Received: 23 October 2024 | Revised: 3 March 2025 | Accepted: 5 March 2025

DOI: 10.1002/oby.24291

ORIGINAL ARTICLE

Obesity A Research Journal | THE OBESITY SOCIETY | WILEY

Uncovering covariance patterns across energy balance traits enables the discovery of new obesity-related genes

Davorka Gulisija[1] | Agustin Gonzalez-Reymundez[2] | Jenifer I. Fenton[3] | Gustavo de los Campos[2,4,5] | Molly S. Bray[6] | Ana I. Vazquez[2,4]

[1]Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA

[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, USA

[3]Department of Food Science and Human Nutrition, Michigan State University, East Lansing, Michigan, USA

[4]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, Michigan, USA

[5]Department of Statistics & Probability, Michigan State University, East Lansing, Michigan, USA

[6]Department of Nutritional Sciences, The University of Texas at Austin, Austin, Texas, USA

Correspondence
Ana I. Vazquez
Email: avazquez@msu.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01DK119836, R03HG011674

Abstract

Objective: Effective solutions to obesity remain elusive, partly owing to its root in a positive energy balance (EB), which stems from the interplay of numerous traits spanning body size and composition, diet, physical activity, and metabolic profile. Nevertheless, EB-contributing traits are typically studied in isolation. We integrate numerous EB-related traits measured in the UK Biobank to uncover the underlying patterns of EB and associated genes in study participants.

Methods: We used sparse factor analysis to integrate traits and performed genome-wide association analyses on the integrated phenotypes to elucidate EB-related genes and metabolic pathways. We performed pleiotropy analyses on candidate single-nucleotide polymorphisms to uncover the genetic basis of EB.

Results: We identified multiple genes and genomic regions associated with EB, including many that have previously not been directly associated with obesity measures (e.g., *MIR5591*, *FNDC3B*, *ANAPC10*, *SULT1A1*, *AXIN1*, *SKIDA1*, *ERLIN1*, *DOCK7*), which we validated using an independent subset of the UK Biobank dataset along with data from the Atherosclerosis Risk in Communities cohort. We found that the covariances in EB traits are primarily driven by genome-wide pleiotropic associations.

Conclusions: We offer new insight into EB patterns and the genetic basis of EB.

INTRODUCTION

Excessive calorie intake and low physical activity contribute to a positive energy balance (EB) [1], the primary cause of obesity. In turn, obesity increases the risk for many health conditions including cardiovascular diseases [2], metabolic syndrome [2], cancer [2, 3], kidney disease [2], complications with infectious diseases such as COVID-19 [4], and mechanical issues [2]. Studying EB is challenging because it involves numerous traits acting in concert, including physical activity, diet, and body size traits [1]. Analyses of these traits in isolation have unveiled many genes (e.g., *FTO*, *APOB*, *GHR*, *PCSK1*) that contribute each of the different components of EB [5]. However, because EB is a compound of many traits that need to be addressed simultaneously, many obesity-related genes are yet to be identified.

Traits contributing to EB are known to have a genetic basis [5]. Animal studies have demonstrated inter-strain differences in the predisposition to physical activity [6] and feeding behavior [7]. In humans, many genes influence food preference and physical activity adherence [6]. However, the multivariate representation of EB, and the comprehensive set of EB-related genes, remains undefined.

See Commentary, pg. X.

Studying patterns of variation among numerous EB-related traits simultaneously may elucidate how the candidate genes shape EB and enhance the ability to detect yet unknown genes associated with this compound trait.

Additionally, a multivariate analysis of EB-related traits may be particularly valuable for detecting pleiotropic genes that affect several traits [8]. For example, some EB-related genes (e.g., *FTO*) exhibit pleiotropic associations with several EB components, including diet, physical activity, body size, and obesity risk [9]. Alleles at pleiotropic loci affect several phenotypes, producing genetic correlations. Notably, EB-related traits tend to be genetically correlated with other EB-related traits [10, 11]. Therefore, analyzing patterns of variation and the combined effects of EB-related traits can provide additional power to detect these otherwise elusive genes.

Studying EB traits simultaneously will aid in uncovering unknown genetic roots of EB. This can be achieved by constructing new variables that capture compound effects from multiple traits via latent factor analysis. Previous studies have focused on specific categories of traits such as diet [12, 13] or physical activity [14]. Joo et al. used sparse latent factor to describe dietary patterns [12] and examine their relationship with physical activity [13]. Johnson et al. [15] integrated dietary traits to investigate their correlation with metabolites linked to islet autoimmunity (type I diabetes) using reduced rank regression, whereas Xu et al. [14] identified patterns within physical activity using functional principal components. Advantages of these integrative studies include lower type I errors and increased power to detect underlying loci than other multivariate methods, such as multivariate ANOVA and separate regressions [8]. However, none of these studies have addressed an extensive set of complex traits across different categories that contribute to EB or examined genetic variants contributing to the combined traits.

In this study, we aim to identify patterns of EB (PEBs) from healthy individuals and use these patterns to uncover the underlying genetics. In order to achieve this, we integrated an extensive set of EB-contributing traits, categorized as body size and composition, diet, physical activity, and metabolic profiles, from the UK Biobank (UKB) [16]. We also identified genetic variants in association with PEBs and generated a set of candidate genes, including known and new genes with putative effects on EB. Finally, we performed a comprehensive validation using two datasets. Our approach thus robustly accounts for the complex biology of EB and uncovers the combined EB profiles in participants.

Study Importance

What is already known?

- Obesity arises from a complex interplay of numerous traits that contribute to energy balance (EB), many of which have a genetic basis.
- Whereas the genetics of individual traits have been studied, a comprehensive analysis of the integrated traits has not been conducted, to our knowledge.

What does this study add?

- This study highlights the utility of an integrative approach to study multiple correlated traits simultaneously.
- This analysis is the first, to our knowledge, to report integrated patterns of covariation in EB.

How might these results change the direction of research or the focus of clinical practice?

- Association analyses identified genetic variants underlying energy balance patterns and demonstrated that genome-wide pleiotropic associations drive covariation in the traits, which may inform future molecular studies.
- Integrated phenotypes provide new metrics that could aid in understanding individual EB profiles.

METHODS

The data used for this study consisted of genotypes, phenotypes, and demographics of individuals of European ancestry from the UKB. The discovery set included 219,049 distantly related Europeans. For internal validation, we used 72,153 independent European UKB participants, and external validation was performed using data from 9618 individuals of European ancestry from the Atherosclerosis Risk in Communities (ARIC) study [17].

