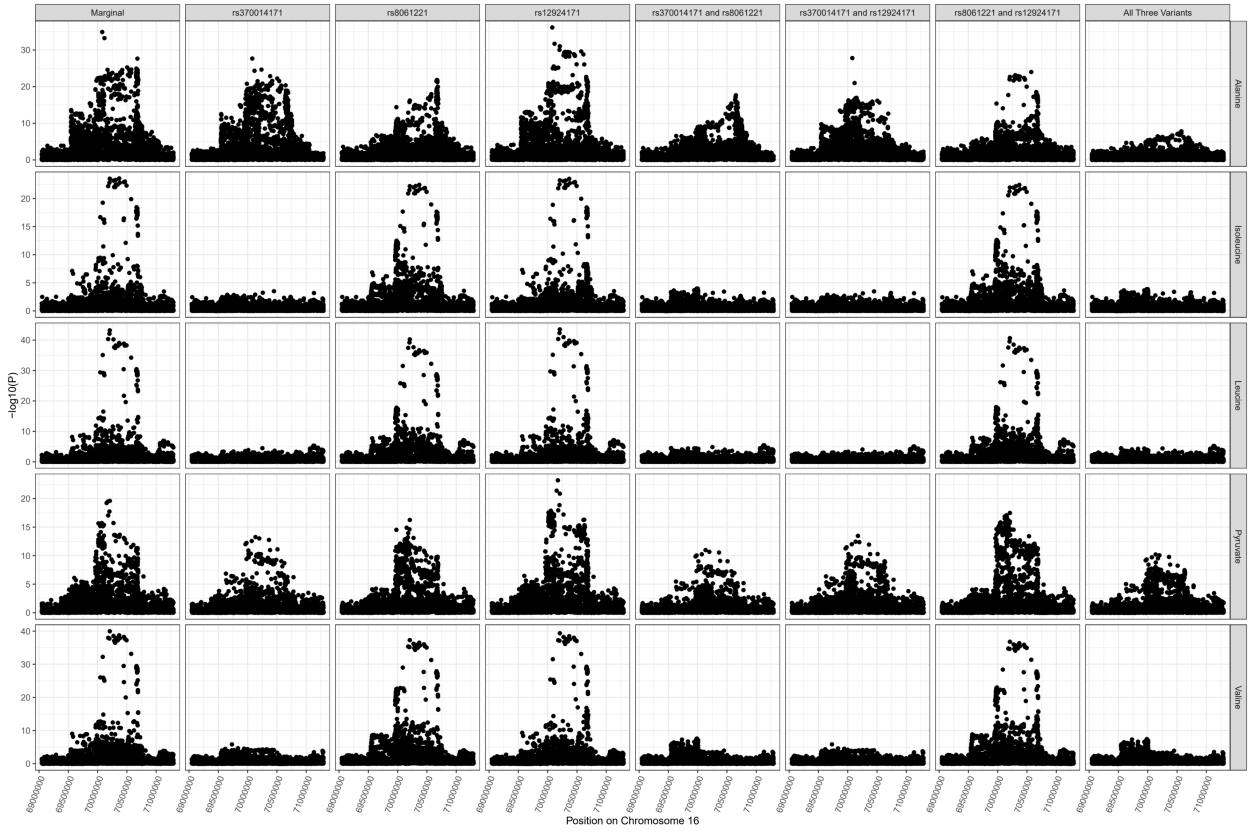

Figure 4 - figure supplement 2: PDPR locus colocalization. *Colocalization results for PDPR locus demonstrating pleiotropic effects of variant rs370014171 on alanine (one of multiple independent associations at this locus), pyruvate, isoleucine, valine, and leucine.*

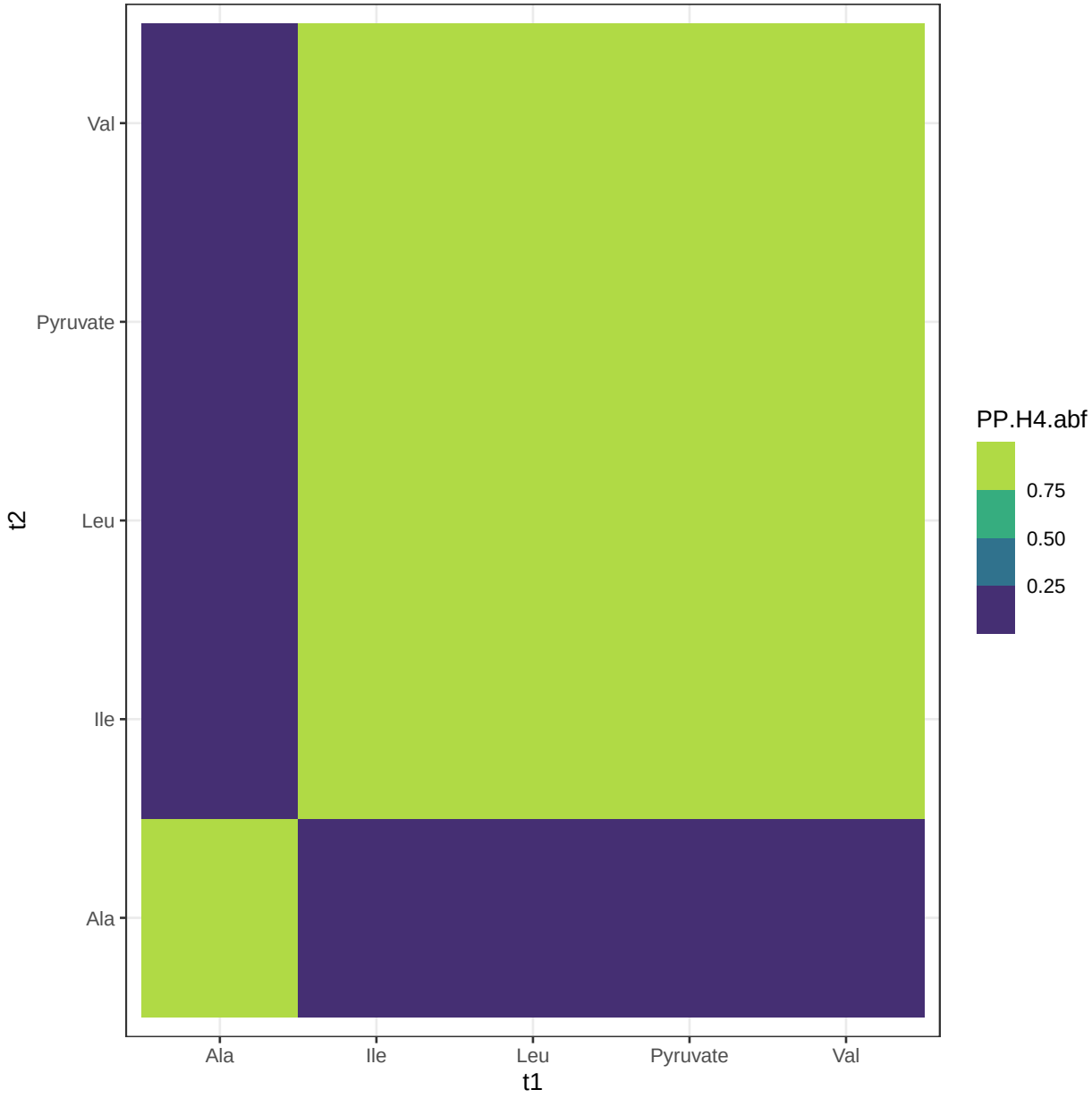

Figure 4 - figure supplement 3: Colocalization at PDPR. *Colocalization probabilities at PDPR for the joint association signals of all metabolites. Note that alanine has multiple associations at the locus (Figure 4 - figure supplement 2) and thus has reduced colocalization, since coloc assumes a single causal variant; see Figure 4 - figure supplement 4 for conditional colocalization.*

