Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight novel adipocyte biology

Yosuke Tanigawa[1*], Jiehan Li[2,3*], Johanne Marie Justesen[1,2,4], Heiko Horn[5,6], Matthew Aguirre[1,7], Christopher DeBoever[1,8], Chris Chang[9], Balasubramanian Narasimhan[1,10], Kasper Lage[5,6,11], Trevor Hastie[1,10], Chong Yon Park[2], Gill Bejerano[1,7,12,13], Erik Ingelsson[2,3*+], Manuel A. Rivas[1*+]

1. Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA.
2. Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA.
3. Stanford Cardiovascular Institute, Stanford University, Stanford, CA 94305.
4. Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
5. Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
6. Broad Institute of MIT and Harvard, Cambridge, MA, USA.
7. Department of Pediatrics, Stanford University School of Medicine, Stanford University, Stanford, CA, USA.
8. Department of Genetics, Stanford University, Stanford, CA, USA.
9. Grail, Inc., 1525 O'Brien Drive, Menlo Park, CA, USA.
10. Department of Statistics, Stanford University, Stanford, CA, USA.
11. Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark.
12. Department of Developmental Biology, Stanford University, Stanford, CA, USA.
13. Department of Computer Science, Stanford University, Stanford, CA, USA.

*These authors contributed equally

+Corresponding authors

Abstract

To characterize latent components of genetic associations, we applied truncated singular value decomposition (DeGAs) to matrices of summary statistics derived from genome-wide association analyses across 2,138 phenotypes measured in 337,199 White British individuals in the UK Biobank study. We systematically identified key components of genetic associations and the contributions of variants, genes, and phenotypes to each component. As an illustration of the utility of the approach to inform downstream experiments, we report putative loss of function variants, rs114285050 (*GPR151*) and rs150090666 (*PDE3B*), that substantially contribute to obesity-related traits, and experimentally demonstrate the role of these genes in adipocyte biology. Our approach to dissect components of genetic associations across human phenotypes will accelerate biomedical hypothesis generation by providing insights on previously unexplored latent structures.

Introduction

Human genetic studies have been profoundly successful at identifying regions of the genome contributing to disease risk[1,2]. Despite these successes, there are challenges to translating findings to clinical advances, much due to the extreme polygenicity and widespread pleiotropy of complex traits[3–5]. In retrospect, this is not surprising given that most common diseases are multifactorial. However, it remains unclear exactly which factors, acting alone or in combination, contribute to disease risk and how those factors are shared across diseases. With the emergence of sequencing technologies, we are increasingly able to pinpoint alleles, possibly rare and with large effects, which may aid in therapeutic target prioritization[6–13]. Furthermore, large population-based biobanks, such as the UK Biobank, have aggregated data across tens of thousands of phenotypes[14]. Thus, an opportunity exists to characterize the phenome-wide landscape of genetic associations across the spectrum of genomic variation, from coding to non-coding, and rare to common.

Singular value decomposition (SVD), a mathematical approach developed by differential geometers[15], can be used to combine information from several (likely) correlated vectors to form basis vectors, which are guaranteed to be orthogonal and to explain maximum variance in the data, while preserving the linear structure that helps interpretation. In the field of human genetics, SVD is routinely employed to infer genetic population structure by calculating principal components using the genotype data of individuals[16].

To address the pervasive polygenicity and pleiotropy of complex traits, we propose an application of truncated SVD (TSVD), a reduced rank approximation of SVD[17–19], to characterize the underlying (latent) structure of genetic associations using summary statistics computed for 2,138 phenotypes measured in the UK Biobank population cohort[14]. We applied our novel approach, referred to as DeGAs – Decomposition of Genetic Associations – to assess associations among latent components, phenotypes, variants, and genes. We highlight its application to body mass index (BMI), myocardial infarction (MI), and gallstones, motivated by high polygenicity in anthropometric traits, global burden, and economic costs, respectively. We assess the relevance of the inferred key components through GREAT genomic region ontology enrichment analysis[20] and functional experiments. Further, we experimentally demonstrate a role of newly discovered obesity-related genes in adipocyte biology.

Results

DeGAs method overview

We generated summary statistics by performing genome-wide association studies (GWAS) of 2,138 phenotypes from the UK Biobank (Fig. 1a, Supplementary Tables S1-S2). We perform variant-level quality control, which includes linkage-disequilibrium (LD) pruning and removal of variants in the MHC region, to focus on 235,907 variants for subsequent analyses. Given the immediate biological consequence, subsequent downstream implications, and medical relevance of predicted protein-truncating variants (PTVs), commonly referred to as loss-of-function variants[12,21,22], we perform separate analyses on two variant sets: (1) all directly-genotyped variants and (2) PTVs (Supplementary Fig. S1). To eliminate unreliable estimates of genetic associations, we selected associations with p-values < 0.001, and standard error of beta value or log odds ratio of less than 0.08 and 0.2, respectively, for each dataset. The Z-scores of these associations were aggregated into a summary statistic matrix W of size $N \times M$, where N and M denote the number of phenotypes and variants, respectively. N and M were 2,138 and

235,907 for the “all” variant group; and 628 and 784 for the PTV group. The rows and columns of W correspond to the GWAS summary statistics of a phenotype and the phenome-wide association study (PheWAS) of a variant, respectively. We applied TSVD to each matrix and obtained a decomposition into three matrices $W = USV^T$ (U: phenotype, S: variance, V: variant). This reduced representation of $K = 100$ components altogether explained 41.9% (all) and 75.5% (PTVs) of the variance in the original summary statistic matrices (Fig. 1b-c, Methods, Supplementary Fig. S2).

To characterize each latent component, we defined phenotype squared cosine score, phenotype contribution score, variant contribution score, and gene contribution score. The squared cosine scores quantifies the relative importance of component for a given phenotype or gene, and are defined based on the squared distance of a component from the origin on the latent space[23] (Fig. 1d, Methods). Contribution scores quantify relative importance of a phenotype, variant, or gene to a given component and is defined based on the squared distance of a phenotype, variant, or gene from the origin (Fig. 1d). We then performed biological characterization of DeGAs latent components with the genomic region enrichment analysis tool (GREAT)[20] followed by functional experiments in adipocytes (Fig. 1e).

Characterization of latent structures of DeGAs

The PCA plots show the projection of phenotypes and variants onto DeGAs latent components. (Fig. 2a-b). For the variant plot, we overlay biplot annotation as arrows to interpret the direction of the components (Fig. 2b). Overall, we find that the first five principal components of genetic associations can be attributed to: 1) fat-free mass that accounts for the “healthy part” of body weight[24] (32.7%, Supplementary Table S3) and two intronic variants in *FTO* (rs17817449: contribution score of 1.15% to PC1, rs7187961: 0.41%); and a genetic variant proximal to *AC105393.1* (rs62106258: 0.46%); 2) whole-body fat mass (61.5%) and the same three *FTO* and *AC105393.1* variants (rs17817449: 0.97%, rs7187961: 0.28%, rs62106258: 0.27%); 3) bioelectrical impedance measurements (38.7%), a standard method to estimate body fat percentage[25,26], and genetic variants proximal to *ACAN* (rs3817428: 0.64%), *ADAMTS3* (rs11729800: 0.31%), and *ADAMTS17* (rs72770234: 0.29%); 4) eye meridian measurements (80.9%), and two intronic variants in *WNT7B* (rs9330813: 5.73%, rs9330802: 1.14%) and a genetic variant proximal to *ATXN2* (rs653178: 0.96%); and 5) bioelectrical impedance and spirometry measures (45.4% and 26.0%, respectively) and genetic variants proximal to *FTO* (rs17817449: 0.17%), *ADAMTS3* (rs11729800: 0.11%), and *PSMC5* (rs13030: 0.11%) (Fig. 2c-d, Supplementary Table S4).

Applying DeGAs components for BMI, MI, and gallstones

To illustrate the application of DeGAs in characterizing the genetics of complex traits, we selected three phenotypes, BMI, MI, and gallstones given the large contribution of anthropometric traits on the first five components, that ischemic heart diseases is a leading global fatal and non-fatal burden, and that gallstones is a common condition with severe pain and large economic costs where polygenic risk factors are largely unknown[27,28]. We identified the top three key components for these three phenotypes with DeGAs using the “all” variants dataset.

For BMI, we find that the top three components of genetic associations (PC2, PC1, and PC30) altogether explained over 69% of the genetic associations (47%, 18%, and 4%, respectively, Supplementary Fig. S3a). The top two components (PC2 and PC1) corresponded to components of body fat (PC2) and fat-free mass measures (PC1), as described above. PC30 was driven by fat mass (28.7%) and fat-free mass (6.8%), but also by non-melanoma skin cancer (7.72%) – linked to BMI in epidemiological studies[29] – and childhood sunburn (7.61%) (Fig. 3a, Supplementary Table S4).

For MI, a complex disease influenced by multiple risk factors[30], we found that the top components were attributed to genetics of lipid metabolism (PC22, high-cholesterol, statin intake, and *APOC1*), alcohol intake (PC100), and sleep duration and food intake (PC83, 25.2%) that collectively corresponded to 36% of the genetic associations (Fig. 3a, Supplementary Fig. S3b, S4-S5, Supplementary Table S4).

Cholelithiasis is a disease involving the presence of gallstones, which are concretions that form in the biliary tract, usually in the gallbladder[31]. We found that the top components contributing to gallstones corresponded to associations with fresh fruit (PC72) and water intake (PC64), as well as bioelectrical impedance of whole body (PC67) corresponding to 51% of genetic associations altogether (Fig. 3a, Supplementary Fig. S3c, S4, S6, Supplementary Table S4).