We assume that EB in healthy individuals is primarily driven by energy in and out of the body indicated by the 32 traits (Table 1) that belonged to one of four categories: blood biomarkers (4 phenotypes), body size and composition (5), food consumption (17), and physical activity (6). Phenotypes were preadjusted by sex; age and age squared; sex-by-age interaction; recruitment center [18]; five single-nucleotide polymorphism (SNP)-derived principal components [19]; and Townsend Deprivation Index, household income, and participant qualifications. Additionally, blood biomarkers were adjusted for the use of medication [20] and the time since the last meal.

Although body mass index (BMI) is not included directly, we have incorporated its components, i.e., body weight and height, into PEBs. Glucose and triglycerides in blood have been included as dietary indicators.

For genotypes, we performed standard quality control and filtering of imputed SNPs with a resulting 13,113,819 SNPs (online Supporting Information Methods).

The PEBs were derived using sparse singular value decomposition on the adjusted and standardized phenotypes [21]. Briefly, this procedure extracts mutually independent features similar to standard principal components; however, the use of a sparsity penalty leads to some of the feature loadings being reduced to zero, which facilitates the interpretation of the patterns and phenotypes linked to them. A 95% confidence interval (CI) was built with the bootstrap procedure.

Genome-wide association analysis on PEBs

In order to identify variants associated with PEBs, we performed standard single-SNP association testing using ordinary least-squares regressions using the R package *BGData* (The R Project for Statistical Computing) [22] on the ~13 million call and imputed variants. This rendered an association p value for each SNP available in the UKB and each PEB; we use a conservative threshold to declare the significance of $\frac{1\times10^{-8}}{q}$ with $q = 5$ to account for multiple testing (PEBs 1–5).

For comparison, we also tested for association with individual phenotypes using multitrait genome-wide association studies (GWAS) (hereinafter referred as to "PLEIO") implemented using the *pleiotest* R package [23]. This package identifies variants associated with five, four, three, two, and single phenotypes.

We grouped the variants associated with phenotypes into 1-megabase (Mb) segments, annotated the closest gene, harboring variants significant for phenotypes using Ensembl's online variant effect prediction tool [24], and identified overlaps between GWAS findings and expression quantitative trail loci (eQTLs) from the Genotype-Tissue Expression (GTEx) Project [26] using *LDlink* [25]. We also identified pathways linked to PEBs (Ingenuity Pathway Analysis, QIAGEN Inc.). Details are in online Supporting Information Methods.

TABLE 1 Characteristics of original traits used in the construction of PEBs for the 219,049 UKB participants.

Numerical variables	Summary statistics: Median (MAD)	Summary statistics: % Missing data
Blood biomarkers		
Glucose, mmol/L	4.9 (0.5)	8.6
Cholesterol, mmol/L	5.8 (1.1)	0.0
Creatinine, μmol/L	70.0 (14)	0.1
Triglycerides, mmol/L	1.5 (0.8)	0.1
Body size and composition		
Weight, kg	77.0 (15)	0.3
Height, cm	170.0 (10)	0.2
Waist circumference, cm	90.0 (13)	0.1
Lean mass, kg	52.0 (13)	1.6
Fat mass, kg	23 (8.0)	1.8
Physical activity		
Moderate activity, MET-min/wk	480 (620)	2.7
Vigorous activity, MET-min/wk	240 (360)	2.7
Walking, MET-min/wk	690 (680)	2.7
Summed activity, MET-min/wk	1800 (1700)	2.7
Summed days of activity	11 (4.4)	0.0
Summed minutes of activity	100 (74.0)	2.7

Dietary variables	% Samples	% Missing data
Cooked vegetable	97	2.2
Salad/raw vegetable	90	5.0
Fresh fruit	94	3.2
Oily fish	57	0.3
Nonoily fish	67	0.9
Processed meat	61	0.1
Poultry	85	0.1
Beef	44	0.2
Lamb/mutton	25	0.3
Pork	26	0.3
Cheese	82	2.2
Bread	98	1.5
Cereal	88	4.0
Tea	85	3.1
Coffee	78	7.0
Water	91	7.1
Alcohol	90	0.1

Note: Numerical measures are summarized with the median (MAD) in their original units. Diet variables are summarized as the percentage of participants consuming the food item at least once a week. The percentage of missing data for all variables is presented in the third column. Most activity variables are given in MET-minutes. A MET-minute is computed by multiplying the MET (a measure of exercise intensity) score by the minutes performed. The summed total of the actual number of minutes of walking, moderate-intensity, and vigorous-intensity activities per week is given in the last row of numerical variables. The summed days of activity variable represents the sum of frequency (days) of walking and moderate-intensity and vigorous-intensity activities per week.

Abbreviations: MAD, median absolute deviation; MET, metabolic equivalent of task; PEB, pattern of energy balance; UKB, UK Biobank.

RESULTS

Construction and characterization of PEBs

Hereafter, we will refer to the traits measured as "original" traits, as opposed to "derived" by the integration of the original traits (i.e., PEBs). Original traits include dietary consumption from the UKB touch screen questionnaire (intake of water, fresh vegetables, fresh fruit, cooked vegetables, beef, pork, lamb, poultry, processed meats, oily and nonoily fish, coffee, tea, bread, alcohol, cereal, and cheese), blood biomarkers (glucose, triglycerides, cholesterol, and creatinine blood levels), body size and composition (weight, height, waist circumference, fat mass, and lean mass), and physical activity (intensity of walking, moderate and vigorous exercise, and frequency and duration of exercise). Summary statistics of original traits are presented in Table 1.

We used sparse factor analysis to uncover PEBs, defined as the first five latent variables following the procedure by Josse and Husson [21]. The analysis uses the correlation structure in data across the multiple traits to generate integrative factors. These factors cumulatively capture 14%, 24%, 30%, 35%, and 39% of the interindividual variance, as PEB1 through PEB5 are added sequentially. The contribution of the original traits to the construction of PEBs was measured by the corresponding factors' loadings (Figure 1). We imposed a sparsity constraint such that loading values near zero were considered a negligible contribution. We characterized each PEB based on the factor loadings for the traits involved. PEB1 primarily captures variation owing to individuals at a higher range of activity and lower body size, fat mass, and lean mass, whereas PEB2 primarily captures variation owing to a lower range of activity and body size, within the data. PEB3 primarily captures variation owing to a diet based on relatively low consumption of red meat with moderate consumption of processed foods, alcohol, cheese, and tea compared to the rest of the data. PEB4 primarily captures variation owing to intermediate height, with relatively high fruit and vegetable consumption. Finally, PEB5 primarily captures variation owing to individuals at a lower height range with relatively higher triglyceride and cholesterol levels. The PEBs summarize the combined variation in numerous traits across five distinct categories.