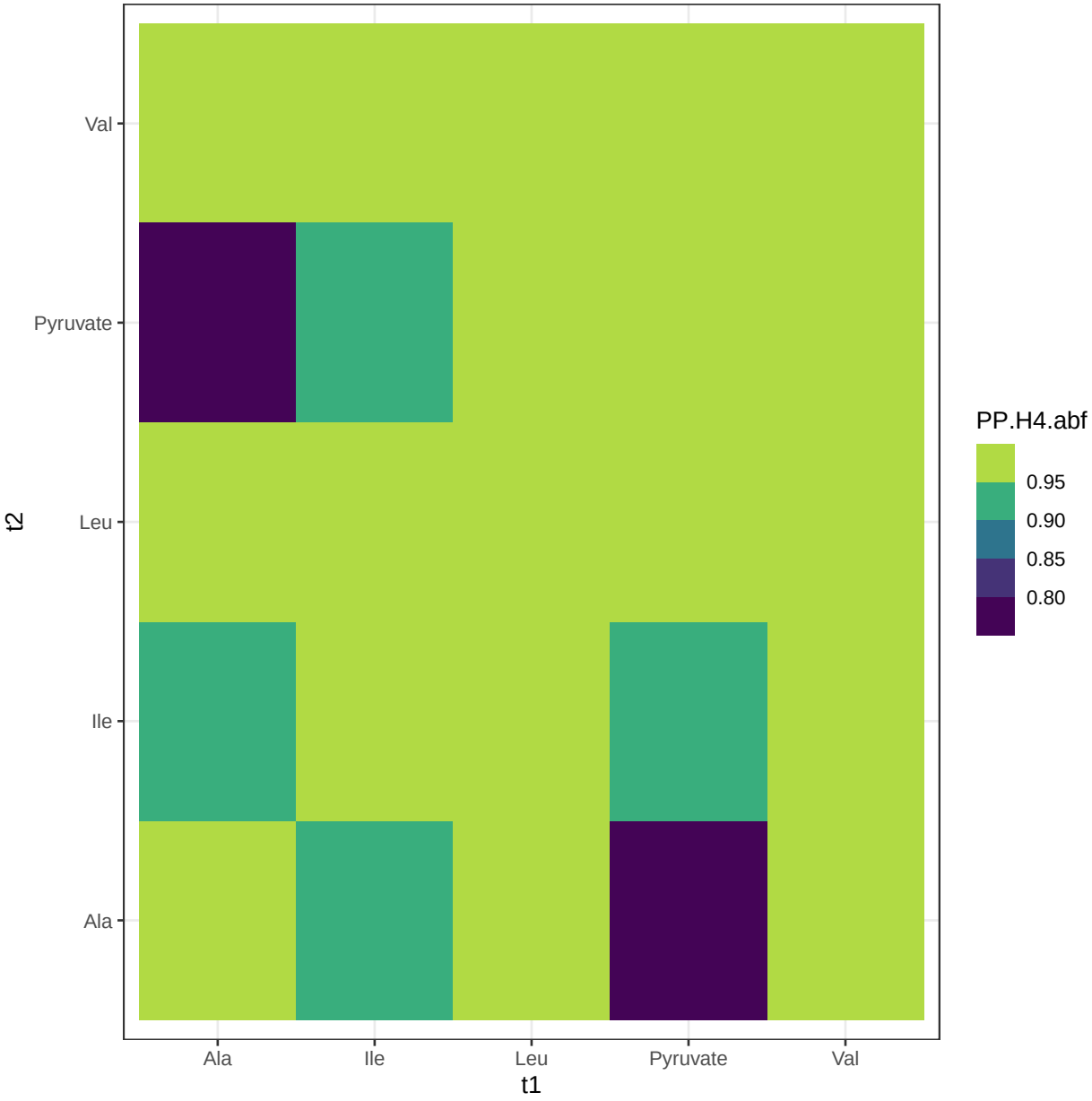

Figure 4 - figure supplement 4: Colocalization at PDPR conditioned on secondary signals. *Colocalization probabilities at PDPR for the joint association signals of all metabolites for summary statistics conditioned on rs8061221 and rs12924171. By conditioning on these two secondary SNPs, there is broad colocalization across traits.*

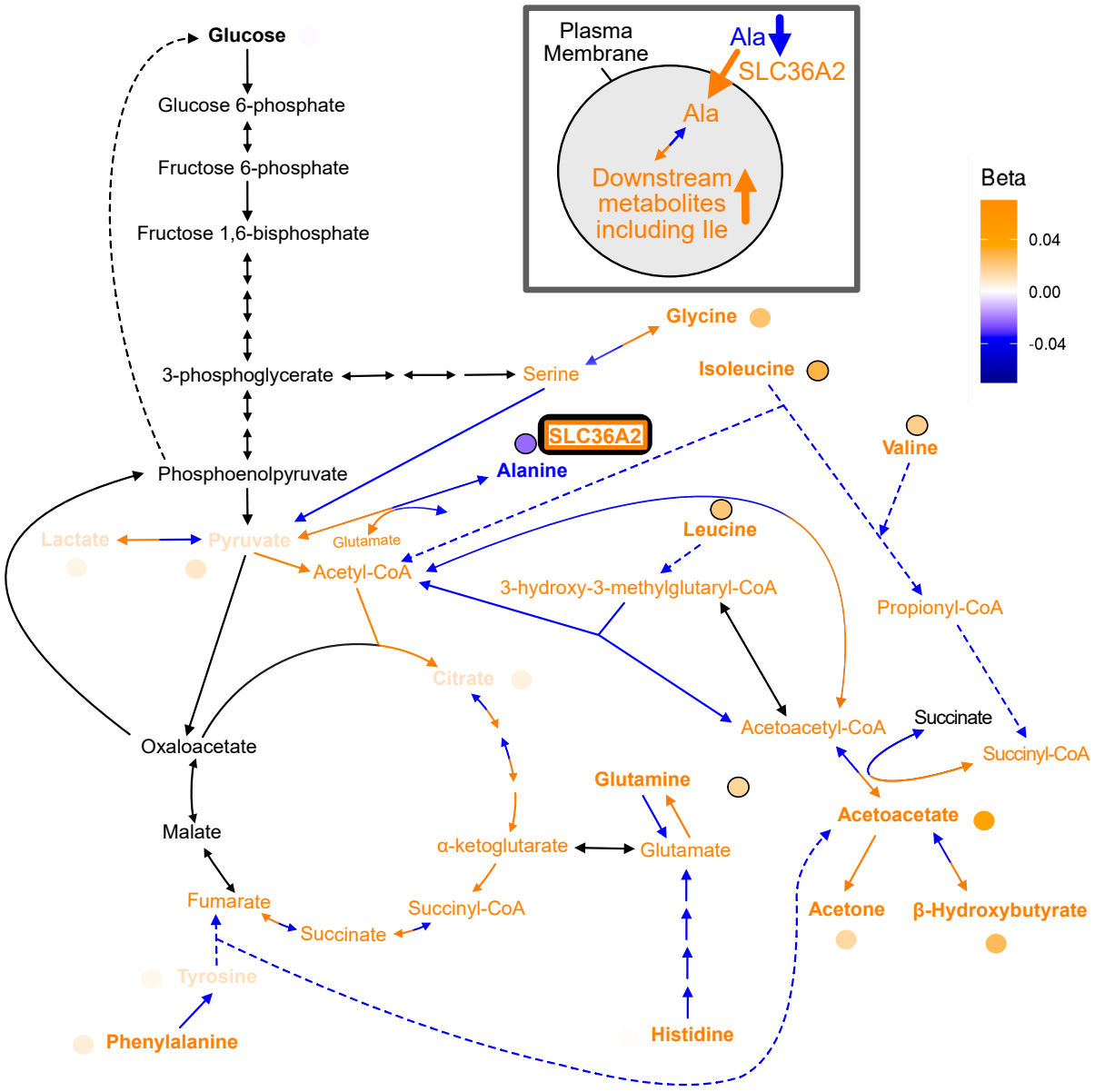

Figure 4 - figure supplement 5: Full Pathway Results for SLC36A2. *Full pathway results and possible mechanism for discordant variant rs77010315 with gene annotation SLC36A2.*

rs77010315 (SLC36A2): The discordant variant rs77010315 is a missense variant in *SLC36A2*, which encodes a transporter for small amino acids such as alanine (Supplementary Expanded Pathway Figure 4 - figure supplement 5). We hypothesize this variant is discordant because it increases the activity of SLC36A2, leading to increased transport of alanine into cells. This results in a decrease in levels of alanine in the blood, but results in increased intracellular conversion of alanine to pyruvate and glutamate. Pyruvate can then be reversibly converted to lactate, resulting in an increase in lactate levels, as well as converted to acetyl-CoA. The increased pyruvate results in an increase in glycine levels because less needs to be broken down to produce it. The increase in acetyl-CoA leads to an increase in its downstream citric acid cycle intermediates including citrate, fumarate, succinyl-CoA and alpha-ketoglutarate. The increased glutamate leads to an increase in glutamine and the increase in fumarate leads to an increase in tyrosine and phenylalanine. The increase in acetyl-CoA also leads to an increase in isoleucine, HMG-CoA and leucine since less needs to be broken down to produce it, and to increased acetoacetyl-CoA since more acetyl-CoA is available to be converted to it. Acetoacetyl-CoA can then be converted to acetoacetate leading to an increase in all three ketone bodies (acetoacetate, acetone and β-hydroxybutyrate) as well as succinyl-CoA. Finally, the increased succinyl-CoA leads to an increase in valine because less valine needs to be broken down to produce it. This variant also had consistent associations consistent in direction of effect with alanine (Z = -4.3, P = 2e-5), glutamine (Z = 3.9, P = 1e-4), valine (Z = 3.7, P = 2e-4), and glycine (Z = 3.4, P = 8e-4) in another dataset [24]. In this dataset, the most significant association with this variant was carnitine (Z = 13.2, P = 2e-39) [24]. Carnitine can be made from methionine which is a metabolite that can be produced from glycine breakdown. Because increased levels of glycine from this variant may result in increased levels of carnitine, based on the hypothesized mechanism for this variant it may make sense that carnitine would have a positive association.