Biological characterization of DeGAs components

To provide biological characterization of the key components, we applied the genomic region enrichment analysis tool (GREAT)[20] to dissect the biological relevance of the identified components with the both coding and non-coding variants. Given the coverage of the manually curated knowledge of mammalian phenotypes, we focused on the mouse genome informatics (MGI) phenotype ontology[32]. For each key component, we applied GREAT and found an enrichment for the mouse phenotypes consistent with the phenotypic description of our diseases of interest[20]. The top component for BMI, identified as the body fat measures component (PC2), showed enrichment of several anthropometric terms including abnormally short feet (brachypodia) (MP:0002772, binomial fold = 9.04, p = 1.3×10^{-23}), increased birth weight (MP:0009673, fold = 6.21, p = 1.3×10^{-11}), and increased body length (MP:0001257, binomial fold = 3.01, p = 1.3×10^{-36}) (Fig. 3B, Supplementary Table S5). For MI, we found enrichment of cardiac terms, such as artery occlusion (PC22, MP:0006134, fold = 15.86, p = 1.14×10^{-25}) and aortitis (PC22, MP:0010139, aorta inflammation, fold = 9.36, p = 3.41×10^{-31}) (Supplementary Fig. 7, Supplementary Table S6). Similarly, for gallstones, the top enrichment was for abnormal circulating phytosterol level (PC72, MP:0010075, fold = 11.54, p = 5.51×10^{-11}), which is known to be involved in gallstone development[33] (Supplementary Fig. 8, Supplementary Table S7).

Protein truncating variants

Predicted PTVs are a special class of genetic variants with possibly strong effects on gene function[9,12,21,34]. More importantly, strong effect PTV-trait associations can uncover promising drug targets, especially when the direction of effect is consistent with protection of human disease. Given the challenges with interpreting genetic associations across thousands of possibly correlated phenotypes, we applied DeGAs to PTV gene-phenotype associations. We identified PC1 and PC3 as the top two key components for BMI, with 28% and 12% of phenotype squared contribution scores, respectively (Supplementary Fig. S9). The major drivers of PC1 were weight-related measurements, including left and right leg fat-free mass (5.0% and 3.7% of phenotype contribution score for PC1, respectively), left and right leg predicted mass (4.9% each), weight (4.6%), and basal metabolic rate (4.6%), whereas the drivers of PC3 included standing height (13.7%), sitting height (8.1%), and high reticulocyte percentage (6.4%) (Fig. 4a, Supplementary Table S4). Top contributing PTVs to PC1 included variants in *PDE3B* (19.0%), *GPR151* (12.3%), and *ABTB1* (8.5%), whereas PC3 was driven by PTVs on *TMEM91* (8.6%), *EML2-AS1* (6.7%), and *KIAA0586* (6.0%) (Fig. 4b, Supplementary Table S4).

Based on stop-gain variants in *GPR151* (rs114285050) and *PDE3B* (rs150090666) being key contributors to the top two components of genetic associations for PTVs and BMI (Fig. 4c), we proceeded to detailed phenome-wide association analysis (PheWAS) assessing associations of these PTVs with anthropometric phenotypes. PheWAS analysis of these variants confirmed strong associations with obesity-related phenotypes including waist circumference (*GPR151,* marginal association beta = -0.065, p = 2.5×10^{-8}), whole-body fat mass (*GPR151,* beta = -0.069, p = 1.4×10^{-7}), trunk fat mass (*GPR151,* beta = -0.071, p = 1.5×10^{-7}), hip circumference (*PDE3B,* beta = 0.248, p = 1.8×10^{-11}), right leg fat-free mass (*PDE3B,* beta = 0.129, p = 4.2×10^{-8}) and body weight (*PDE3B,* beta = 0.177, p = 4.6×10^{-8}) (Fig. 4d, Supplementary Fig. S10, Supplementary Table S8-9). Among 337,199 White British individuals, we found 7,560 heterozygous and 36 homozygous carriers of the *GPR151* variant and 947 heterozygous carriers of *PDE3B* variants. To assess the effect of the PTVs on BMI, a commonly-used measure of obesity, we performed univariate linear regression analysis with age, sex, and the first four genetic PCs as covariates and found that heterozygous and carriers of *GPR151* PTVs showed 0.324 kg/m^2 lower BMI than the average UK Biobank participant (p = 4.13×10^{-7}). We did not find evidence of association with homozygous carriers (N = 28; p = 0.665), presumably due to lack of power (Supplementary Fig. S11). Heterozygous carriers of *PDE3B* PTVs showed 0.647 kg/m^2 higher BMI (p = 2.09×10^{-4}) than the average UK Biobank participant (Supplementary Fig. S12).

Functional experiments for candidate genes in cellular models of adipocytes

We sought to illustrate the potential application of DeGAs in prioritizing therapeutic targets using functional follow-up experiments. Several of our most interesting findings were observed for obesity-related traits, including the top two candidate genes (*PDE3B* and *GPR151*) contributing to PC1 – the leading component associated with obesity. For this reason, we chose to study these two genes in relation to adipocyte biology. Specifically, the expression and function of

PDE3B and *GPR151* were evaluated in mouse 3T3-L1 and human Simpson-Golabi-Behmel Syndrome (SGBS) cells, two well-established preadipocyte models used for studying adipocyte differentiation (i.e. adipogenesis) and function[35,36].

First, we demonstrated that both genes were expressed in preadipocytes, but showed different expression patterns when cells were transforming into mature adipocytes: *PDE3B* increased dramatically during both mouse and human adipogenesis, while *GPR151* maintained a low expression level throughout the differentiation (Fig. 5a-b). Next, to explore the causal relationships between gene expression and adipogenesis, we introduced short interfering RNA (siRNA) against *Pde3b* and *Gpr151*, respectively, into 3T3-L1 preadipocytes and monitored the impact of gene knockdown on conversion of preadipocytes to adipocytes. Knockdown of *Gpr151* (Fig. 5c) drastically impaired adipocyte differentiation, as evidenced by lowered expression of adipogenesis markers (*Pparg, Cebpa and Fabp4*) (Fig. 5d), as well as the reduced formation of lipid-containing adipocytes (Fig. 5e-f). Further, to test the functional capacity of the fat cells lacking *Gpr151*, we performed a lipolysis assay - an essential metabolic pathway of adipocytes and thus, a key indicator of adipocyte function - on mature adipocytes derived from preadipocytes transfected with either scrambled siRNA (scRNA) or si*Gpr151*. Not surprisingly, *Gpr151*-deficient lipid-poor adipocytes showed dramatically lower lipolysis, along with impaired capability of responding to isoproterenol (ISO), a β-adrenergic stimulus of lipolysis (Fig. 5g). These data suggest that *GPR151* knockdown in adipocyte progenitor cells may block their conversion into mature adipocytes; thus, preventing the expansion of adipose tissue. These results are directionally consistent with our DeGAs and univariate regression analysis showing that *GPR151* PTVs are associated with lower obesity and fat mass, especially central obesity (e.g. waist circumference and trunk fat mass) (Fig. 4d).

To further analyze the functional impact of GPR151 in adipocytes, we generated an overexpression model of GPR151 by infecting 3T3-L1 preadipocytes with virus expressing Flag-tagged human *GPR151* driven by either EF1α or aP2 promotor (Supplementary Fig. S13a). Overexpression of *GPR151* by both constructs were confirmed at the gene and protein levels (Supplementary Fig. S13b-d). However, despite the substantial effect of *Gpr151* knockdown on adipogenesis (Fig. 5), overexpression of *GPR151* in preadipocytes failed to influence adipocyte differentiation significantly, as shown by similar levels of adipogenic markers compared to the non-infected controls (Supplementary Fig. S13e-f). To eliminate the potential masking effects of any unperturbed cells in the partially infected cell population, we specifically selected *GPR151*-overexpressing cells by staining Flag-*GPR151* positive cells with APC-conjugated flag antibody (Supplementary Fig. S13g-h) and sorted APC+ and APC- cells from the differentiating adipocyte cultures. In both EF1α- and aP2-driven *GPR151* overexpression models, *GPR151* mRNA levels were enriched in APC+ cells compared to APC- cells. However, APC+ cells expressed genes characteristics of differentiating adipocytes in a similar level to that of APC- cells (Supplementary Fig. S13i-j). These data conclude that overexpression of GPR151 in preadipocytes cannot further enhance adipogenesis, suggesting that the endogenous level of GPR151 in preadipocytes may be sufficient to maintain the normal differentiation potential of preadipocytes. Although GPR151 is predominantly expressed in the brain, especially in hypothalamic neurons that control appetite and energy expenditure[37], we identified for the first time that the GPR151 protein is present in both subcutaneous and visceral adipose tissue from mice (SAT and VAT), albeit in a very low level (Supplementary Fig. S13k). Together with our gain- and loss-of-function studies of *GPR151* in preadipocyte models, we infer that the regulatory role of *GPR151* in body weight may involve both central and peripheral effects. The minimal but indispensable presence of GPR151 in adipose progenitor cells in generating lipid-rich adipocytes seems to represent an important mechanism by which GPR151 promotes obesity.

In contrast to *GPR151*, knockdown of *Pde3b* in 3T3-L1 preadipocytes (Supplementary Fig. S14a) showed no significant influence on adipogenesis and lipolysis (under either basal or β-adrenergic stimulated conditions), as compared to scRNA-transfected controls (Supplementary Fig. S14b-e). Since PDE3B is expressed primarily in differentiated adipocytes (Fig. 5a-b), future research efforts should be concentrated on studying the metabolic role of PDE3B in mature adipocytes. As an essential enzyme that hydrolyzes both cAMP and cGMP, PDE3B is known to be highly expressed in tissues that are important in regulating energy homeostasis, including adipose tissue[38]. *Pde3b* whole-body knockout in mice reduces the visceral fat mass[39] and confers cardioprotective effects[40]. There is a growing body of evidence that cardiometabolic health is linked to improved body fat distribution (i.e. lower visceral fat, higher subcutaneous fat) in a consistent direction[41]. Our PheWAS analysis suggests that *PDE3B* PTVs have the strongest association with subcutaneous and lower-body adiposity (e.g. hip and leg fat mass) (Supplementary Fig. S10). Therefore, understanding the fat depot-specific metabolic effects of PDE3B may help uncover the mechanism underlying the positive relationship of *PDE3B* PTVs with peripheral fat accumulation and favorable metabolic profiles.