The analysis provides scores across the five PEBs for each participant, which together form a unique EB profile per individual, summarizing their dietary patterns, activity, body characteristics, and blood metabolites simultaneously. PEB1 and PEB2 explain more variation than any other factor and capture traits directly related to signals of obesity measures (Figure 1). Individuals with high scores of PEB1 or PEB2 are those in the lower range of body size and composition and can be characterized as relatively lean within the data range. Conversely, the low scores of these two PEBs capture obesity signals. For PEB5, a high score signals metabolic disorder risk, with low height potentially accounting for reduced body surface area and metabolic rate [27], in addition to the contributions of elevated cholesterol and triglycerides to disease risk. Using the participant scores for the five PEBs as traits in downstream genomic analyses thereby allows us to identify EB-related genetic loci.

PEBs are highly consistent across subsets of the data

For validation of PEBs, we analyzed the UKB White related partition consisting of individuals who are genetically related to, but independent of, the UKB sample used in the primary analysis. First, we generated PEBs with the same set of traits as the unrelated participants. In the validation, the same traits contributed to PEBs, and the trait loadings (as given in Figure 1) are similar between discovery and validation analyses (Figure 2, top row).

In addition, we obtained trait-loading CIs via nonparametric bootstrap with 1000 replicates to exclude from the comparison traits with loading intervals flanking zero. Remarkably, the overlap between the significant contributions estimated from the unrelated and related UKB partitions was 100%, 100%, 88%, 91%, and 75% for PEB1 to PEB5, suggesting that the PEBs are generalizable beyond the subset on which they were constructed. This serves as the first evidence of

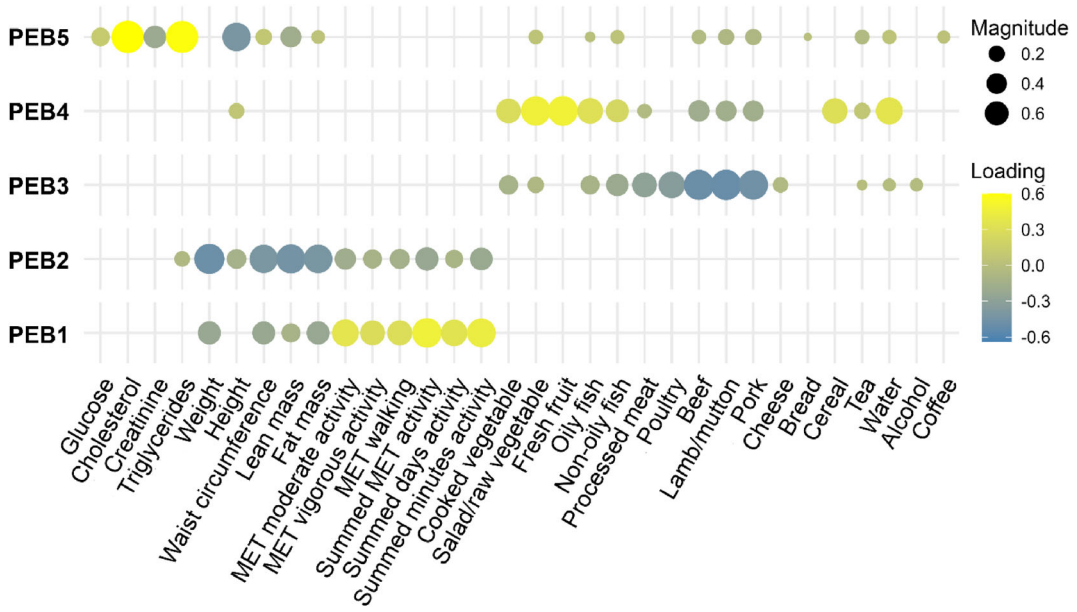

FIGURE 1 PEBs are defined by the contribution of distinct sets of EB-contributing traits. The size of the circles indicates the magnitude of the contribution (in terms of the absolute value of PEB loadings) of the original traits to PEBs. The color scheme indicates the direction of the contribution: positive values represent positive relationships between traits and PEBs (i.e., PEB values increase with trait phenotype values), whereas negative values indicate the opposite (i.e., PEB decreases with trait phenotype values). For example, PEB1 is characterized by low values across the body size and composition measures. EB, energy balance; PEB, pattern of energy balance.