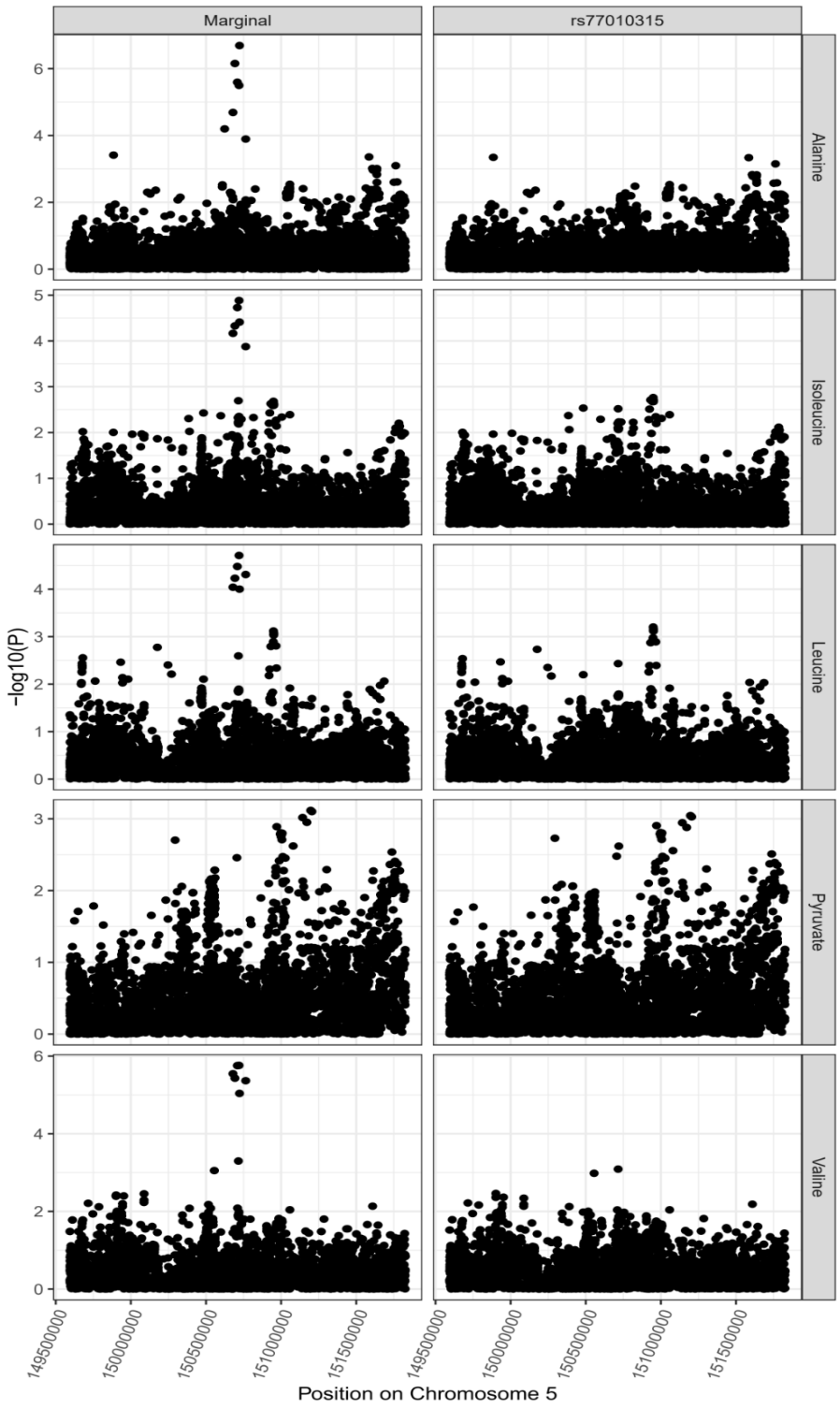

Figure 4 - figure supplement 6: SLC36A2 locus colocalization. *Colocalization results for SLC36A2 locus demonstrating pleiotropic effects of variant rs77010315 on alanine, isoleucine, valine and leucine. There is limited evidence for an association with pyruvate levels at the locus, suggesting the potential for larger sample sizes to resolve the contributions of changes to pyruvate levels at SLC36A2.*

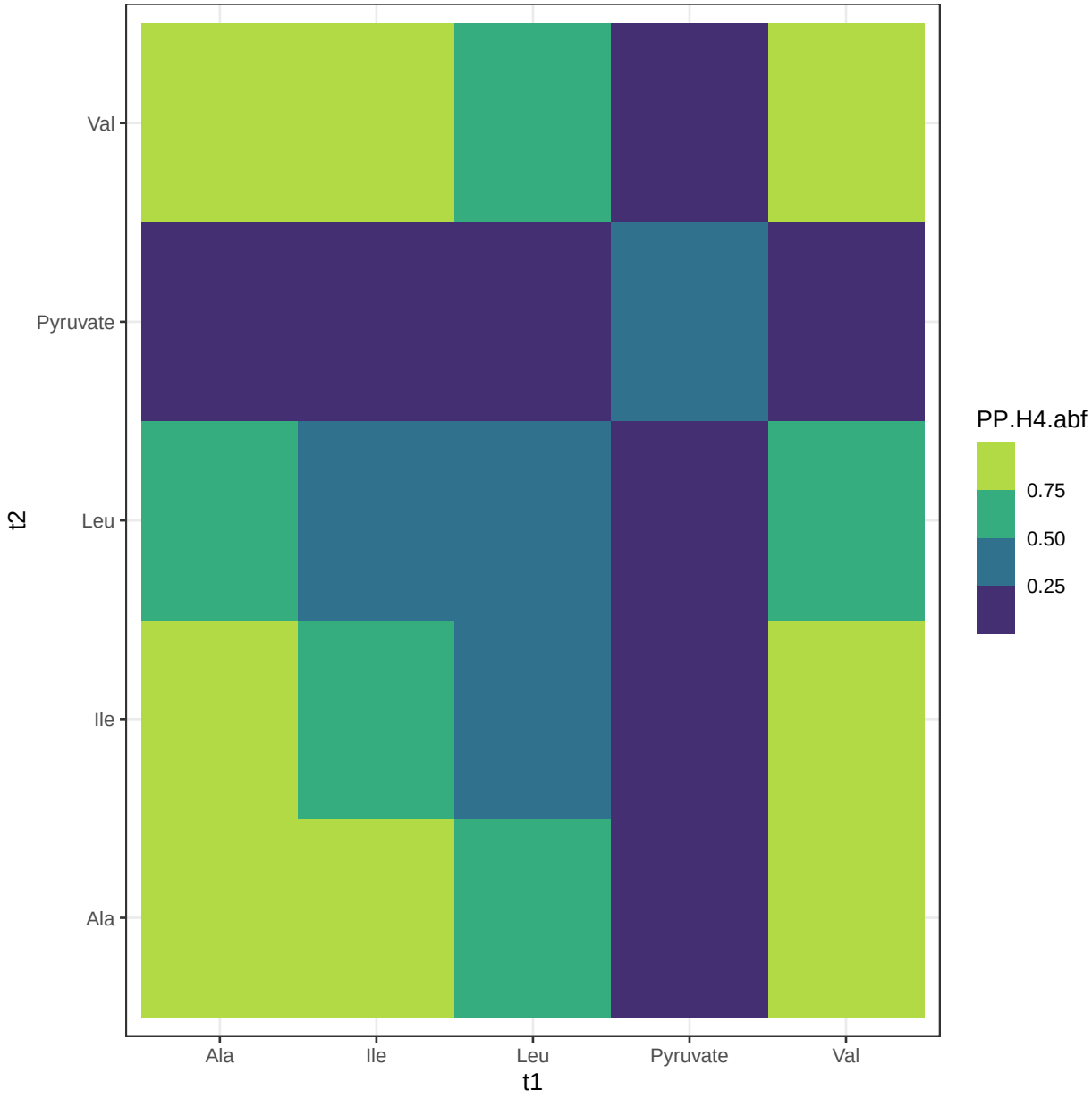

Figure 4 - figure supplement 7: Colocalization at SLC36A2. *Colocalization probabilities at SLC36A2 for the joint association signals of all metabolites. Note that pyruvate has a weak or non-existent marginal association at the locus, which likely contributes to the lack of colocalization; see Figure 4 - figure supplement 6.*