Discussion

We developed DeGAs, an application of TSVD, to decompose genome-and phenome-wide summary statistic matrix from association studies of thousands of phenotypes for systematic characterization of latent components of genetic associations. Applying DeGAs, we identified key latent components characterized with disease outcomes, risk factors, comorbidity structures, and environmental factors, with corresponding sets of genes and variants, providing insights on their context specific functions. With additional biological characterization of latent components using GREAT, we find enrichment of relevant phenotypes in mouse phenotype ontology. This replication across species highlights the ability of DeGAs to capture functionally relevant sets of both coding and non-coding variants in each component.

Given that DeGAs is applied on summary statistics and does not require individual level data, there is substantial potential to dissect genetic components of the human phenome when applied to data from population-based biobanks around the globe[14,42–45]. As a proof of concept, we report novel potential therapeutic targets against obesity or its complications based on combination of quantitative results from DeGAs, phenome-wide analyses in the UK Biobank, and functional studies in adipocytes.

Taken together, we highlight the directional concordance of our experimental data with the quantitative results from DeGAs and PTV-phenotype associations: *GPR151* inhibition may reduce total body and central fat, while deletion of *PDE3B* may favor subcutaneous, rather than visceral, fat deposition; both are expected to have beneficial effects on cardiometabolic health. Although these two genes were recently reported to be associated with obesity in another recent study based on the UK Biobank[46], we are the first to experimentally identify *GPR151* as a promising therapeutic target to treat obesity, partly due to its requisite role in regulating adipogenesis. We also suggest *PDE3B* as a potential target of adipocyte-directed therapy. In this study, we focused on evaluating the functional effects of these genes on adipocyte function and development. We do not exclude the contribution nor the importance of other tissues or mechanisms underlying body weight changes. Indeed, some lines of evidence support additional effects of *GPR151* on obesity via the central nervous system – possibly on appetite regulation[37].

The resource made available with this study, including the DeGAs app, an interactive web application in the Global Biobank Engine[47], provides a starting point to investigate genetic components, their functional relevance, and new therapeutic targets. These results highlight the benefit of comprehensive phenotyping on a population and suggest that systematic characterization and analysis of genetic associations across the human phenome will be an important part of efforts to understand biology and develop novel therapeutic approaches.

Methods

Study population

The UK Biobank is a population-based cohort study collected from multiple sites across the United Kingdom. Information on genotyping and quality control has previously been described[14]. In brief, study participants were genotyped using two similar arrays (Applied Biosystems UK BiLEVE Axiom Array (807,411 markers) and the UK Biobank Axiom Array (825,927 markers)), which were designed for the UK Biobank study. The initial quality control was performed by the UK Biobank analysis team and designed to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches, using two similar novel arrays[14].

Genotype data preparation

We used genotype data from the UK Biobank dataset release version 2[14] and the hg19 human genome reference for all analyses in the study. To minimize the variabilities due to population structure in our dataset, we restricted our analyses to include 337,199 White British individuals based on the following five criteria reported by the UK Biobank in the file "ukb_sqc_v2.txt":

1. self- reported white British ancestry ("in_white_British_ancestry_subset" column)
2. used to compute principal components ("used_in_pca_calculation" column)
3. not marked as outliers for heterozygosity and missing rates ("het_missing_outliers" column)
4. do not show putative sex chromosome aneuploidy ("putative_sex_chromosome_aneuploidy" column)
5. have at most 10 putative third-degree relatives ("excess_relatives" column).

We annotated variants using the VEP LOFTEE plugin (https://github.com/konradjk/loftee) and variant quality control by comparing allele frequencies in the UK Biobank and gnomAD (gnomad.exomes.r2.0.1.sites.vcf.gz) as previously described[12].

We focused on variants outside of major histocompatibility complex (MHC) region (chr6:25477797-36448354) and performed LD pruning using PLINK with "--indep 50 5 2". Furthermore, we selected variants according to the following rules:

- Missingness of the variant is less than 1%.
- Minor-allele frequency is greater than 0.01%.
- The variant is in the LD-pruned set.
- Hardy-Weinberg disequilibrium test p-value is greater than 1.0×10^{-7}.
- Manual cluster plot inspection. We investigated cluster plots for subset of our variants and removed 11 variants that has unreliable genotype calls as previously described[12].
- Passed the comparison of minor allele frequency with gnomAD dataset as previously described[12].

These variant filters are summarized in Supplementary Fig. S1.

Phenotype data preparation

We organized 2,138 phenotypes from the UK Biobank in 11 distinct groups (Supplementary Table 1). We included phenotypes with at least 100 cases for binary phenotypes and 100 individuals with non-missing values for quantitative phenotypes. For disease outcome phenotypes, cancer, and family history, we used the same definitions as previously described[12]. We used specific data fields and data category from the UK Biobank to define the phenotypes in the following categories as well as 19 and 42 additional miscellaneous binary and quantitative phenotypes: medication, imaging, physical measurements, assays, and binary and quantitative questionnaire (Supplementary Table 1-2).

Some phenotype information from the UK Biobank contains three instances, each of which corresponds to (1) the initial assessment visit (2006-2010), (2) first repeat assessment visit (2012-2013), and (3) imaging visit (2014-). For binary phenotype, we defined "case" if the participants are classified as case in at least one of their visits and "control" otherwise. For quantitative phenotype, we took a median of non-NA values. In total, we defined 1,196 binary phenotypes and 943 quantitative phenotypes.

Genome-wide association analyses of 2,138 phenotypes

Association analyses for single variants were applied to the 2,138 phenotypes separately. For binary phenotypes, we performed Firth-fallback logistic regression using PLINK v2.00a (17 July 2017) as previously described[12,48]. For quantitative phenotypes, we applied generalized linear model association analysis with PLINK v2.00a (20 Sep. 2017). We applied quantile normalization for phenotype (--pheno-quantile-normalize option), where we fit a linear model with covariates and transform the phenotypes to normal distribution $N(0, 1)$ while preserving the original rank. We used the following covariates in our analysis: age, sex, types of genotyping array, and the first four genotype principal components computed from the UK Biobank.

Summary statistic matrix construction and variant filters

We constructed two Z-score summary statistic matrices. Each element of the matrix corresponds to summary statistic for a particular pair of a phenotype and a variant. We imposed different sets of variant filters.

- Variant quality control filter: Our quality control filter described in the previous section on genotype data preparation.
- Non-MHC variant filter: All variants outside of major histocompatibility complex region. With this filter, variants in chr6:25477797-36448354 were excluded from the summary statistic matrix.
- PTVs-only: With this filter, we subset to include only the variants having the VEP LOFTEE predicted consequence of: stop gain, frameshift, splice acceptor, or splice donor.

By combining these filters, we defined the following sets of variants

- All-non-MHC: This is a combination of our variant QC filter and non-MHC filter.
- PTVs-non-MHC: This is a combination of our variant QC filter, non-MHC filter, and PTVs filter.

In addition to phenotype quality control and variant filters, we introduced value-based filters based on statistical significance to construct summary statistic matrices only with confident values. We applied the following criteria for the value filter:

- P-value of marginal association is less than 0.001.
- Standard error of beta value or log odds ratio is less than 0.08 for quantitative phenotypes and 0.2 for binary phenotypes.

With these filters, we obtained the following two matrices:

- All-non-MHC dataset that contains 2,138 phenotypes and 235,907 variants. We label this dataset as **"all" dataset**.
- "PTVs-non-MHC" dataset that contains 628 phenotypes and 784 variants. We label this dataset as **"PTVs only" dataset**. This contains a fewer number of phenotypes because not all the phenotypes have statistically significant associations with PTVs.

The effects of variant filters are summarized in Fig. S1. Finally, we transformed the summary statistics to Z-scores so that each vector that corresponds to a particular phenotype has zero mean with unit variance.

Truncated singular value decomposition of the summary statistic matrix

For each summary statistic matrix, we applied truncated singular value decomposition (TSVD). The matrix, which we denote as W, of size $N \times M$, where N denotes the number of phenotypes and M denotes the number of variants, is the input data. With TSVD, W is factorized into a product of three matrices: U, S, and V^T: $W = USV^T$, where $U = (u_{i,k})_{i,k}$ is an orthonormal matrix of size $N \times K$ whose columns are phenotype (left) singular vectors, S is a diagonal matrix of size $K \times K$ whose elements are singular values, and $V = (v_{j,k})_{j,k}$ is an orthonormal matrix of size $M \times K$ whose columns are variant (right) singular vectors. While singular values in S

represent the magnitude of the components, singular vectors in U and V summarizes the strength of association between phenotype and component and variant and component, respectively. With this decomposition, the k-th latent component (principal component, PC k) are represented as a product of k-th column of U, k-th diagonal element in S, and k-th row of V^T. We used implicitly restarted lanczos bidiagonalization algorithm (IRLBA)[49] (https://github.com/bwlewis/irlba) implemented on SciDB[50] to compute the first K components in this decomposition.

Relative variance explained by each of the components

A scree plot (Fig. S1) quantify the variance explained by each component: variance explained by k-th component = ${s_k}^2/\,\mathrm{Var}_{\mathrm{Tot}}(W)$ where, s_k is the k-th diagonal element in the diagonal matrix S and $\mathrm{Var}_{\mathrm{Tot}}(W)$ is the total variance of the original matrix before DeGAs is applied.

Selection of number of latent components in TSVD

In order to apply TSVD to the input matrix, the number of components should be specified. We apply $K = 100$ for our analysis for both datasets. We computed the expected value of squared eigenvalues under the null model where the distribution of variance explained scores across the full-ranks are uniform. This can be computed with the rank of the original matrix, which is equal to the number of phenotypes in our datasets:

$$E[\text{Variance explained by } k\text{-th component under the null}] = {}^{1}\!/_{(\mathrm{Rank}(W)^2)}$$

For all of the two datasets, we found that that of 100-th component is greater than the expectation. This indicates even the 100-th components are informative to represent the variance of the original matrix. In the interest of computational efficiency, we set $K = 100$.