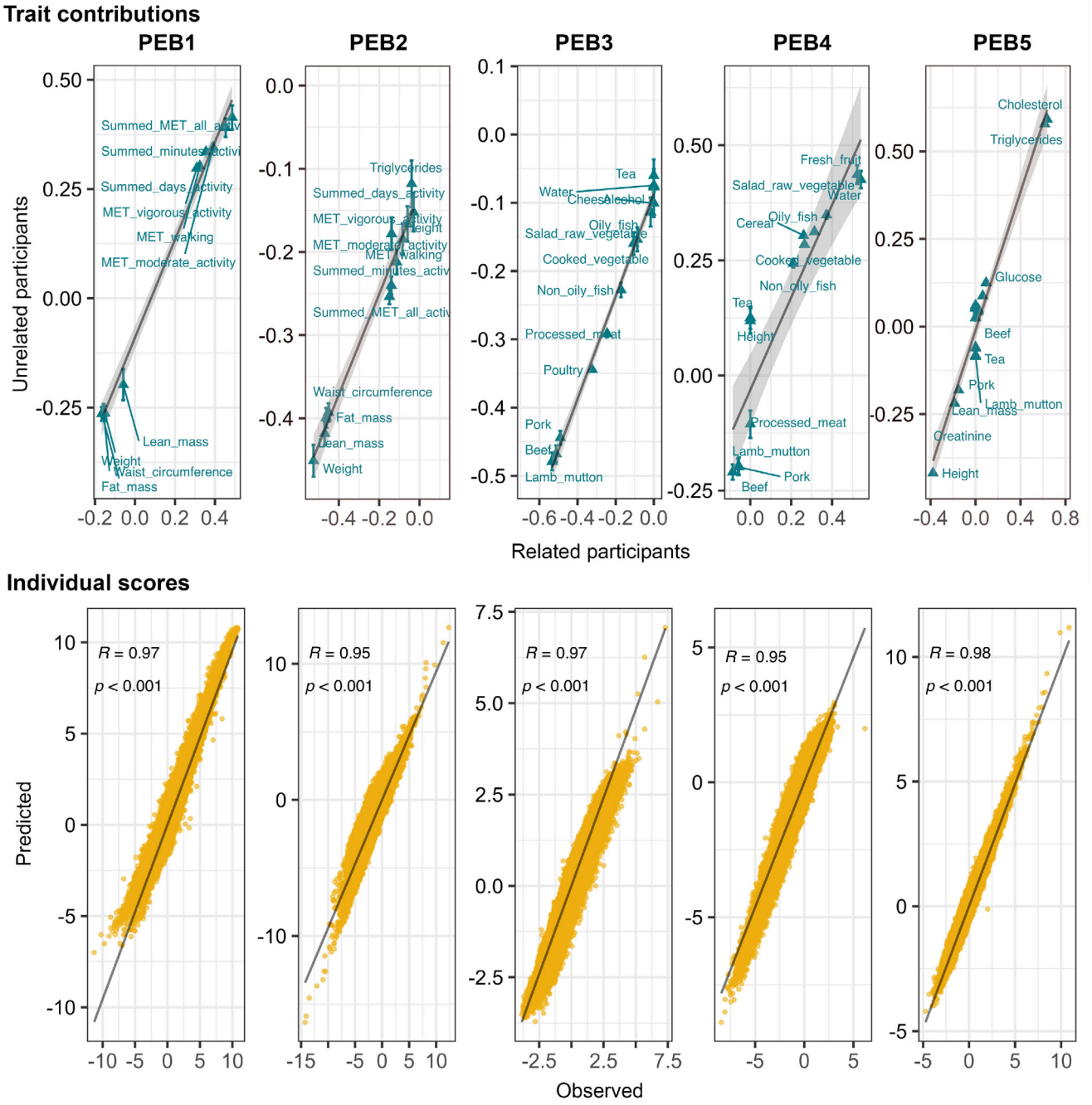

FIGURE 2 PEBs from the UKB White unrelated cohort (used for discovery) are highly predictive of the PEBs in the related UKB White cohort. The top row shows the contributions of original traits to PEB construction in two independent analyses: one in UKB unrelated participants (y-axis) and another analysis in UKB related participants (x-axis). The trait contributions (i.e., the loadings) were highly correlated in the two analyses. The bootstrap analysis shows the significant overlap between the traits with PEB loadings in unrelated and related sets. The overlap was 100%, 100%, 88%, 91%, and 75% for PEB1 to PEB5, respectively. The shaded area shows the 95% CI. The bottom row shows scatterplots, correlation, and *p* value between predicted and observed PEB scores for related UKB participants. PEB scores in related participants were predicted based on loadings obtained in the unrelated participant analysis (y-axis), whereas the observed PEB scores were obtained in independent factor analysis on the related participant data (x-axis). This figure shows that PEB construction is highly reliable across the two independent partitions of UKB. PEB, pattern of energy balance; UKB, UK Biobank.

the robustness of the construction of the EB profile, a set of unique participants' PEB scores.

Next, we predicted PEBs in the UKB related dataset using loadings from the unrelated participants (predicted) and compared them with those calculated independently on the UKB related sample (observed). The predicted and observed PEBs in the related participants were highly correlated ($r \geq 0.95$). Therefore, the unrelated participant loadings exhibited high power to predict PEBs in the related