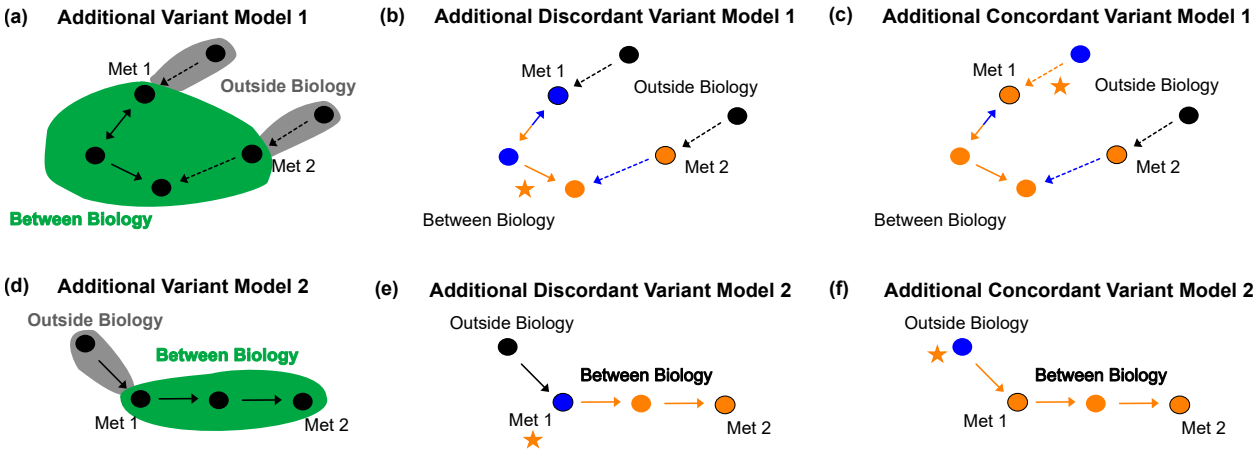

Figure 5 - figure supplement 1: Additional Variant Models. *Additional Variant Model 1 defining of between vs outside biology* **a.**, *example discordant variant,* **b.**, *and example concordant variant* **c.** *for a more complicated example with a pathway like that in Figure 4c and 4d. Additional Variant Model 2 defining of between vs outside biology* **d.**, *example discordant variant,* **e.**, *and example concordant variant* **f.** *for a variant affecting a transmembrane transporter, labeled next to the metabolite the transporter acts on.*

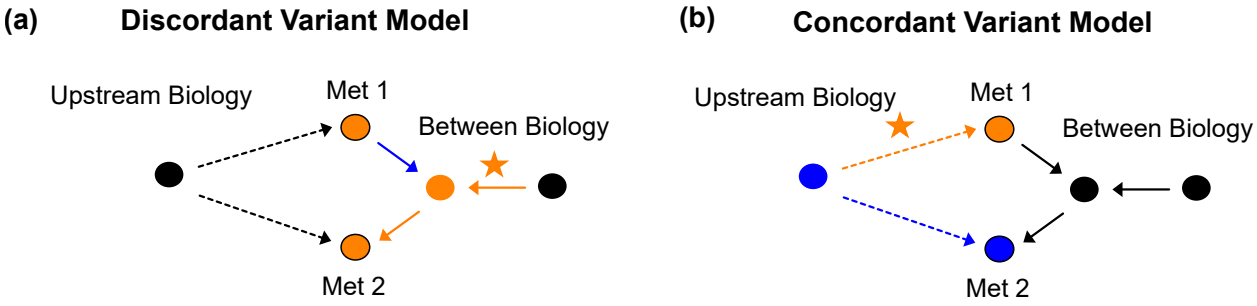

Figure 5 - figure supplement 2: Negative Genetic Correlation Variant Models. *Model for an example discordant* **a.** *and an example concordant* **b.** *variant when there is an overall negative genetic correlation.*

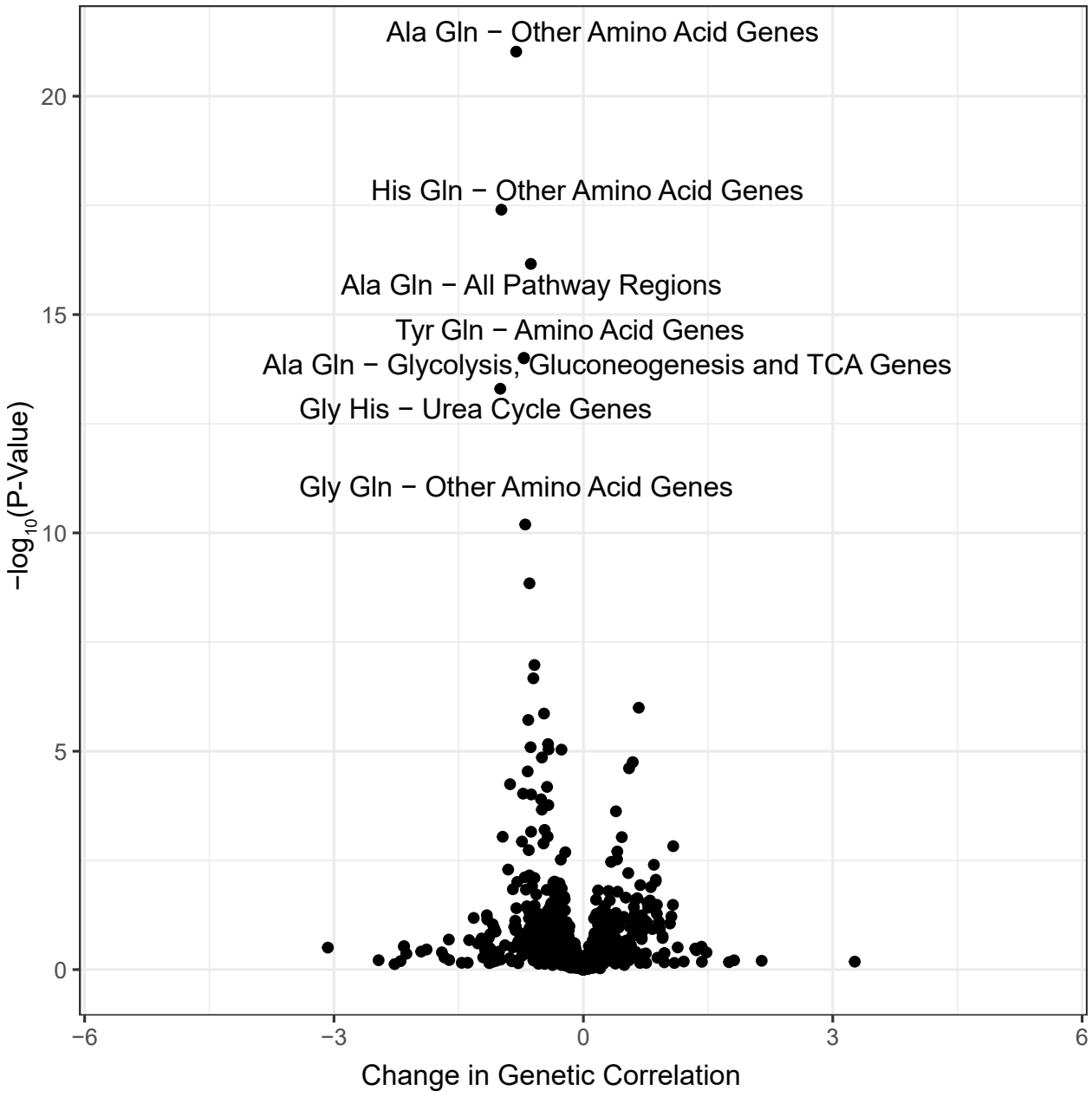

Figure 6 - figure supplement 1: Genetic correlation results. *P-values for all metabolite pairs for all pathway regions for Haseman-Elston regression genetic correlation analysis.*