Factor scores

From these decomposed matrices, we computed **factor score** matrices for both phenotypes and variants as the product of singular vector matrix and singular values. We denote the one for phenotypes as $F_p = ({f_{i,j}}^p)_{i,j}$ the one for variants as $F_v = ({f_{i,j}}^v)_{i,j}$ and defined as follows:

$$F_p = US$$

$$F_v = VS$$

Since these factor scores are mathematically the same as principal components in principal component analysis (PCA), one can investigate the contribution of the phenotypes or variants for specific principal components by simply plotting factor scores[23] (Fig. 2a-b). Specifically, phenotype factor score is the same as phenotype principal components and variant factor score is the same as variant principal components. By normalizing these factor scores, one can compute contribution scores and cosine scores to quantify the importance of phenotypes, variants, and principal components as described below.

Scatter plot visualization with biplot annotations

To investigate the relationship between phenotype and variants in the TSVD eigenspace, we used a variant of biplot visualization[51,52]. Specifically, we display phenotypes projected on phenotype principal components ($F_p = US$) as a scatter plot. We also show variants projected on variant principal components ($F_v = VS$) as a separate scatter plot and added phenotype singular vectors (U) as arrows on the plot using sub-axes (Fig. 2b, 4c, S5-6). In scatter plot with biplot annotation, the inner product of a genetic variant and a phenotype represents the direction and the strength of the projection of the genetic association of the variant-phenotype pair on the displayed latent components. For example, when a variant and a phenotype share the same direction on the annotated scatter plot, that means the projection of the genetic associations of the variant-phenotype pair on the displayed latent components is positive. When a variant-phenotype pair is projected on the same line, but on the opposite direction, the projection of the genetic associations on the shown latent components is negative. When the variant and phenotype vectors are orthogonal or one of the vectors are of zero length, the projection of the genetic associations of the variant-phenotype pair on the displayed latent components is zero. Given the high dimensionality of the input summary statistic matrix, we selected relevant phenotypes to display to help interpretation of genetic associations in the context of these traits.

Contribution scores

To quantify the contribution of the phenotypes, variants, and genes to a given component, we compute **contribution scores**. We first define **phenotype contribution score** and **variant contribution score**. We denote phenotype contribution score and variant contribution score for some component k as $\mathrm{cntr}_k^{\mathrm{phe}}(i)$ and $\mathrm{cntr}_k^{\mathrm{var}}(j)$, respectively. They are defined by squaring the left and right singular vectors and normalizing them by Euclidian norm across phenotypes and variants:

$$\mathrm{cntr}_k^{\mathrm{phe}}(i) = \left(u_{i,k}\right)^2$$

$$\mathrm{cntr}_k^{\mathrm{var}}(j) = \left(v_{i,k}\right)^2$$

where, i and j denote indices for phenotype and variant, respectively. Because U and V are orthonormal, the sum of phenotype and variant contribution scores for a given component are guaranteed to be one, i.e. $\sum_i \mathrm{cntr}_k^{\mathrm{phe}}(i) = \sum_j \mathrm{cntr}_k^{\mathrm{var}}(j) = 1$.

Based on the variant contribution scores for the k-th component, we define the **gene contribution score** for some component k as the sum of variant contribution scores for the set of variants in the gene:

$$\mathrm{cntr}_k^{\mathrm{gene}}(g) = \sum_{j \in g} \mathrm{cntr}_k^{\mathrm{var}}(j)$$

where, g denotes indices for the set of variants in gene g. To guarantee that gene contribution scores for a given component sum up to one, we treat the variant contribution score for the non-coding variants as gene contribution scores. When multiple genes, g_1, g_2,..., g_n are sharing the same variants, we defined the gene contribution score for the union of multiple genes rather than each gene:

$$\mathrm{cntr}_k^{\mathrm{gene}}(\{g_i \mid i \in [1, n]\}) = \sum_{\{j \mid j \in g_1 \wedge j \in g_2 \wedge \cdots \wedge j \in g_n\}} \mathrm{cntr}_k^{\mathrm{var}}(j)$$

With these contribution score for a given component, it is possible to quantify the relative importance of a phenotype, variant, or gene to the component. Since DeGAs identifies latent components using unsupervised learning, we interpret each component in terms of the driving phenotypes, variants, and genes, i.e. the ones with large contribution scores for the component.

The top 20 driving phenotypes, variants, and genes (based on contribution scores) for the top five TSVD components and the top three key components for our phenotypes of interest are summarized in Supplementary Table S3.

We used stacked bar plots for visualization of the contribution score profile for each of the components. We represent phenotypes, genes, or variants with large contribution scores as colored segments and aggregated contributions from the remaining ones as “others” in the plot (Fig. 2c-d, 3a, 4a-b, Supplementary Fig. S4). To help interpretation of the major contributing factors for the key components, we grouped phenotypes into categories, such as “fat”, “fat-free” phenotypes, and showed the sum of contribution scores for the phenotype groups. The list of phenotype groups used in the visualization is summarized in Supplementary Table S3.

Squared cosine scores

Conversely, we can also quantify the relative importance of the latent components for a given phenotype or variant with **squared cosine scores**. We denote phenotype squared cosine score for some phenotype i and variant squared cosine score for some variant j as $\cos^2{}_i^{\mathrm{phe}}(k)$ and $\cos^2{}_j^{\mathrm{var}}(k)$, respectively. They are defined by squaring of the factor scores and normalizing them by Euclidian norm across components:

$$\cos^2{}_i^{\mathrm{phe}}(k) = {\left(f_{i,k}{}^{p}\right)^2} \Big/ {\sum_{k'} \left(f_{i,k'}{}^{p}\right)^2}$$

$$\cos^2{}_j^{\mathrm{var}}(k) = {\left(f_{j,k}{}^{v}\right)^2} \Big/ {\sum_{k'} \left(f_{j,k'}{}^{v}\right)^2}$$

By definition, the sum of squared cosine scores across a latent component for a given phenotype or variant equals to one, i.e. $\sum_k \cos^2{}_i^{\mathrm{phe}}(k) = \sum_k \cos^2{}_j^{\mathrm{var}}(k) = 1$. While singular values in the diagonal matrix S quantify the importance of latent components for the global latent structure, the phenotype or variant squared cosine score quantifies the relative importance of each component in the context of a given phenotype or a variant. The squared cosine scores for the phenotypes highlighted in the study is summarized in Fig. S3 and Supplementary Fig. S9.

Note that squared cosine scores and contribution scores are two complementary scoring metrics to quantify the relationship among phenotypes, components, variants, and genes. It does not necessarily have inverse mapping property. For example, it is possible to see a situation, where for a given phenotype p, phenotype squared cosine score identifies k as the top key component, but phenotype contribution score for k identifies p' ($p' \neq p$) as the top driving phenotype for the component k. This is because the two scores, contribution score and squared cosine score, are both defined by normalizing singular vector and principal component vector matrices, respectively, but with respect to different slices: one for row and the other for column.

Genomic region enrichment analysis with GREAT

We applied the genomic region enrichment analysis tool (GREAT version 4.0.3) to each DeGAs components[20]. We used the mouse genome informatics (MGI) phenotype ontology, which contains manually curated knowledge about hierarchical structure of phenotypes and genotype-phenotype mapping of mouse[32]. We downloaded their ontologies on 2017-09-28 and mapped MGI gene identifiers to Ensembl human gene ID through unambiguous one-to-one homology mapping between human and mouse Ensembl IDs. We removed ontology terms that were labelled as “obsolete”, “bad”, or “unknown” from our analysis. As a result, we obtained 709,451 mapping annotation spanning between 9,554 human genes and 9,592 mouse phenotypes.

For each DeGAs component, we selected the top 5,000 variants according to their variant contribution score and performed enrichment analysis with the default parameter as described elsewhere[20]. Since we included the non-coding variants in the analysis, we focused on GREAT binomial genomic region enrichment analysis based on the size of regulatory domain of genes, and quantified the significance of enrichment in terms of binomial fold enrichment and binomial p-value. Given that we have 9,561 terms in the ontology, we set a Bonferroni p-value threshold of 5×10^{-6}. To illustrate the results of the genomic region enrichment analysis for the phenotypes of our interest, we made circular bar plots using the R package ggplot2, where each of the key components are displayed in the innermost track with their phenotype squared cosine score to be proportional to their angle, and the resulted significant ontology terms are represented as the bars. The binomial fold change is represented as the radius and the binomial p-value is represented as color gradient in a log scale in the plot (Fig. 3b, Supplementary Fig. S7-8, Supplementary Table S5-7).

Quality control of variant calling with intensity plots

To investigate the quality of variant calling for the two PTVs highlighted in the study, we manually inspected intensity plots. These plots are available on Global Biobank Engine.

- https://biobankengine.stanford.edu/intensity/rs114285050
- https://biobankengine.stanford.edu/intensity/rs150090666

Phenome-wide association analysis

To explore the functional roles of the two PTVs across thousands of potentially correlated phenotypes, we performed a phenome-wide association study (PheWAS). We report the statistically significant ($p < 0.001$) associations with phenotypes with at least 1,000 case count (binary phenotypes) or 1,000 individuals with measurements with non-missing values (quantitative phenotypes) (Fig. 3d, Supplementary Fig. S10). The results of this PheWAS are also available as interactive plots as a part of Global Biobank Engine.

- https://biobankengine.stanford.edu/variant/5-145895394
- https://biobankengine.stanford.edu/variant/11-14865399

Univariate regression analysis for the identified PTVs

To quantify the effects of the two PTVs on obesity, we performed univariate regression analysis. We extracted individual-level genotype information for the two PTVs with the PLINK2 pgen Python API (http://www.cog-genomics.org/plink/2.0/)[48]. After removing individuals with missing values for BMI and genotype, we performed linear regression for BMI (http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21001) with age, sex, and the first four genomic PCs as covariates:

$$\text{BMI} \sim 0 + \text{age} + \text{as.factor(sex)} + \text{PC1} + \text{PC2} + \text{PC3} + \text{PC4} + \text{as.factor(PTV)}$$

where, PC1-4 denotes the first four components of genomic principal components, PTV ranges in 0, 1, or 2 and it indicates the number of minor alleles that the individuals have.