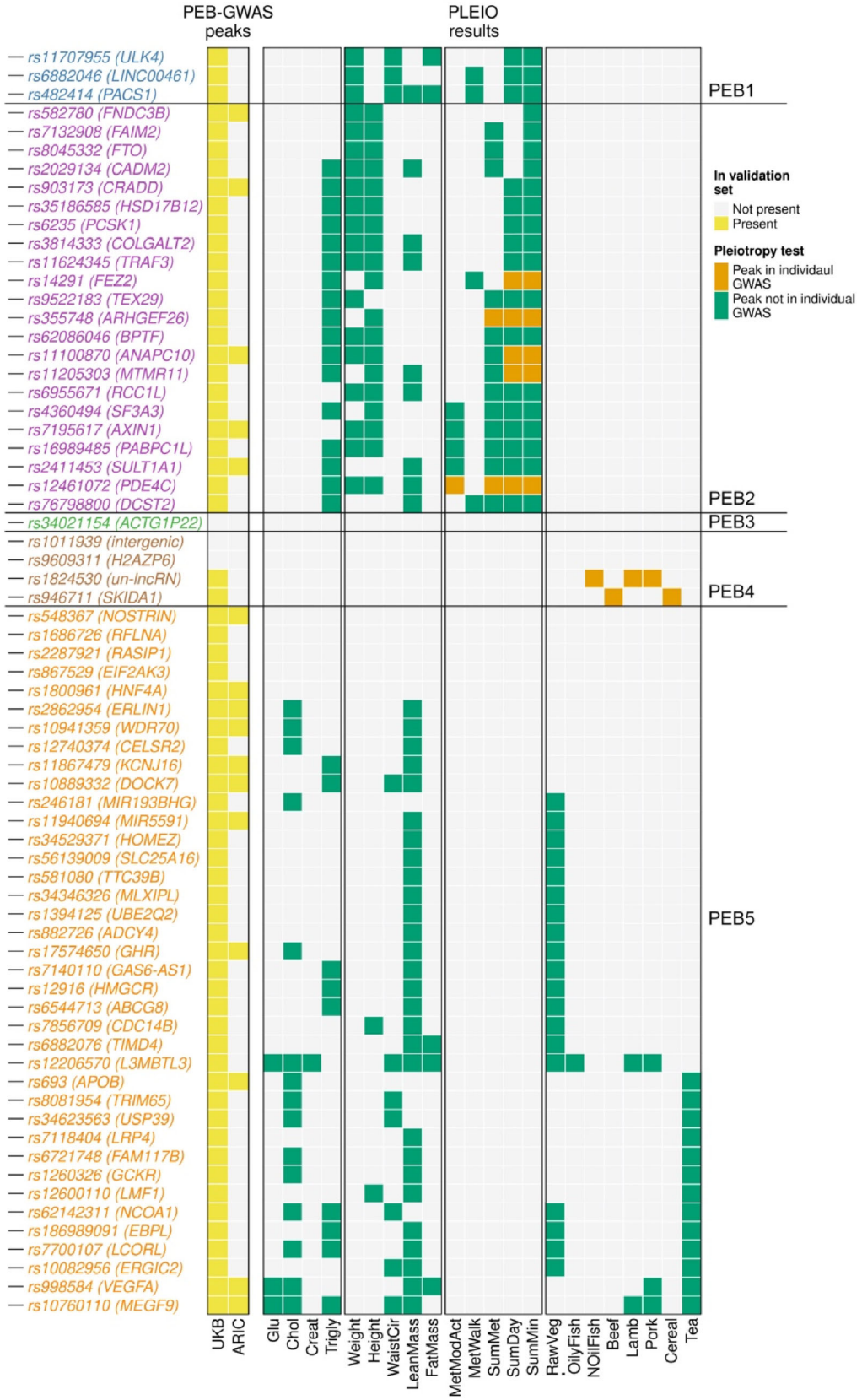

FIGURE 3 Legend on next page.

subset of participants, providing additional evidence that our findings are applicable in constructing personal profiles in participants with the same traits even when those participants were not used in the factor analysis.

Phenotypic variation is partitioned orthogonally across PEBs, ensuring near-zero phenotypic correlations between integrated traits ($<1e^{-5}$). These correlations arise from genetic and environmental covariance. Although most genetic and environmental correlations are negligible, a few show slightly larger absolute values (Table S1). Notably, PEB1, PEB2, and PEB5 share variation from overlapping traits, leading to shared significant SNPs and nontrivial genetic correlations among them.

GWAS of PEBs uncover pleiotropic genes that underlie variation in EB

We estimated heritability to gain insight into the degree of genetic control of PEBs. Then we performed GWAS to identify the following: 1) loci associated with PEBs; and (2) their pleiotropic effects on the original traits, for comparison.

Each PEB represents a novel heritable complex trait. Estimated heritabilities of 0.11, 0.30, 0.06, 0.08, and 0.37 for PEB1 through PEB5, respectively (p <0.05; Table S1), demonstrate that a substantial portion of phenotypic variation is driven by genetics. The relatively low heritability of PEB3 (0.06) likely reflects its composition of consumption of processed foods, alcohol, cheese, and tea, behavioral traits that are generally considered to have modest genetic influence. Conversely, the higher heritability of PEB2 and PEB5 (0.30 and 0.37) corresponds to their inclusion of structural traits with a stronger genetic basis, such as some body size traits (e.g., height).

Using PEB scores for the unrelated UKB participants as a discovery set, we conducted a series of GWAS analyses (PEB-GWAS) using the multivariate PEBs as outcomes. We found 41,020 SNPs that were significantly associated with PEBs (p <$[1 \times 10^{-8}]/5.0$ to adjust for multiple testing and five PEBs), indicating that more than 5% of genetic variants are associated with correlations in traits (i.e., genetic correlations). Because multiple PEBs were affected by the same original traits, and these traits are expected to be affected by multiple variants, we narrowed the characterization of SNPs to those significantly associated only with one PEB (PEB-specific variants). This restriction resulted in 14 PEB-specific SNPs for PEB1; 8903 for PEB2; 2 for PEB3; 289 for PEB4; and 27,313 for PEB5, with more SNPs identified for PEBs with higher heritability, as expected. We then lumped SNPs into groups ("peaks") and reported the most significant SNP within a 1-Mb window (Table S2). The lumping procedure identified 3 distinct peaks for PEB1, 149 for PEB2, 1 for PEB3, 4 for PEB4, and 239 for PEB5. Manhattan plots for the PEB-GWAS analyses are shown in Figures S1–S5.

A pleiotropy analysis of the original traits revealed that many genomic variants associated with PEBs (PEB-GWAS peaks) exhibited pleiotropic effects on the original traits. We refer to these results as "PLEIO." In particular, we found 356 peaks with pleiotropy. In particular, 106 peaks were associated with two traits at the same time, 75 were associated with three traits, 56 with four traits, 54 with five traits, 40 with six traits, 17 with seven traits, 7 with eight traits, and 1 peak was associated with ten original traits (Table S2).

Results from PLEIO were consistent with each PEB (Figure 3). PEB1 and PEB2 contributing variants exhibited pleiotropic associations with variables of both body size/composition and physical activity. PEB4 peaks exhibited pleiotropic associations with the consumption of nonoily fish, lamb, and pork and associations with the consumption of cereal and beef, whereas peaks at PEB5 were associated with body measurements, blood biomarkers, and, in some cases, dietary intake. We did not discover significant pleiotropic effects of variants in PEB3. These results confirm that phenotypes that are affected by the same pleiotropic alleles also group to form PEBs. Notably, the majority of pleiotropic genes detected by data integration and PLEIO were not detected when original traits were analyzed individually as a control (green vs. orange squares; Figure 3). Therefore, capturing the signal contained in several correlated traits increases the detection of otherwise weak associations of a pleiotropic locus via individual trait analysis. Moreover, integrated trait GWAS allow for uncovering pleiotropic genes underlying numerous positively correlated EB-contributing traits in a more efficient manner than PLEIO analysis that involves hypothesis testing for multiple traits.

GWAS peaks validated in the ARIC study and UKB related participant data

For the GWAS validation in the ARIC study, we identified eight traits corresponding to the discovery study (32 traits in UKB). The subset of traits for validation consisted of blood biomarkers (glucose, cholesterol, and triglycerides in blood), weight, height, waist circumference, and self-reported physical activity (intensity of moderate and vigorous exercise). Next, we conducted a new independent analysis in which we derived PEBs and conducted GWAS. Of the total 396 PEB-GWAS

FIGURE 3 PEB-GWAS and PLEIO study. The PEB-GWAS peaks are represented by their lead SNPs. The different colors of reference SNP IDs (left) indicate PEB1 through PEB5, with PEB1 in blue, PEB2 in violet, PEB3 in green, PEB4 in brown, and PEB5 in orange. Functional elements (e.g., coding, noncoding, intergenic regions) are included within the parentheses next to the SNP IDs. The first panel shows validation in the UKB related dataset and the ARIC study (yellow squares). The following panels summarize the results of PLEIO analysis, with the original trait labels at the bottom and green squares indicating peaks only detected in PLEIO analysis but not individual GWAS (control), given across four main categories of traits used in the analysis. ARIC, Atherosclerosis Risk in Communities; GWAS, genome-wide association study; ID, identifier; PEB, pattern of energy balance; PLEIO, individual phenotypes using multitrait GWAS; SNP, single-nucleotide polymorphism; UKB, UK Biobank.

peaks, 39 were validated in the ARIC study, i.e., 15 for PEB2 and 24 for PEB5. Despite a far smaller number of traits and participants in the ARIC study, we confirmed 39 PEB peaks offering strong indications of their biological importance as opposed to spurious correlations. The list of the validated peaks, associated traits, and gene sets can be found in Figure 3 and Table S2. A list of lead SNPs associated with 39 peaks validated in the ARIC study data is given in Table S3.