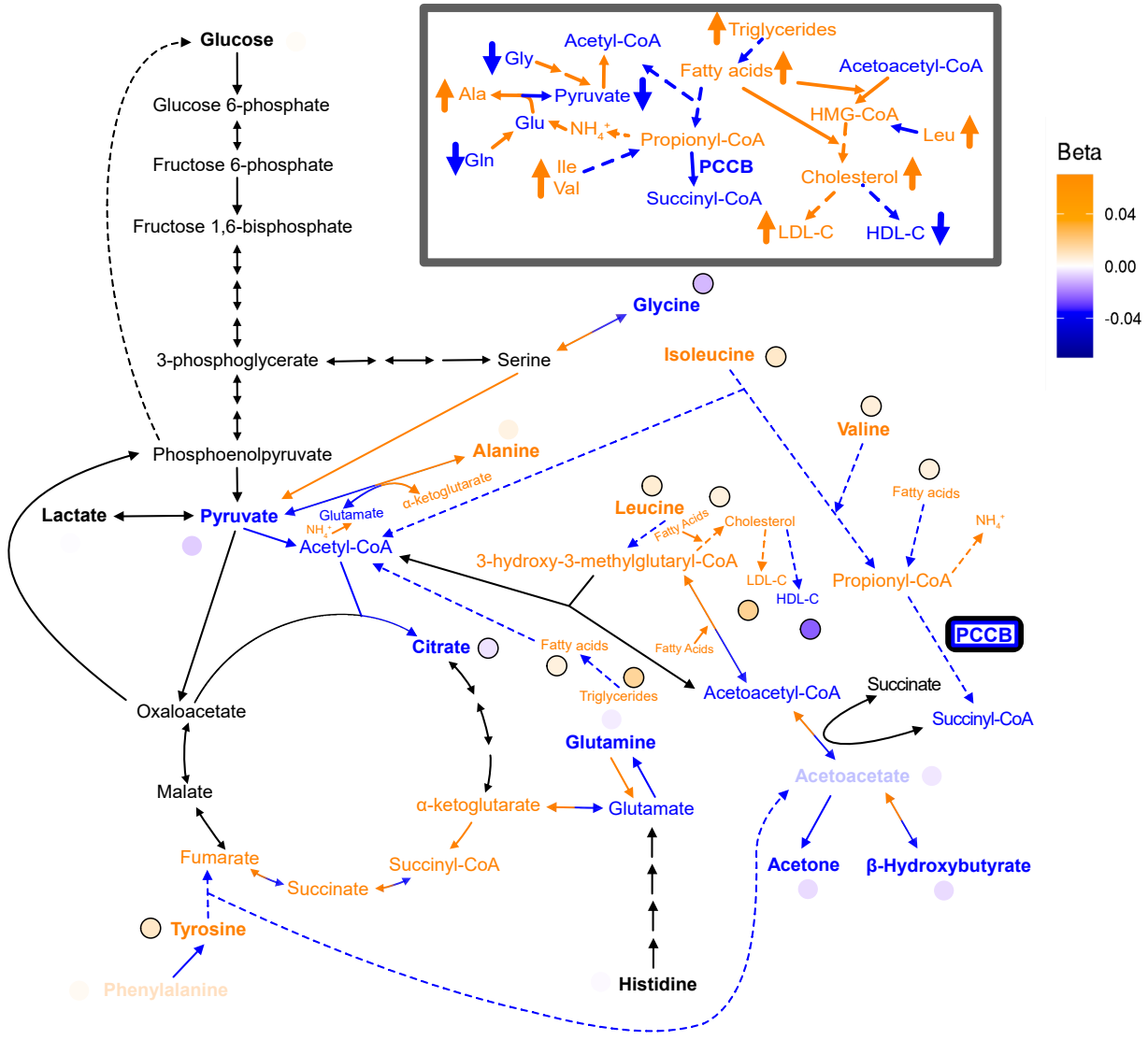

Figure 7 - figure supplement 1: Full Pathway Results for PCCB. *Full pathway results and possible mechanism for disease-associated variant rs61791721 with gene annotation PCCB.*

rs61791721 (PCCB): The variant rs61791721 has gene assignment *PCCB*, which encodes a protein that catalyzes the conversion of propionyl-CoA to succinyl-CoA (Supplementary Expanded Pathway Figure 7 - figure supplement 1). The hypothesized mechanism is that the variant decreases PCCB activity, resulting in lower levels of succinyl-CoA and increased propionyl-CoA. The increased propionyl-CoA results in excess ammonium being produced because propionyl-CoA inhibits N-acetylglutamate synthase, which is an important cofactor for the enzyme (carbamoyl phosphate synthetase) that catalyzes the first step in the urea cycle for ammonium capture. Alanine is a reservoir for nitrogen waste so there is an increase in pyruvate and glutamate to alanine and alpha-ketoglutarate to capture the toxic ammonium [34, 35]. More glycine is then broken down in response to the decrease in pyruvate levels and more glutamine is converted to glutamate, leading to a decrease in glycine and glutamine levels. Conversely, the increased levels of propionyl-CoA mean less valine and isoleucine need to be broken down, resulting in an increase in their levels. In addition, the increased alpha-ketoglutarate increases downstream citric acid cycle intermediates meaning less tyrosine needs to be broken down to produce fumarate. Also in response to increased propionyl-CoA, fewer fatty acids are broken down, resulting in decreased acetyl-CoA, which is typically the downstream product, decreased citrate, which is one step downstream of acetyl-CoA, increased total fatty acid levels and increased total triglyceride levels. This increase may stimulate the activity of HMG-CoA reductase and synthase, resulting in increased conversion of ketone bodies to acetoacetyl-CoA to HMG-CoA and cholesterol. This results in decreased levels of ketone bodies and increased HMG-CoA and cholesterol. Increased HMG-CoA leads to increased leucine, while increased cholesterol leads to an increase in LDL-C and a decrease in HDL-C. While it is not necessarily intuitive that an increase in cholesterol levels would decrease HDL-C, it is possible that this increase activates the Rho A signal transduction pathway and suppresses peroisome proliferator-activated receptor alpha which then decreases the amount of ApoA1, as the reverse may explain why patients on statins may experience an increase in HDL-C despite a decrease in total cholesterol [65]. ApoA1 is an essential protein for HDL and thus a decrease in its levels would result in a decrease in HDL-C. We also found that *PCCB* was the strongest eGene in GTEx for rs61791721, with an effect shared across most tissues (Supplementary Figure 7 - figure supplement 5). Note, we additionally ran GWAS for total triglycerides, total fatty acids, HDL cholesterol, and LDL cholesterol as part of the interpretation of the *PCCB* variant because they are important metabolites in the discussion of cardiometabolic disease (Supplementary Figure 7 - figure supplement 6).

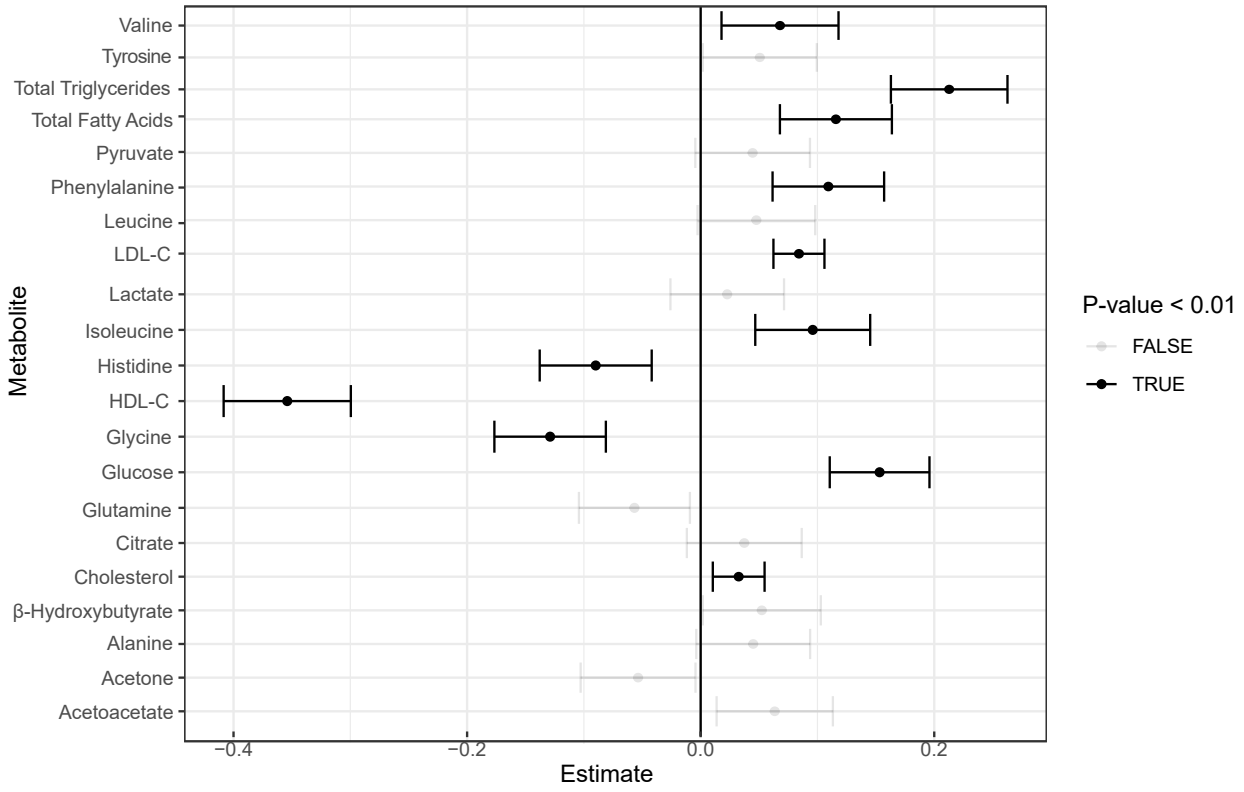

Figure 7 - figure supplement 2: Incident analysis. *Forest plot for CAD incident analysis for the 16 metabolites as well as HDL-C, LDL-C, total triglycerides, cholesterol, and total fatty acids in the UK Biobank.*