Mouse 3T3-L1 cell culture and differentiation

3T3-L1 preadipocytes were cultured in Dulbecco's Modified Eagle's Medium (DMEM) containing 10% fetal bovine serum (FBS) and antibiotics (100 U/mL of penicillin G and 100 μg/mL of streptomycin) at 37°C in a humidified atmosphere containing 5% CO_2. To obtain fully differentiated adipocytes, 3T3-L1 preadipocytes were grown into 2-day post-confluence, and then differentiation was induced by using a standard differentiation cocktail containing 0.5 mM of IBMX, 1 μm of dexamethasone, 1 μg/mL of insulin, and 10% FBS. After 48 h, medium was changed into DMEM supplemented with 10% FBS and 1 μg/mL of insulin and replenished every 48 h for an additional 6 days.

Human SGBS cell culture and differentiation

SGBS cells were cultured in DMEM/F12 containing 33 μM biotin, 17 μM pantothenate, 0.1 mg/mg streptomycin and 100 U/mL penicillin (0F medium) supplemented with 10% FBS in a 5% CO_2 incubator. To initiate differentiation, confluent cells were stimulated by 0F media supplemented with 0.01 mg/mL human transferrin, 0.2 nM T3, 100 nM cortisol, 20 nM insulin, 250 μM IBMX, 25 nM dexamethasone and 2 μM rosiglitazone. After day 4, the differentiating cells were kept in 0F media supplemented with 0.01 mg/mL human transferrin, 100 nM cortisol, 20 nM insulin and 0.2 nM T3 for additional 8-10 days until cells were fully differentiated.

siRNA knockdown in 3T3-L1 preadipocytes

At 80% confluence, 3T3-L1 preadipocytes were transfected with 50 nM siRNA against *Gpr151* (Origene #SR412988), *Pde3b* (Origene #SR422062), or scrambled negative control (Origene #SR30004) using Lipofectamine™ RNAiMAX Transfection Reagent (Invitrogen) following the manufacturer's protocol. The transfected cells were incubated for 48 h and then subjected to differentiation.

Reverse transcription (RT) and qPCR analysis

Total RNA was extracted using TRIzol reagent (Invitrogen), following the manufacturer's instruction. RNA was converted to cDNA using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Quantitative PCR reactions were prepared with TaqMan™ Fast Advanced Master Mix (Thermo Fisher Scientific) and performed on ViiA 7 Real-Time PCR System (Thermo Fisher Scientific). All data were normalized to the content of Cyclophilin A (PPIA), as the endogenous control. TaqMan primer information for RT-qPCR is listed below: *GPR151* (Hs00972208_s1), *Gpr151* (Mm00808987_s1), *PDE3B* (Hs00265322_m1), *Pde3b* (Mm00691635_m1), *Pparg* (Mm00440940_m1), *Cebpa* (Mm00514283_s1), *Fabp4* (Mm00445878_m1), *PPIA* (Hs04194521_s1), *Ppia* (Mm02342430_g1).

Oil Red O staining and quantification

Cells were washed twice with PBS and fixed with 10% formalin for 1 h at room temperature. Cells were then washed with 60% isopropanol and stained for 15 min with a filtered Oil Red O solution (mix six parts of 0.35% Oil Red O in isopropanol with four parts of water). After washing with PBS 4 times, cells were maintained in PBS and visualized by inverted microscope. After taking pictures, Oil Red O stain was extracted with 100% isopropanol and the absorbance was measured at 492 nm by a multi-well spectrophotometer (Bio-Rad).

Lipolysis assay

Glycerol release into the culture medium was used as an index of lipolysis. Fully differentiated 3T3-L1 adipocytes were serum starved overnight and then treated with either vehicle (DMSO) or the lipolytic stimuli isoproterenol (ISO, 10μM) for 3 h. The culture medium was collected and the glycerol content in the culture medium was measured using an adipocyte lipolysis assay kit (ZenBio #LIP-1-NCL1). Glycerol release into the culture medium was normalized to the protein content of the cells from the same plate.

Overexpression of *GPR151* in 3T3-L1 preadipocytes

The *GPR151* construct was obtained from Addgene (#66327). This construct includes a cleavable HA signal to promote membrane localization, a FLAG epitope sequence for cell surface staining followed by codon-optimized human *GPR151* sequence[53]. We PCR-amplified the above sequence with stop codon and assembled it into a lentiviral plasmid (Addgene #85969) with either EF1α promoter (Addgene # 11154) or aP2 promoter (Addgene # 11424). EF1α-*GPR151* or aP2-*GPR151* lentiviral plasmid were transfected into human embryonic kidney 293T cells, together with the viral packaging vectors pCMV-dR8.91 and pMD2-G. 72 h after transfection, virus-containing medium was collected, filtered through a 0.45-μm pore-size syringe filter, and frozen at -80°C. 3T3-L1 preadipocytes at 50% confluence were infected with the lentivirus stocks containing 8 μg/mL polybrene. Two days after transduction, lentivirus-infected 3T3-L1 preadipocytes were subject to differentiation.

Flow cytometry analysis

Day 6 differentiating 3T3-L1 adipocytes were collected and washed with ice cold FACS buffer (PBS containing 2% BSA). Cells were first resuspended into FACS staining buffer (BioLegend # 420201) at ~1M cells/100µl and incubated with anti-mouse CD16/CD32 Fc Block (BioLegend # 101319) at room temperature for 10-15 min. Cells were then incubated with APC-conjugated FLAG antibody (BioLegend # 637307) for 20-30 min at room temperature in the dark. Following washing and centrifugation, cells were resuspended in FACS buffer and sorted using a BD Influx^{TM} Cell Sorter. Cells without FLAG antibody staining were used to determine background fluorescence levels. Cells were sorted based on APC fluorescence and collected directly into TRIzol reagent for RNA extraction.

Western Blot Analysis

Lysate aliquots containing 50µg of proteins were denatured, separated on a 4-10% SDS-polyacrylamide gel, and transferred to nitrocellulose membranes using a Trans-Blot® SD Semi-Dry Transfer Cell (Bio-Rad). Membranes were blocked in 5% non-fat milk and incubated overnight at 4 °C with primary antibodies: anti-GPR151 (LSBio # LS-B6760-50) or anti-beta-actin (Cell Signaling #3700). Subsequently, the membranes were incubated for 1 h at room temperature with IRDye® 800CW goat-anti-mouse antibody (LI-COR #926-32210). Target proteins were visualized using Odyssey® Fc Imaging System (LI-COR).

Statistical analysis of functional data

Data are expressed as mean ± SEM. Student’s t test was used for single variables, and one-way ANOVA with Bonferroni post hoc correction was used for multiple comparisons using GraphPad Prism 7 software.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 24983. We thank all the participants in the UK Biobank study. Y.T. is supported by Funai Overseas Scholarship from Funai Foundation for Information Technology and the Stanford University School of Medicine. J.M.J. was funded by grant NNF17OC0025806 from the Novo Nordisk Foundation and the Stanford Bio-X Program. M.A.R. and C.D. are supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). C.D. is supported by a postdoctoral fellowship from the Stanford Center for Computational, Evolutionary, and Human Genomics and the Stanford ChEM-H Institute. The primary and processed data used to generate the analyses presented here are available in the UK Biobank access management system (https://amsportal.ukbiobank.ac.uk/) for application 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf), and the results are displayed in the Global Biobank Engine (https://biobankengine.stanford.edu). We obtained clip-arts for Fig. 1b from Irasutoya (https://www.irasutoya.com/) by following their terms and conditions. The copyright of the original clip-arts belongs to Mr. Takashi Mifune. We would like to thank the Customer Solutions Team from Paradigm4 who helped us implement efficient databases for queries and application of inference methods to the data, and also implemented optimized versions of truncated singular value decomposition.

Funding:

This work was supported by National Human Genome Research Institute (NHGRI) and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the National Institutes of Health (NIH) under awards R01HG010140 and R01DK106236. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author information

Author contributions

M.A.R. and E.I. conceived and designed the study. Y.T. and M.A.R. carried out the statistical and computational analyses with advice from J.M.J., H.H., M.A., C.D., B.N., K.L, T.H., G.B., and E.I. J.L., C.Y.P., and E.I. carried out the functional experiments. Y.T., M.A., and C.D. carried out quality control of the data. C.C. optimized and implemented computational methods. Y.T. and M.A.R. developed the DeGAs app in Global Biobank Engine. M.A.R. supervised computational and statistical aspects of the study. E.I. supervised experimental aspects of the study. The manuscript was written by Y.T., J.L., J.M.J., E.I., and M.A.R; and revised by all the co-authors. All co-authors have approved of the final version of the manuscript.

Competing financial interests

None.

Data availability:

Data is displayed in the Global Biobank Engine (https://biobankengine.stanford.edu).

References

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** 1001–1006 (2014).

2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).

3. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).

4. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–190 (2014).

5. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).

6. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12,** 581–594 (2013).

7. Waring, R. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *N Engl J Med* 9 (2006).

8. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37,** 161–165 (2005).

9. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43,** 1066–1073 (2011).

10. Rivas, M. A. *et al.* A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. *Nat. Commun.* **7,** 12342 (2016).

11. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. *Science* **324,** 387–389 (2009).

12. DeBoever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9,** 1612 (2018).

13. Tipney, H. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 1–7 (2015).

14. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).

15. Stewart, G. On the Early History of the Singular Value Decomposition. *SIAM Rev.* **35,** 551–566 (1993).

16. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).

17. Varah, J. M. On the Numerical Solution of Ill-Conditioned Linear Systems with Applications to Ill-Posed Problems. *SIAM J. Numer. Anal.* **10,** 257–267 (1973).

18. Hanson, R. J. A Numerical Method for Solving Fredholm Integral Equations of the First Kind Using Singular Values. *SIAM J. Numer. Anal.* **8,** 616–622 (1971).