In addition, we conducted a second validation of the genetic associations using the White related UKB participants, with the same traits as the discovery data. With the 100% coverage of traits, the majority of the peaks inferred in the unrelated UKB participants were confirmed in the related UKB partition (Table S2).

Genomic variants associated with PEBs mapped onto previously known and unknown genomic regions

The entire set of annotated PEB-GWAS peaks is presented in Table S2 and eQTLs in Figure S6. Two PEB1 peaks mapped onto the following genes: *ULK4* and *PACS1*. *PACS1*, represented by rs482414, has been previously associated with obesity [28] and [29] sedentarism. The *ULK4* gene (e.g., rs11707955) has been associated with BMI [30], and its expression (Figure S6) has been associated with physical activity-related traits [31].

Five peaks of PEB2 that were validated in the ARIC study and the UKB related participants mapped onto *FNDC3B*, *ANAPC10*, *CRADD*, *SULT1A1*, and *AXIN1*. Two of the significant SNPs for PEB2 showed previous associations with body composition, physical activity, and blood triglycerides; rs582780 (in *FNDC3B*) showed association with hip circumference [32]; rs11100870 (*ANAPC10*) with BMI [33]; rs2411453 (*SULT1A1*) with body fat [32] and BMI [34]; *FNDC3B* with physical activity [35]; *ANAPC10* with lung capacity [36] and lean mass (thereby indirectly to exercise performance); and *AXIN1* with BMI [37].

Two PEB4 peaks mapped onto coding regions of *H2AZP6* and *SKIDA1*.

For PEB5, 11 peaks were validated in both datasets and mapped onto *DOCK7*, *APOB*, *NOSTRIN*, *MIR5591*, *GHR*, *WDR70*, *VEGFA*, *MEGF9*, *ERLIN1*, *KCNJ16*, and *HNF4A*. Many of these peaks were associated with obesity biomarkers, cholesterol, triglycerides, body composition, and dietary intake. rs2862954 (*ERLIN1*) was associated with cholesterol, lean mass, and high-density lipoprotein cholesterol [38]. The peak at rs10889332 (*DOCK7*), associated with triglycerides, lean mass, and waist circumference in our study, has been previously associated with serum lipid levels [39], perhaps reflecting an indirect effect of lipid metabolism and lean mass.

Genes associated with PEBs are enriched for numerous functional classes

PEB-GWAS peaks are typically mapped onto intronic (44%) or noncoding transcript variants (16%) and downstream (12%) or upstream gene variants (9%). In addition, most peaks were missense (76%), followed by synonymous (23%) and nonsense variants (1%). Details of the peaks and genomic regions are in Figure S6. PEB-GWAS peaks were enriched for 30 canonical pathways (false discovery rate p <0.05; Table S4). The 10 pathways with a lower p value (all with a false discovery rate p $<1 \times 10^{-4}$) were as follows: FXR/RXR activation; cell cycle: G1/S checkpoint regulation; protein kinase A signaling; insulin secretion signaling pathway; maturity onset diabetes of young signaling; cyclins and cell cycle regulation; aryl hydrocarbon receptor signaling; LXR/RXR activation; role of JAK2 in hormone-like cytokine signaling; and autophagy.

DISCUSSION

Addressing the obesity epidemic has proved to be challenging, partly because obesity arises from the interplay of numerous factors, including exogenous as well as behavioral phenotypes, that result in a positive EB in individuals. Therefore, understanding EB and mapping genetic loci associated with it may require considering multiple phenotypes jointly. Most obesity studies, however, have analyzed EB-related traits in isolation. Herein, we used sparse factor analysis to integrate an extensive set of EB-related phenotypes and identified multiple genetic loci associated with obesity through the compound proxy measure of EB, some of them previously known and others that are new associations.

Previous studies have integrated related traits but have been limited to a category, such as diet or physical activity [14]. Our comprehensive approach integrated traits across categories, i.e., dietary intake, physical activity, body size and composition, and metabolic profile, resulting in five main PEBs. Interestingly, the traits across distinct categories tend to covary, yielding unique cross-category integrative PEBs, which were previously unknown. PEB1 and PEB2 primarily integrated body size and composition with activity, distinguishing variation among generally active and inactive individuals. PEB3 integrated numerous diet traits, primarily driven by consumption of red meat, whereas PEB4 was primarily driven by produce consumption, capturing dietary preferences. Finally, PEB5 integrated traits across several categories and primarily captured variation owing to body size and consumption of various food items with triglyceride and cholesterol levels. Jointly, the five main PEBs summarized variation patterns in the extensive UKB data across numerous EB-related traits from distinct categories, profiling EB across participants.

The sparse factor analysis identified traits that have a partially shared genetic basis. For instance, the summed metabolic equivalent of task, summed days of activity, and summed minutes of activity all contribute to PEB1. The traits that primarily contribute to PEBs have similar variation patterns indicating likely common underlying environmental and/or genetic effects. For example, we found that different physical activities, from walking to vigorous exercise, co-occur, capturing generally active or inactive individuals. Interestingly, elevated blood lipids coincided with moderate consumption of a wide class of unprocessed meats.

This research additionally demonstrates the utility of sparse latent factor analysis [21] in summarizing numerous continuous and

categorical traits into compound phenotypes. The PEBs offered additional insight into the genetic basis of EB, extending beyond the results of numerous single-trait analyses. Previous multitrait studies, including canonical correlation analysis [40], multivariate multiple regression [41], and principal component analysis [8], have suggested that the integration may increase efficiency controlling for false positives that arise owing to multiple comparisons when testing traits independently. This control is achieved without sacrificing the power to detect true associations, such as in the case of multiple comparison corrections. Our approach produced meaningful integrative traits by imposing sparsity on factor loadings, resulting in factors composed of only a subset of nontrivial contributions. Therefore, the sparse latent factor provides a promising tool to advance the study of the genetics of complex and correlated traits by complementing results from individually based studies.