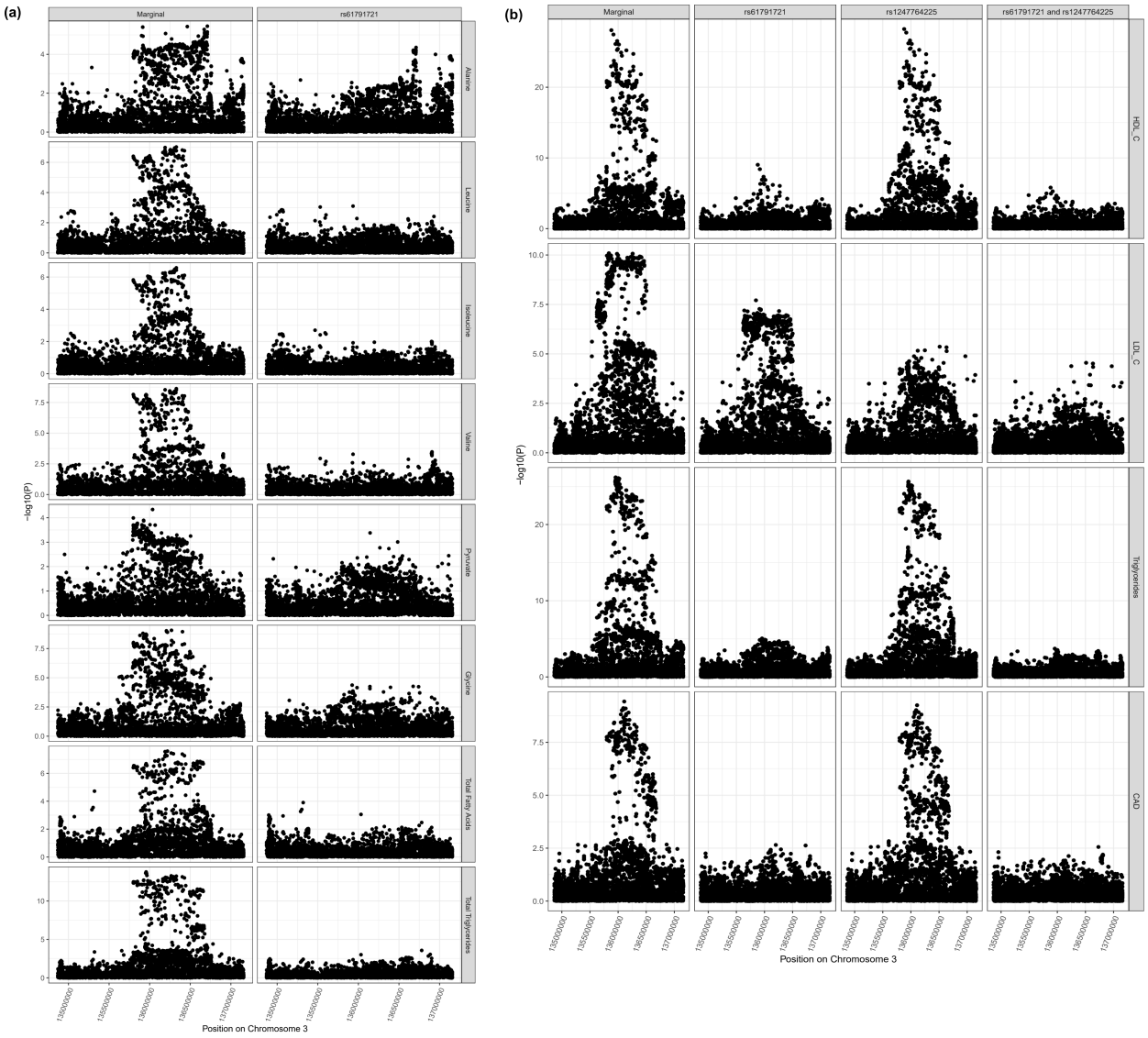

Figure 7 - figure supplement 3: PCCB locus colocalization. *Colocalization results for PCCB locus demonstrating pleiotropic effects of variant rs61791721 on alanine, glycine, pyruvate, isoleucine, valine, leucine, total fatty acids, total triglycerides, total free cholesterol, the UK Biobank clinical measures of HDL-C, LDL-C and triglycerides, and CAD.*

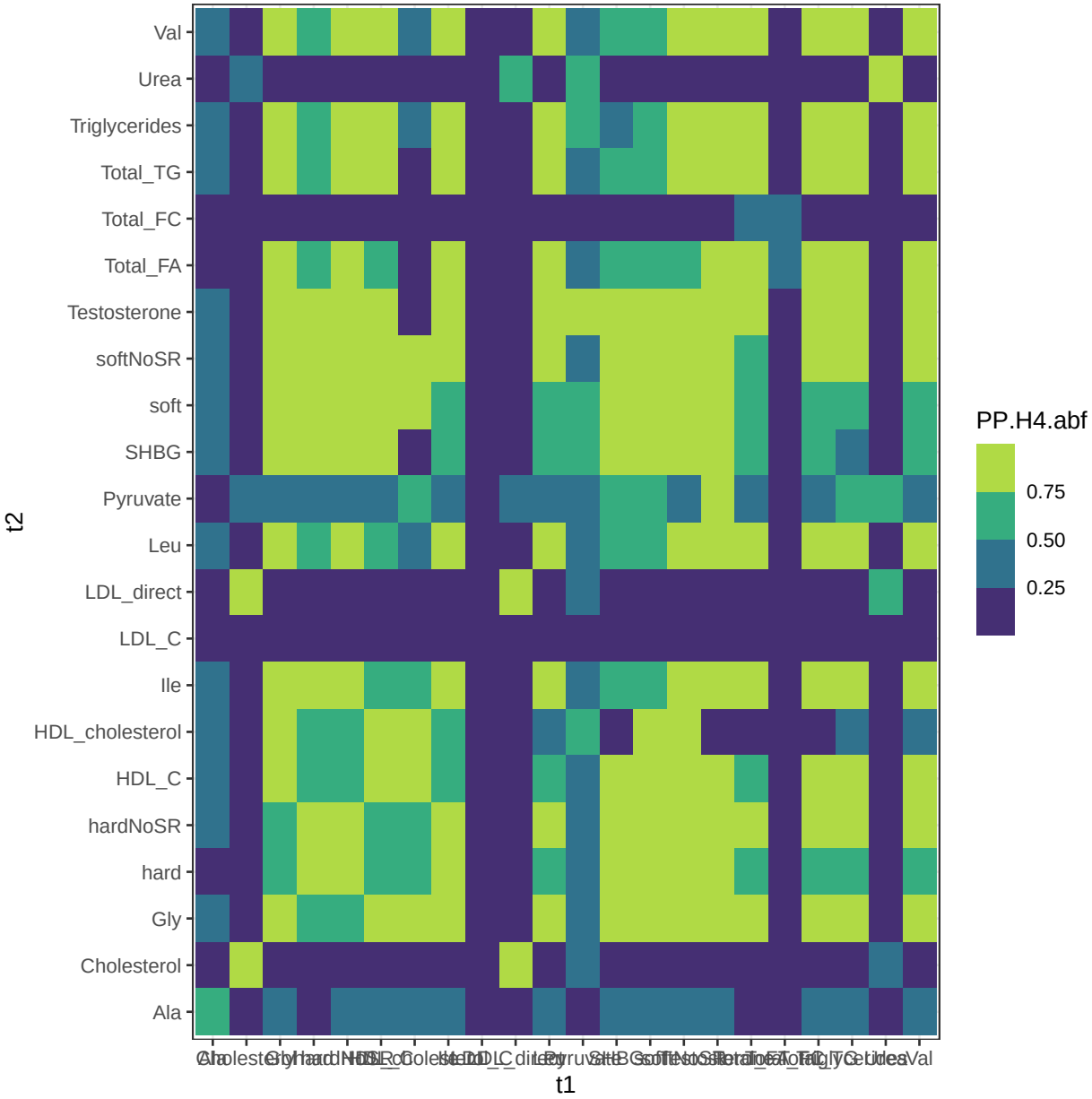

Figure 7 - figure supplement 4: Colocalization at PCCB for metabolites. *Colocalization probabilities at PCCB for the joint association signals of all metabolites and clinical traits. hard, hard CAD; soft, soft CAD (with angina); SR, self-reported cases included. All traits except Total free cholesterol* $(Total_FC), Urea, LDL, and total cholesterol colocalize together; LDL was initially well powered enough to detect multi$ $figure supplement 3), supporting the alternative coloc pattern.$