19. Hansen, P. C. The truncatedSVD as a method for regularization. *BIT Numer. Math.* **27,** 534–553 (1987).

20. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).

21. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348,** 666–669 (2015).

22. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335,** 823–828 (2012).

23. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2,** 433–459 (2010).

24. Bigaard, J. *et al.* Body Fat and Fat-Free Mass and All-Cause Mortality. *Obes. Res.* **12,** 1042–1049 (2004).

25. Foster, K. R. & Lukaski, H. C. Whole-body impedance--what does it measure? *Am. J. Clin. Nutr.* **64,** 388S-396S (1996).

26. Talma, H. *et al.* Bioelectrical impedance analysis to estimate body composition in children and adolescents: a systematic review and evidence appraisal of validity, responsiveness, reliability and measurement error. *Obes. Rev. Off. J. Int. Assoc. Study Obes.* **14,** 895–905 (2013).

27. Moran, A. E. *et al.* The Global Burden of Ischemic Heart Disease in 1990 and 2010. *Circulation* (2014).

28. Lammert, F. *et al.* Gallstones. *Nat. Rev. Dis. Primer* **2,** 16024 (2016).

29. Zhou, D., Wu, J. & Luo, G. Body mass index and risk of non-melanoma skin cancer: cumulative evidence from prospective studies. *Sci. Rep.* **6,** 37691 (2016).

30. Khera, A. V. & Kathiresan, S. Is Coronary Atherosclerosis One Disease or Many?: Setting Realistic Expectations for Precision Medicine. *Circulation* **135,** 1005–1007 (2017).

31. Bennion, L. J. & Grundy, S. M. Risk Factors for the Development of Cholelithiasis in Man. *N. Engl. J. Med.* **299,** 1161–1167 (1978).

32. Smith, C. L. & Eppig, J. T. Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *J. Biomed. Semant.* **6,** 11 (2015).

33. Krawczyk, M. *et al.* Phytosterol and cholesterol precursor levels indicate increased cholesterol excretion and biosynthesis in gallstone disease. *Hepatology* **55,** 1507–1517 (2012).

34. Abul-Husn, N. S. *et al.* A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *N. Engl. J. Med.* **378,** 1096–1106 (2018).

35. Green, H. & Kehinde, O. An established preadipose cell line and its differentiation in culture II. Factors affecting the adipose conversion. *Cell* **5,** 19–27 (1975).

36. Wabitsch, M. *et al.* Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *Int. J. Obes.* **25,** 8–15 (2001).

37. Broms, J. *et al.* Monosynaptic retrograde tracing of neurons expressing the G-protein coupled receptor Gpr151 in the mouse brain. *J. Comp. Neurol.* **525,** 3227–3250 (2017).

38. Shakur, Y. *et al.* Regulation and function of the cyclic nucleotide phosphodiesterase (PDE3) gene family. in *Progress in Nucleic Acid Research and Molecular Biology* **66,** 241–277 (Elsevier, 2000).

39. Chung, Y. W. *et al.* White to beige conversion in PDE3B KO adipose tissue through activation of AMPK signaling and mitochondrial function. *Sci. Rep.* **7,** 40445 (2017).

40. Chung, Y. W. *et al.* Targeted disruption of PDE3B, but not PDE3A, protects murine heart from ischemia/reperfusion injury. *Proc. Natl. Acad. Sci.* 201416230 (2015). doi:10.1073/pnas.1416230112

41. Emdin, C. A. *et al.* Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA* **317,** 626–634 (2017).

42. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

43. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40,** 1652–1666 (2011).

44. Christensen, H., Nielsen, J. S., Sorensen, K. M., Melbye & Brandslund, I. New national Biobank of The Danish Center for Strategic Research on Type 2 Diabetes (DD2). *Clin. Epidemiol.* 37 (2012). doi:10.2147/CLEP.S33042

45. Avlund, K. *et al.* Copenhagen Aging and Midlife Biobank (CAMB): An Introduction. *J. Aging Health* **26,** 5–20 (2014).

46. Emdin, C. A. *et al.* Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. *Nat. Commun.* **9,** 1613 (2018).

47. McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *bioRxiv* 304188 (2018). doi:10.1101/304188

48. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).

49. Baglama, J. & Reichel, L. Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. *SIAM J. Sci. Comput.* **27,** 19–42 (2005).

50. Brown, P. G. Overview of sciDB: large scale array storage, processing and analysis. in *Proceedings of the 2010 international conference on Management of data - SIGMOD '10* 963 (ACM Press, 2010). doi:10.1145/1807167.1807271

51. Gower, J., Lubbe, S. & Roux, N. le. *Understanding Biplots*. (John Wiley & Sons, Ltd, 2011). doi:10.1002/9780470973196

52. Gabriel, K. R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* **58,** 453 (1971).

53. Kroeze, W. K. *et al.* PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome. *Nat. Struct. Mol. Biol.* **22,** 362–369 (2015).

Figures

Figure 1

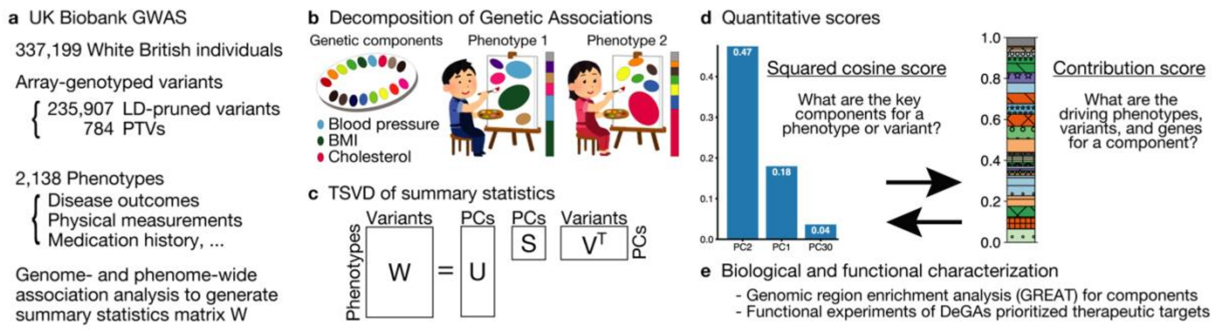

a UK Biobank GWAS
337,199 White British individuals
Array-genotyped variants
235,907 LD-pruned variants
784 PTVs
2,138 Phenotypes
Disease outcomes
Physical measurements
Medication history, ...
Genome- and phenome-wide association analysis to generate summary statistics matrix W
b Decomposition of Genetic Associations
Genetic components
Phenotype 1
Phenotype 2
Blood pressure
BMI
Cholesterol
c TSVD of summary statistics
Variants
PCs
PCs
Variants
Phenotypes
W
=
U
S
V^T
PCs
d Quantitative scores
0.47
0.18
0.04
0.0
0.1
0.2
0.3
0.4
PC2
PC1
PC30
Squared cosine score
What are the key components for a phenotype or variant?
1.0
0.8
0.6
0.4
0.2
0.0
Contribution score
What are the driving phenotypes, variants, and genes for a component?
e Biological and functional characterization
- Genomic region enrichment analysis (GREAT) for components
- Functional experiments of DeGAs prioritized therapeutic targets

Fig.1 Illustrative study overview. **a** Summary of the UK Biobank genotype and phenotype data used in the study. We included White British individuals and analyzed LD-pruned and quality-controlled variants in relation to 2,138 phenotypes with a minimum of 100 individuals as cases (binary phenotypes) or non-missing values (quantitative phenotypes) (Supplementary Table S1-2). **b** Decomposition of Genetic Associations (DeGAs) characterizes latent genetic components, which are represented as different colors on the palette, with an unsupervised learning approach – truncated singular value decomposition (TSVD), followed by identification of the key components for each phenotype of our interest (painting phenotypes with colors) and annotation of each of the components with driving phenotypes, variants, and genes (finding the meanings of colors). **c** TSVD applied to decompose genome- and-phenome-wide summary statistic matrix W to characterize latent components. U, S, and V represent resulting matrices of singular values and vectors. **d** We used the squared cosine score and the contribution score, to quantify compositions and biomedical relevance of latent components. **e** We applied the genomic region enrichment analysis tool (GREAT) for biological characterization of each component and performed functional experiments focusing on adipocyte biology.