Our approach found genetic variants for EB that have already been reported for obesity-related traits and uncovered new variants underlying integrated EB-related traits that have not been directly associated with compound obesity-related phenotypes such as BMI. These results indicated a common genetic basis, potentially pleiotropic for covarying traits. Interestingly, the results of PLEIO analysis perfectly align with PEB loadings, i.e., the traits that strongly affected the same PEB tend to be underlined by the same pleiotropic gene(s), pointing to a partially shared genetic basis. Many of these pleiotropic loci were not previously associated with obesity measures, such as *MIR5591*, *FNDC3B*, *ANAPC10*, *SULT1A1*, *AXIN1*, *SKIDA1*, *ERLIN1*, and *DOCK7*. In addition, although we found a substantial agreement among GWAS results using PEB-GWAS and those using multivariate analysis through PLEIO test, the majority of newly discovered pleiotropic loci were not discovered in association analyses of individual traits.

Importantly, most of the regions harboring variants reaching GWAS-significant associations fall into protein-coding regions, many affecting splicing, including retained introns and nonsense mutations, which induce the production of nonfunctional or potentially incomplete proteins. Many of the detected variants are in genes involved in canonical EB pathways. Several genes were linked to insulin secretion signaling, LXR/RXR activation, and JAK2, which are established obesity pathways [42], inflammation [15], and adipocyte metabolism [43]. Additional genes mapped onto pathways proposed as emergent targets to combat obesity, such as FXR/RXR activation, aryl hydrocarbon receptor signaling, and autophagy [44]. Other genes mapped to gustation pathways (*ADCY4*), supporting the role of neurological processing of gustatory stimuli in food selection [45]. Some genes have been previously reported as cis-eQTLs. Other significant SNPs are in the protein-coding regions and exhibit pleiotropic effects.

The caveat of this approach is that a portion of the phenotypic variation is captured by the main PEBs. Particularly, PEB1 through PEB5 explained 39% of the total phenotypic variation owing to covariation among traits. Such analysis thus complements and does not replace the individual GWAS approach. Nevertheless, we identified new genes involved in EB and genes already known.

The study caveats include limited cross-validation, precorrections for socioeconomic status, and unsegregated sexes. We integrated an extensive set of EB traits, leveraging the UKB's large number of participants, wide breadth of traits, and consistent data collection. Our first tier of validation was on an independent partition of the UKB, which provided strong support for the predictability of PEBs. Nonetheless, the second tier of validation relied on an independent dataset with a significantly smaller number of participants and traits, adding a caveat to the study. Another caveat consists of our correction for socioeconomic status to account for the different access to lifestyle choices (e.g., access to a gym, quality food). This correction may reduce the power to discover SNPs associated with socioeconomic status. Finally, this study investigated the sexes together, limiting its ability for gender-specific effects.O

AUTHOR CONTRIBUTIONS

Ana I. Vazquez, Gustavo de los Campos, and Molly S. Bray developed the research plan and designed the analytic strategy for the study. Davorka Gulisija, Ana I. Vazquez, Agustin Gonzalez-Reymundez, and Gustavo de los Campos performed the statistical analyses and figure generation. Ana I. Vazquez, Molly S. Bray, Davorka Gulisija, and Agustin Gonzalez-Reymundez drafted the manuscript. Jenifer I. Fenton and Molly S. Bray contributed to the interpretation of results. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We acknowledge the UKB (application 15326) and ARIC (dbGap) studies and thank the participants and coordinators. This manuscript was not prepared in collaboration with investigators of these studies and does not necessarily reflect their opinions or conclusions.

CONFLICT OF INTEREST STATEMENT

The authors declared no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in UK Biobank at the following link: https://www.ukbiobank.ac.uk/. These data were derived from the following resources available in the public domain.

ORCID

Ana I. Vazquez https://orcid.org/0000-0001-6618-7486

REFERENCES

1. Hall KD, Heymsfield SB, Kemnitz JW, Klein S, Schoeller DA, Speakman JR. Energy balance and its components: implications for body weight regulation. *Am J Clin Nutr*. 2012;95:989-994. doi:10.3945/ajcn.112.036350
2. GBD 2015 Obesity Collaborators; Afshin A, Forouzanfar MH, Reitsma MB, et al. Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med*. 2017;377:13-27. doi:10.1056/NEJMoa1614362
3. Falk Libby E, Azrad M, Novak L, Vazquez AI, Wilson TR, Demark-Wahnefried W. Obesity is associated with higher 4E-BP1 expression in endometrial cancer. *Curr Biomark Find*. 2014;4:1-7.
4. Drucker DJ. Diabetes, obesity, metabolism, and SARS-CoV-2 infection: the end of the beginning. *Cell Metab*. 2021;33:479-498. doi:10.1016/j.cmet.2021.01.016