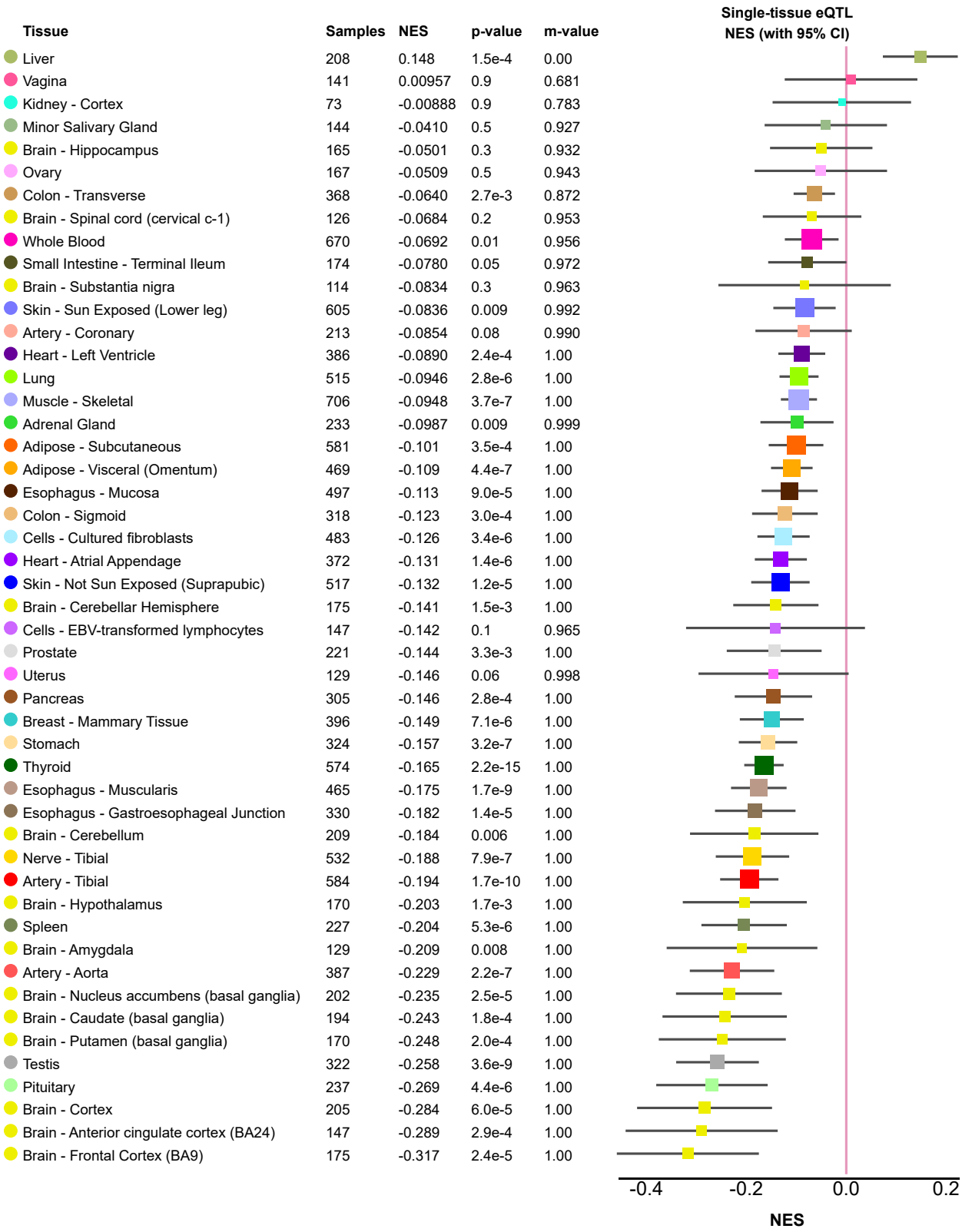

Figure 7 - figure supplement 5: rs61791721 eQTL effects. *Tissue eQTL effects for the PCCB variant (rs61791721) in GTEx.*

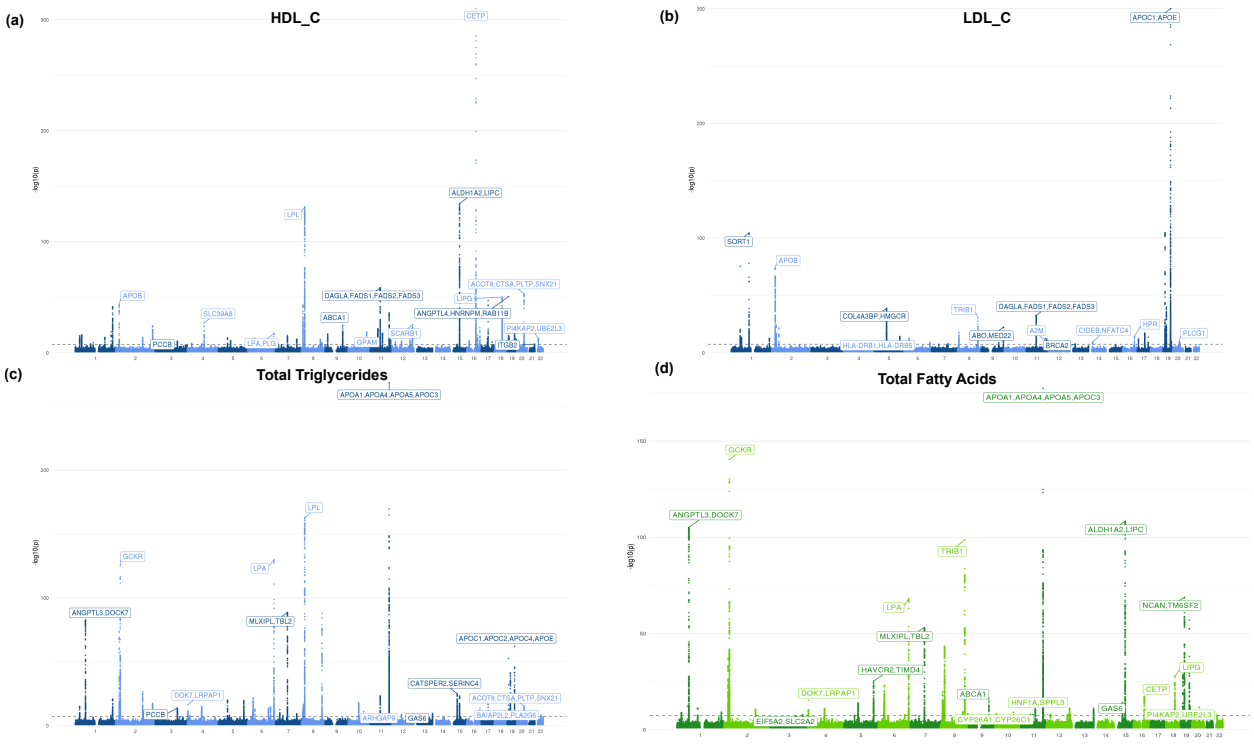

Figure 7 - figure supplement 6: Additional manhattan plots. *Manhattan plots for (a) HDL_C, (b) LDL_C, (c) total triglycerides, and (d) total fatty acids. Blue coloring represents metabolites belonging to the lipid group and green belongs to fatty acid group.*

Supplementary Files

Supplementary File 1: Metabolite HESS heritabilities. *Heritibility results from HESS for each metabolite.*
(File: hess_scaled_heritabilities.tsv)

Supplementary File 2: Gene function sources. *Sources for biochemical characterization of genes mentioned in the variant vignettes.*
(File: gene_biochemistry_sources.xlsx)

Supplementary File 3: Metabolite GWAS hits annotation. *Annotation for the 213 metabolite GWAS hits, including the assigned gene, assigned gene type, variant classification, and nearest gene.*
(File: rigidmetabolites_sig_pruned_snplist_manual_annotation_closestgene_disttss.tsv)

Supplementary File 4: Gene biochemical groups. *The number of significant ($P < 1e\text{-}4$) metabolite associations each metabolite GWAS gene had for each biochemical group. For a given gene, only biochemical groups that had at least one significant metabolite association were listed.*
(File: gene_metgroup_assignments_all.tsv)

Supplementary File 5: Ancestry-inclusive GWAS hits. *List of additional metabolite GWAS hits from the ancestry-inclusive analysis that were not present in the European-only GWAS results.*
(File: newsnps_annotated_rigid_novel_withsumstats.tsv)

Supplementary File 6: Discordant variant annotation. *List of each discordant variant-metabolite association, including the variant annotations and relevant metabolite pair genetic correlation and GWAS summary statistics.*
(File: disvar_annotation.tsv)

Supplementary File 7: Local genetic correlation results. *Combined results for different methods of calculating the local genetic correlation for different pathways for alanine and glutamine, demonstrating the consistency across the different approaches.*
(File: combined_localgencor_methods.tsv)

Supplementary File 8: Metabolite associations with CAD. *Literature evidence and citations for metabolite associations with CAD.*
(File: met_to_disease.xlsx)

Supplementary File 9: PCCB GWAS results. *GWAS summary statistics for rs61791721 (PCCB) in the 16 metabolites.*
(File: PCCB_sumstats.tsv)

841

842

References

[1] Nadia Solovieff, Chris Cotsapas, Phil H. Lee, Shaun M. Purcell, and Jordan W. Smoller. Pleiotropy in complex traits: challenges and strategies. , 14(7):483–495.

[2] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. , 47(3):291–295.

[3] Huwenbo Shi, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. , 101(5):737–751.

[4] Curtis R. Warren, John F. O'Sullivan, Max Friesen, et al. Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease. , 20(4):547–557.e7.