Figure 2

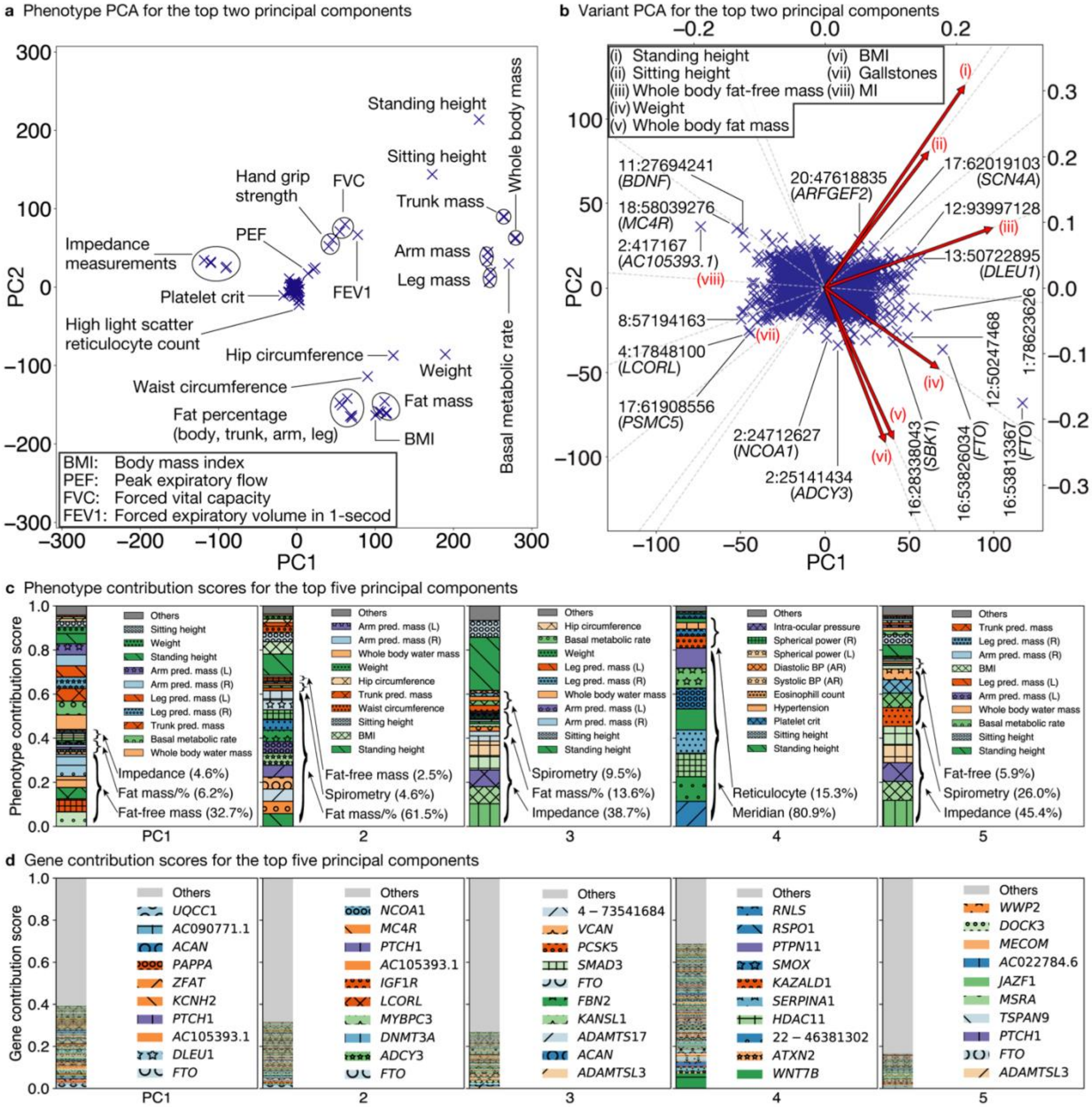
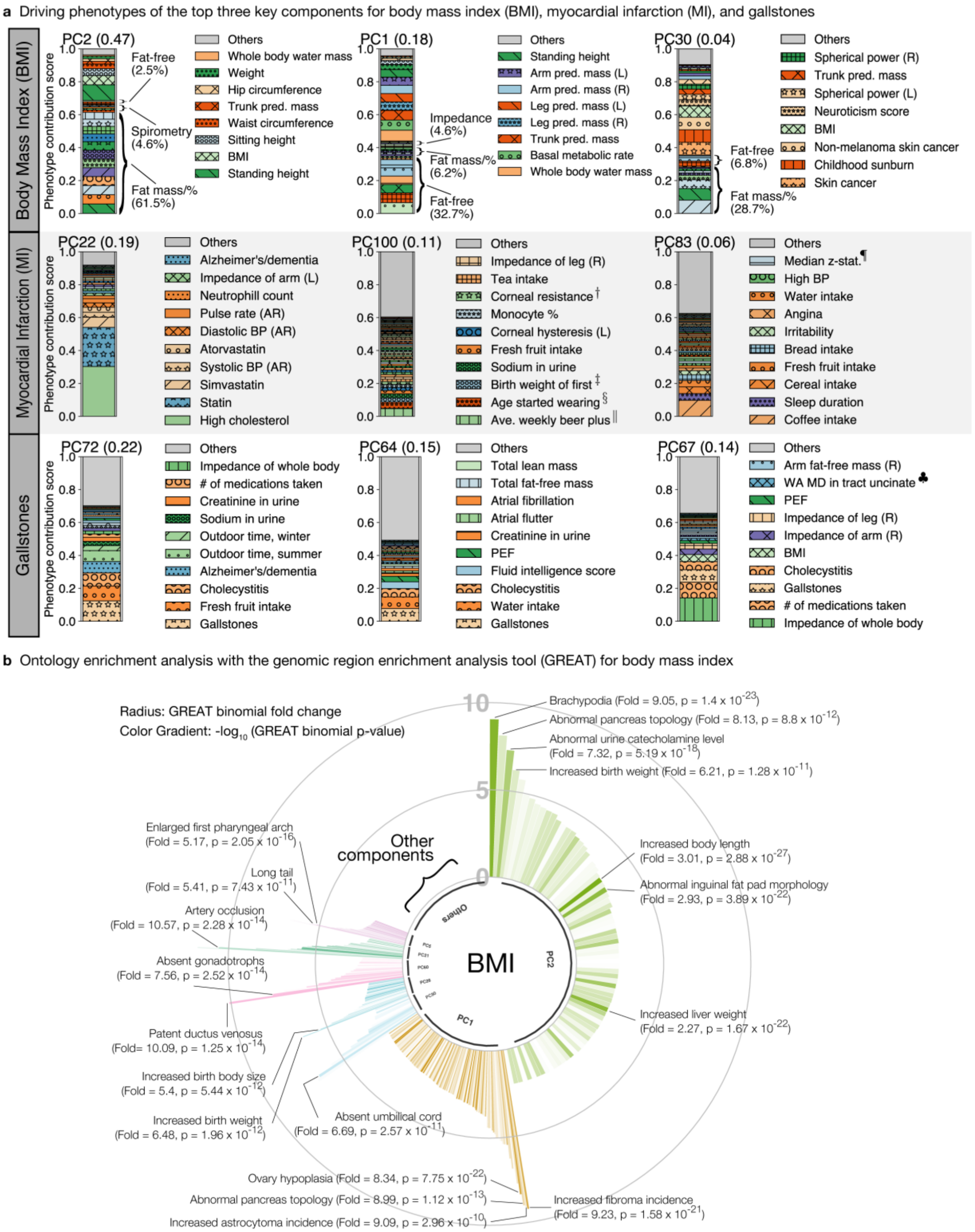


Fig. 2 Characterization of latent structures of genetic associations from genome- and-phenome-wide association summary statistics with DeGAs. **a**-**b** Components from truncated singular value decomposition (TSVD) corresponds to principal components in the phenotype (**a**) and variant (**b**) spaces. The first two components of all variants, excluding the MHC region, and relevant phenotypes are shown. **b** For variant PCA, we show biplot arrows (red) for selected phenotypes to help interpretation of the direction of principal components (Methods). The variants are labeled based on the genomic positions and the corresponding gene symbols. For example, "16:53813367 (*FTO*)" indicates the variant in gene *FTO* at position 53813367 on chromosome 16. **c**-**d** Phenotype (**c**) and gene (**d**) contribution scores for the first five components. PC1 is driven by largest part of the body mass that accounts for the "healthy part" (main text) including whole-body fat-free mass and genetic variants on *FTO* and *DLEU1*, whereas PC2 is driven by fat-related measurements, PC3 is driven by bioelectrical impedance measurements, PC4 is driven by eye measurements, and PC5 is driven by bioelectrical impedance and spirometry measurements along with the corresponding genetic variants (main text, Supplementary Table S3-4). Each colored segment represents a phenotype or gene with at least 0.5% and 0.05% of phenotype and gene contribution scores, respectively, and the rest is aggregated as others on the top of the stacked bar plots. The major

contributing phenotype groups (Methods, Supplementary Table S3) and additional top 10 phenotypes and the top 10 genes for each component are annotated in **c** and **d**, respectively. Abbreviations. pred.: predicted, %: percentage, mass/% mass and percentage, BP: blood pressure, AR: automated reading, L: left, R: right.

Figure 3



a Driving phenotypes of the top three key components for body mass index (BMI), myocardial infarction (MI), and gallstones

b Ontology enrichment analysis with the genomic region enrichment analysis tool (GREAT) for body mass index

Fig.3 The top three key latent components from DeGAs of coding and non-coding variants for body mass index (BMI), myocardial infarction (MI), and gallstones. **a** The top three key components for each phenotype are identified by phenotype squared cosine scores and characterized with the driving phenotypes by phenotype contribution scores (Methods). Each colored segment

represents a phenotype with at least 0.5% of phenotype contribution scores for each of the component and the rest of the phenotypes are aggregated as others and shown as the gray bar on the top. For BMI, additional phenotype grouping is applied (Methods, Supplementary Table S3). **b** Biological characterization of driving non-coding and coding variants of the key components for BMI with GREAT. The key components are shown proportional to their squared cosine score along with significantly enriched terms in mouse MGI phenotype ontology. The radius represents binomial fold change and the color gradient represents p-value from GREAT ontology enrichment analysis. Abbreviations. pred.: predicted, #: number, %: percentage, mass/% mass and percentage, BP: blood pressure, AR: automated reading, L: left, R: right, WA: weighted average. †: Corneal resistance factor (right), ‡: Birth weight of first child, §: Age started wearing glasses or contact lenses, ‖: Average weekly beer plus cider intake, ¶: Median z-statistic (in group-defined mask) for shapes activation, ♣: Weighted-mean MD in tract uncinate fasciculus (right).