5. Jiang L, Penney KL, Giovannucci E, Kraft P, Wilson KM. A genome-wide association study of energy intake and expenditure. *PLoS One*. 2018;13:e0201555. doi:10.1371/journal.pone.0201555
6. Lightfoot JT, Turner MJ, Daves M, Vordermark A, Kleeberger SR, Timothy J. Genetic influence on daily wheel running activity level. *Physiol Genomics*. 2004;19:270-276. doi:10.1152/physiolgenomics.00125
7. Ellacott KLJ, Morton GJ, Woods SC, Tso P, Schwartz MW. Assessment of feeding behavior in laboratory mice. *Cell Metab*. 2010;12:10-17. doi:10.1016/j.cmet.2010.06.001
8. Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*. 2014;94:662-676. doi:10.1016/j.ajhg.2014.03.016
9. Rankinen T, Rice T, Teran-Garcia M, Rao DC, Bouchard C. FTO genotype is associated with exercise training-induced changes in body composition. *Obesity (Silver Spring)*. 2010;18:322-326. doi:10.1038/oby.2009.205
10. Reynolds RJ, Irvin MR, Bridges SL, et al. Genetic correlations between traits associated with hyperuricemia, gout, and comorbidities. *Eur J Hum Genet*. 2021;29:1438-1445. doi:10.1038/s41431-021-00830-z
11. Lupi AS, Sumpter NA, Leask MP, et al. Local genetic covariance between serum urate and kidney function estimated with Bayesian multitrait models. *G3 (Bethesda)*. 2022;12:jkac158. doi:10.1093/g3journal/jkac158
12. Joo J, Williamson SA, Vazquez AI, Fernandez JR, Bray MS. Advanced dietary patterns analysis using sparse latent factor models in young adults. *J Nutr*. 2018;148:1984-1992. doi:10.1093/jn/nxy188
13. Joo J, Williamson SA, Vazquez AI, Fernandez JR, Bray MS. The influence of 15-week exercise training on dietary patterns among young adults. *Int J Obes (Lond)*. 2019;43:1681-1690. doi:10.1038/s41366-018-0299-3
14. Xu SY, Nelson S, Kerr J, et al. Modeling temporal variation in physical activity using functional principal components analysis. *Stat Biosci*. 2019;11:403-421. doi:10.1007/s12561-019-09237-3
15. Johnson RK, Vanderlinden L, DeFelice BC, et al. Metabolite-related dietary patterns and the development of islet autoimmunity. *Sci Rep*. 2019;9:14819. doi:10.1038/s41598-019-51251-4
16. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z
17. Tang W, Basu S, Kong X, et al. Genome-wide association study identifies novel loci for plasma levels of protein C: the ARIC study. *Blood*. 2010;116:5032-5036. doi:10.1182/blood-2010-05-283739
18. Abdellaoui A, Dolan CV, Verweij KJH, Nivard MG. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nat Genet*. 2022;54:1345-1354. doi:10.1038/s41588-022-01158-0
19. Galinsky KJ, Loh P-R, Mallick S, Patterson NJ, Price AL. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. *Am J Hum Genet*. 2016;99:1130-1139. doi:10.1016/j.ajhg.2016.09.014
20. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*. 2021;53:185-194. doi:10.1038/s41588-020-00757-z
21. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016;70:1-31. doi:10.18637/jss.v070.i01
22. Grueneberg A, de los Campos G. BGData - a suite of R packages for genomic analysis with big data. *G3 (Bethesda)*. 2019;9:1377-1383. doi:10.1534/g3.119.400018
23. Aguate FM, Vazquez AI, Merriman TR, de los Campos G. Mapping pleiotropic loci using a fast-sequential testing algorithm. *Eur J Hum Genet*. 2021;29:1762-1773. doi:10.1038/s41431-021-00911-z
24. Kersey PJ, Allen JE, Allot A, et al. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucl Acids Res*. 2018;46:D802-D808. doi:10.1093/nar/gkx1011
25. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31:3555-3557. doi:10.1093/bioinformatics/btv402
26. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) Project. *Nat Genet*. 2013;45:580-585. doi:10.1038/ng.2653
27. Verbraecken J, Van de Heyning P, De Backer W, Van Gaal L. Body surface area in normal-weight, overweight, and obese adults. A comparison study. *Metabolism*. 2006;55:515-524.
28. Wheeler E, Huang N, Bochukova EG, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet*. 2013;45(5):513-517. doi:10.1038/NG.2607
29. van de Vegte YJ, Said MA, Rienstra M, van der Harst P, Verweij N. Genome-wide association studies and Mendelian randomization analyses for leisure sedentary behaviours. *Nat Commun*. 2020;11:1770. doi:10.1038/S41467-020-15553-W
30. Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet*. 2021;53:1415-1424. doi:10.1038/s41588-021-00931-x
31. Kichaev G, Bhatia G, Loh PR, et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am J Hum Genet*. 2019;104(1):65-75. doi:10.1016/j.ajhg.2018.11.008
32. Christakoudi S, Evangelou E, Riboli E, Tsilidis KK. GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis, adrenal cell renewal and cancer. *Sci Rep*. 2021;11:10688. doi:10.1038/S41598-021-89176-6
33. Liu Y, Zhang X, Lee J, et al. Genome-wide association study of neck circumference identifies sex-specific loci independent of generalized adiposity. *Int J Obes (Lond)*. 2021;45:1532-1541. doi:10.1038/s41366-021-00817-2
34. Hübel C, Gaspar HA, Coleman JRI, et al. Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am J Med Genet B Neuropsychiatr Genet*. 2019;180(6):428-438. doi:10.1002/ajmg.b.32709
35. Graff M, Scott RA, Justice AE, et al. Genome-wide physical activity interactions in adiposity —— a meta-analysis of 200,452 adults. *PLoS Genet*. 2017;13(4):130. doi:10.1371/journal.pgen.1006528
36. Shrine N, Guyatt AL, Erzurumluoglu AM, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet*. 2019;51:481-493. doi:10.1038/S41588-018-0321-7
37. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518(7538):197-206. doi:10.1038/NATURE14177
38. Richardson TG, Sanderson E, Palmerid TM, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med*. 2020;17:e1003062. doi:10.1371/JOURNAL.PMED.1003062
39. Guo T, Yin RX, Lin WX, Wang W, Huang F, Pan SL. Association of the variants and haplotypes in the DOCK7, PCSK9 and GALNT2 genes and the risk of hyperlipidaemia. *J Cell Mol Med*. 2016;20:243-265. doi:10.1111/JCMM.12713
40. Tang CS, Ferreira MAR. A gene-based test of association using canonical correlation analysis. *Bioinformatics*. 2012;28:845-850. doi:10.1093/bioinformatics/bts051

41. Basu S, Zhang Y, Ray D, Miller MB, Iacono WG, McGue M. A rapid gene-based genome-wide association test with multivariate traits. *Hum Hered*. 2014;76:53-63. doi:10.1159/000356016
42. Qatanani M, Lazar MA. Mechanisms of obesity-associated insulin resistance: many choices on the menu. *Genes Dev*. 2007;21:1443-1455. doi:10.1101/gad.1550907
43. Richard AJ, Stephens JM. The role of JAK-STAT signaling in adipose tissue function. *Biochim Biophys Acta*. 2014;1842:431-439. doi:10.1016/j.bbadis.2013.05.030
44. Filho CDDM, Downes M, Evans RM. Farnesoid X receptor an emerging target to combat obesity. *Dig Dis Sci*. 2017;35:185-190. doi:10.1159/000450909
45. Lemon CH. It's all a matter of taste: gustatory processing and ingestive decisions. *Mo Med*. 2010;107:247-251.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gulisija D, Gonzalez-Reymundez A, Fenton JI, de los Campos G, Bray MS, Vazquez AI. Uncovering covariance patterns across energy balance traits enables the discovery of new obesity-related genes. *Obesity (Silver Spring)*. 2025;1-11. doi:10.1002/oby.24291