[5] Nasa Sinnott-Armstrong, Isabel S. Sousa, Samantha Laber, et al. A regulatory variant at 3q21.1 confers an increased pleiotropic risk for hyperglycemia and altered bone mineral density. , 33(3):615–628.e13.

[6] Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, et al. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. , 4(11):e1000282.

[7] Nasa Sinnott-Armstrong, Sahin Naqvi, Manuel Rivas, and Jonathan K Pritchard. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. , 10:e58615.

[8] Mathias Woidy, Ania C. Muntau, and Søren W. Gersting. Inborn errors of metabolism and the human interactome: a systems medicine approach. , 41(3):285–296.

[9] So-Youn Shin, Eric B. Fauman, Ann-Kristin Petersen, et al. An atlas of genetic influences on human blood metabolites. , 46(6):543–550.

[10] Victoria Au Yeung. *Common 'Inborn Errors' of Metabolism in the General Population*. Thesis, University of Cambridge.

[11] Heli Julkunen, Anna Cichońska, P Eline Slagboom, Peter Würtz, and Nightingale Health UK Biobank Initiative. Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. , 10:e63033.

[12] Lori Laffel. Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes. , 15(6):412–426.

[13] Marta Guasch-Ferré, José L Santos, Miguel A Martínez-González, et al. Glycolysis/gluconeogenesis- and tricarboxylic acid cycle–related metabolites, Mediterranean diet, and type 2 diabetes. , 111(4):835–844.

[14] P. Newsholme, K. Bender, A. Kiely, and L. Brennan. Amino acid metabolism, insulin secretion and diabetes. , 35(5):1180–1186.

[15] Aldons J. Lusis and James N. Weiss. Cardiovascular Networks. , 121(1):157–170.

[16] Matthew J Watt, Paula M Miotto, William De Nardo, and Magdalene K Montgomery. The Liver as an Endocrine Organ—Linking NAFLD and Insulin Resistance. , 40(5):1367–1393.

[17] Rozenn N. Lemaitre, Toshiko Tanaka, Weihong Tang, et al. Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. , 7(7):e1002193.

[18] Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, et al. Human metabolic individuality in biomedical and pharmaceutical research. , 477(7362):54–60.

[19] Tanya M Teslovich, Daniel Seung Kim, Xianyong Yin, et al. Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study. , 27(9):1664–1674.

[20] Sarah E. Graham, Shoa L. Clarke, Kuan-Han H. Wu, et al. The power of genetic diversity in genome-wide association studies of lipids. , pages 1–11.

[21] Rico Rueedi, Roger Mallol, Johannes Raffler, David Lamparter, Nele Friedrich, Peter Vollenweider, Gérard Waeber, Gabi Kastenmüller, Zoltán Kutalik, and Sven Bergmann. Metabomatching: Using genetic association to identify metabolites in proton NMR spectroscopy. , 13(12):e1005839.

[22] Johannes Kettunen, Ayşe Demirkan, Peter Würtz, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. , 7(1):11122.

[23] Laura B. L. Wittemans, Luca A. Lotta, Clare Oliver-Williams, et al. Assessing the causal association of glycine with risk of cardio-metabolic diseases. , 10(1):1060.

[24] Luca A. Lotta, Maik Pietzner, Isobel D. Stewart, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. , 53(1):54–64.

[25] Janne Pott, Yoon Ju Bae, Katrin Horn, et al. Genetic Association Study of Eight Steroid Hormones and Implications for Sexual Dimorphism of Coronary Artery Disease. , 104(11):5008–5023.

[26] Anna Cichonska, Juho Rousu, Pekka Marttinen, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. , 32(13):1981–1989.

[27] Sanni E. Ruotsalainen, Juulia J. Partanen, Anna Cichonska, et al. An expanded analysis framework for multivariate GWAS connects inflammatory biomarkers to functional variants and disease. , 29(2):309–324.

[28] Guanghao Qi and Nilanjan Chatterjee. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. , 14(10):e1007549.

[29] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. , 47(3):284–290.

[30] Jian Yang, Beben Benyamin, Brian P. McEvoy, et al. Common SNPs explain a large proportion of the heritability for human height. , 42(7):565–569.

[31] Brendan Bulik-Sullivan, Hilary K. Finucane, Verneri Anttila, et al. An atlas of genetic correlations across human diseases and traits. , 47(11):1236–1241.

[32] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika K. Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. , 104(1):65–75.

[33] Satoshi Koyama, Kaoru Ito, Chikashi Terao, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. , 52(11):1169–1177.

[34] Parith Wongkittichote, Nicholas Ah Mew, and Kimberly A. Chapman. Propionyl-CoA carboxylase – A review. , 122(4):145–152.

[35] L. D. Smith and U. Garg. Chapter 5 - Urea cycle and other disorders of hyperammonemia. In Uttam Garg and Laurie D. Smith, editors, Biomarkers in Inborn Errors of Metabolism, pages 103–123. Elsevier, San Diego.

[36] Nan Wu, Lindsei K. Sarna, Sun-Young Hwang, Qingjun Zhu, Pengqi Wang, Yaw L. Siow, and Karmin O. Activation of 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase during high fat diet feeding. , 1832(10):1560–1568.

[37] W H Salam, H G Wilcox, L M Cagen, and M Heimberg. Stimulation of hepatic cholesterol biosynthesis by fatty acids. Effects of oleate on cytoplasmic acetoacetyl-CoA thiolase, acetoacetyl-CoA synthetase and hydroxymethylglutaryl-CoA synthase. , 258(2):563–568.

[38] Therese Tillin, Alun D. Hughes, Qin Wang, et al. Diabetes risk and amino acid profiles: cross-sectional and prospective analyses of ethnicity, amino acids and diabetes in a South Asian and European cohort from the SABRE (Southall And Brent REvisited) Study. , 58(5):968–979.

[39] Raimo Jauhiainen, Jagadish Vangipurapu, Annamaria Laakso, Teemu Kuulasmaa, Johanna Kuusisto, and Markku Laakso. The Association of 9 Amino Acids With Cardiovascular Events in Finnish Men in a 12-Year Follow-up Study. , 106(12):3448–3454.

[40] Anubha Mahajan, Daniel Taliun, Matthias Thurner, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. , 50(11):1505–1513.

[41] Eeva Sliz, Sylvain Sebert, Peter Würtz, et al. NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. , 27(12):2214–2223.

[42] Clare Bycroft, Colin Freeman, Desislava Petkova, et al. The UK Biobank resource with deep phenotyping and genomic data. , 562(7726):203–209.

[43] Cristen J. Willer, Ellen M. Schmidt, Sebanti Sengupta, et al. Discovery and Refinement of Loci Associated with Lipid Levels. , 45(11):1274–1283.

[44] Céline Bellenguez, Amy Strange, Colin Freeman, Peter Donnelly, and Chris C.A. Spencer. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. , 28(1):134–135.

[45] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. , 4(1):s13742–015–0047–8.

[46] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: tool for the unification of biology. , 25(1):25–29.

[47] The Gene Ontology Consortium, Seth Carbon, Eric Douglass, et al. The Gene Ontology resource: enriching a GOld mine. , 49(D1):D325–D334.

[48] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. , 28(1):27–30.

[49] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. , 27(12):1739–1740.

[50] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. , 102(43):15545–15550.

[51] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. , 54(1).

[52] Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in UK Biobank. , 562(7726):210–216.

[53] Maya Ghoussaini, Edward Mountjoy, Miguel Carmona, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. , 49(D1):D1311–D1320.

[54] Edward Mountjoy, Ellen M. Schmidt, Miguel Carmona, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. , 53(11):1527–1533.

[55] Pathways of Human Metabolism Map. .

[56] Mingcong Xu, Xuefeng Bai, Bo Ai, et al. TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. , 50(D1):D402–D412.

[57] Genevieve L. Wojcik, Mariaelisa Graff, Katherine K. Nishimura, et al. Genetic analyses of diverse populations improves discovery for complex traits. , 570(7762):514–518.

[58] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. , 10(5):e1004383.

[59] Chris Wallace. Statistical Testing of Shared Genetic Control for Potentially Related Traits. , 37(8):802–813.

[60] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. , 99(1):139–153.

[61] Jack Bowden, Wesley Spiller, Fabiola Del Greco M, Nuala Sheehan, John Thompson, Cosetta Minelli, and George Davey Smith. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. , 47(4):1264–1278.

[62] Nasa Sinnott Armstrong, Yosuke Tanigawa, David Amar, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. , 53(2):185–194.

[63] Matthew Stephens. False discovery rates: a new deal. , 18(2):275–294.

[64] Michael Inouye, Gad Abraham, Christopher P. Nelson, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. , 72(16):1883–1893.

[65] Geneviève Martin, Hélène Duez, Christophe Blanquart, Vincent Berezowski, Philippe Poulain, Jean-Charles Fruchart, Jamila Najib-Fruchart, Corine Glineur, and Bart Staels. Statin-induced inhibition of the Rho-signaling pathway activates PPAR and induces HDL apoA-I. , 107(11):1423–1432.