Figure 4

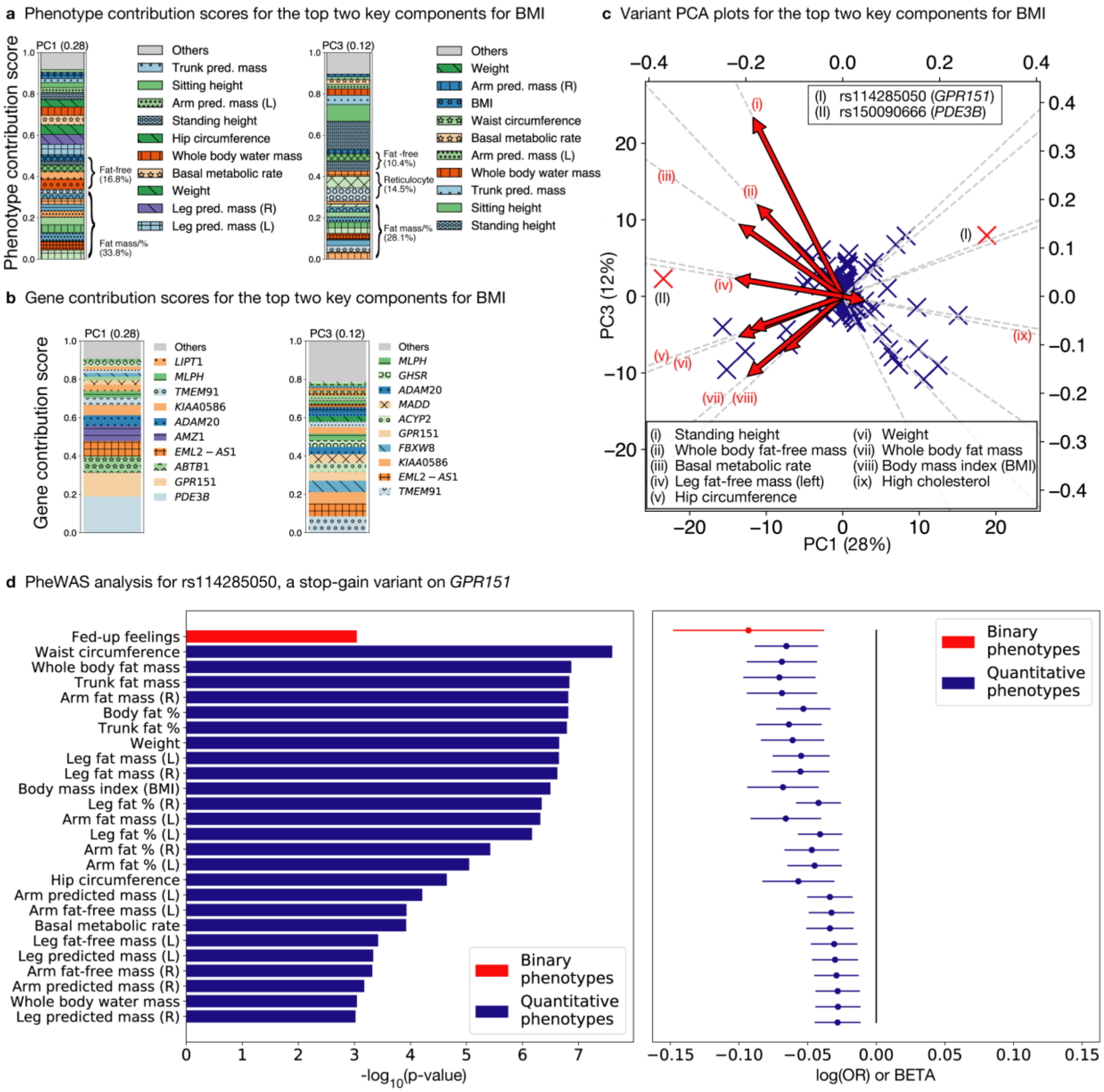


Fig. 4 DeGAs applied to the protein-truncating variants (PTVs) dataset. **a**-**b** Phenotype (**a**) and gene (**b**) contribution scores for the top key components associated with BMI based on phenotype grouping (Methods, Supplementary Table S3). **c** Variant PCA plot with biplot annotations for the top two components (Methods). The identified targets for functional follow-up (main text) are marked as (I) rs114285050 (a stop-gain variant on *GPR151*) and (II) rs150090666 (*PDE3B*). **d** Phenome-wide association analysis for *GPR151* rs114285050. The p-values (left) and log odds ratio (OR) (binary phenotypes, shown as red) or beta (quantitative phenotypes, shown as blue) (right) along with 95% confidence interval are shown for the phenotypes with minimum case count of 1,000 (binary phenotypes) or 1,000 individuals with non-missing values (quantitative phenotypes) and strong association (p < 0.001) and with this variants among all the phenotypes used in the study. Abbreviations: L: left, R: right, %: percentage, pred: predicted.

Figure 5

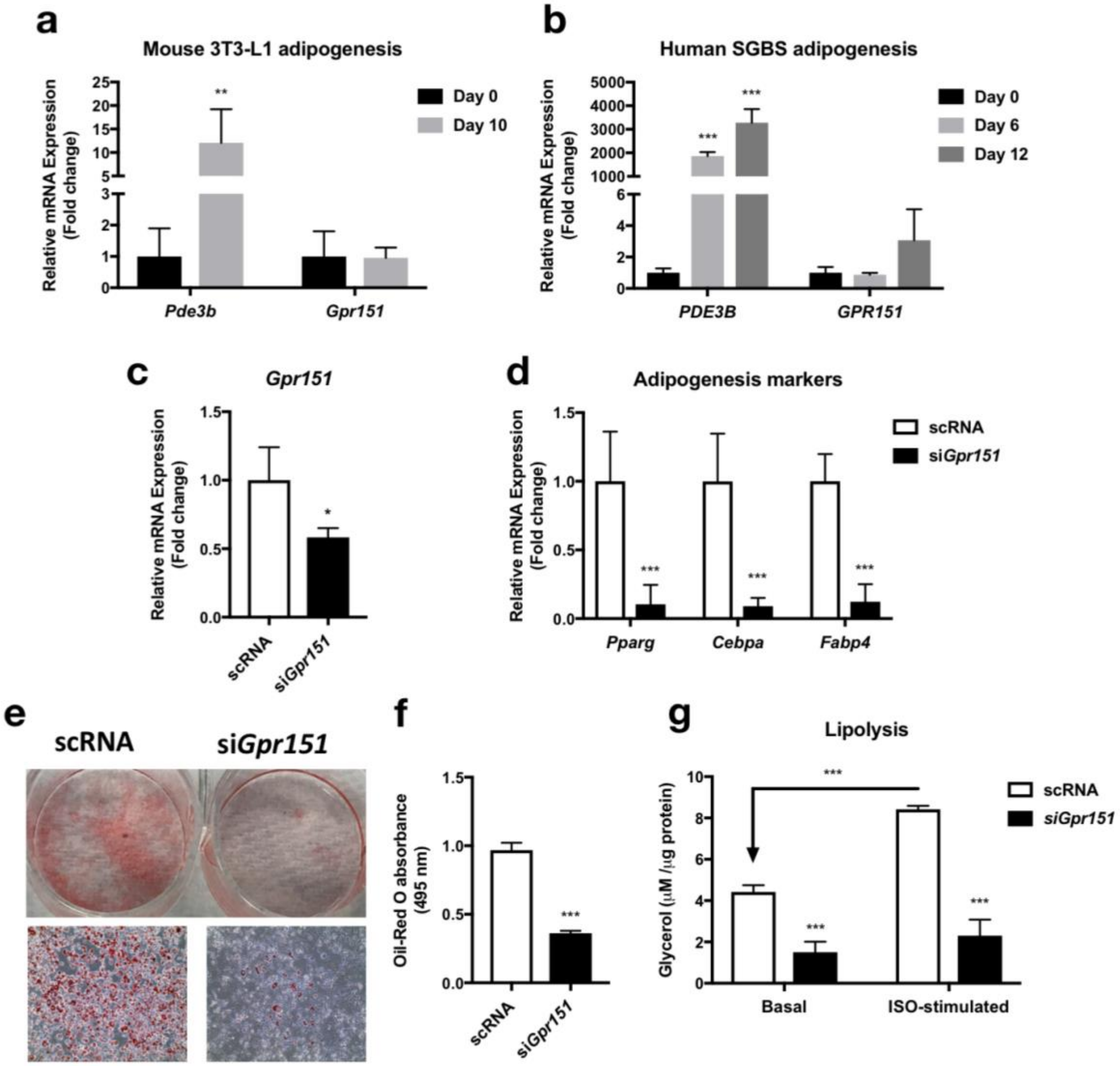


Fig. 5 Experimental validation of *GPR151* and *PDE3B* function in cellular models of adipogenesis. **a**-**b** qPCR analysis of gene expression patterns of *PDE3B* and *GPR151* during (**a**) mouse 3T3-L1 adipogenesis and (**b**) human SGBS adipogenesis. **c** qPCR analysis of *Gpr151* mRNA knockdown in 3T3-L1 preadipocytes. **d** qPCR analysis of the effect of si*Gpr151* knockdown on adipogenesis markers, *Pparg*, *Cebpa* and *Fabp4*. **e-g** Oil-Red O staining (**e**), quantification of lipid droplets (**f**), and lipolysis (**g**) in scRNA- or si*Gpr151*-tansfected adipocytes. Means ± SEM are shown (***p-value<0.001, **p-value<0.01, *p-value <0.05). scRNA: scrambled siRNA. ISO: isoproterenol.

Supplementary Materials

List of supplementary materials

- Fig. S1: Variant filtering workflow
- Fig. S2: Scree plot
- Fig. S3: Squared cosine score (all variants dataset)
- Fig. S4: Gene contribution score (all variants dataset)
- Fig. S5: Variant PCA plot for MI
- Fig. S6: Variant PCA plot for gallstones
- Fig. S7: GREAT enrichment analysis for MI
- Fig. S8: GREAT enrichment analysis for gallstones
- Fig. S9: Squared cosine score of BMI (PTVs dataset)
- Fig. S10: PheWAS analysis for *PDB3B*
- Fig. S11: Univariate regression analysis for *GPR151*
- Fig. S12: Univariate regression analysis for *PDE3B*
- Fig. S13: *GPR151* overexpression
- Fig. S14: Effects of *Pde3b* knockdown in 3T3-L1 adipogenesis
- Table S1: List of phenotype categories
- Table S2: List of phenotypes
- Table S3: Phenotype groupings for visualization
- Table S4: Summary of contribution scores for the key components
- Table S5: GREAT enrichment analysis for BMI
- Table S6: GREAT enrichment analysis for MI
- Table S7: GREAT enrichment analysis for gallstones
- Table S8: PheWAS analysis for rs114285050 (*GPR151*)
- Table S9: PheWAS analysis for rs150090666 (*PDE3B*)

Fig. S1: Variant filtering workflow

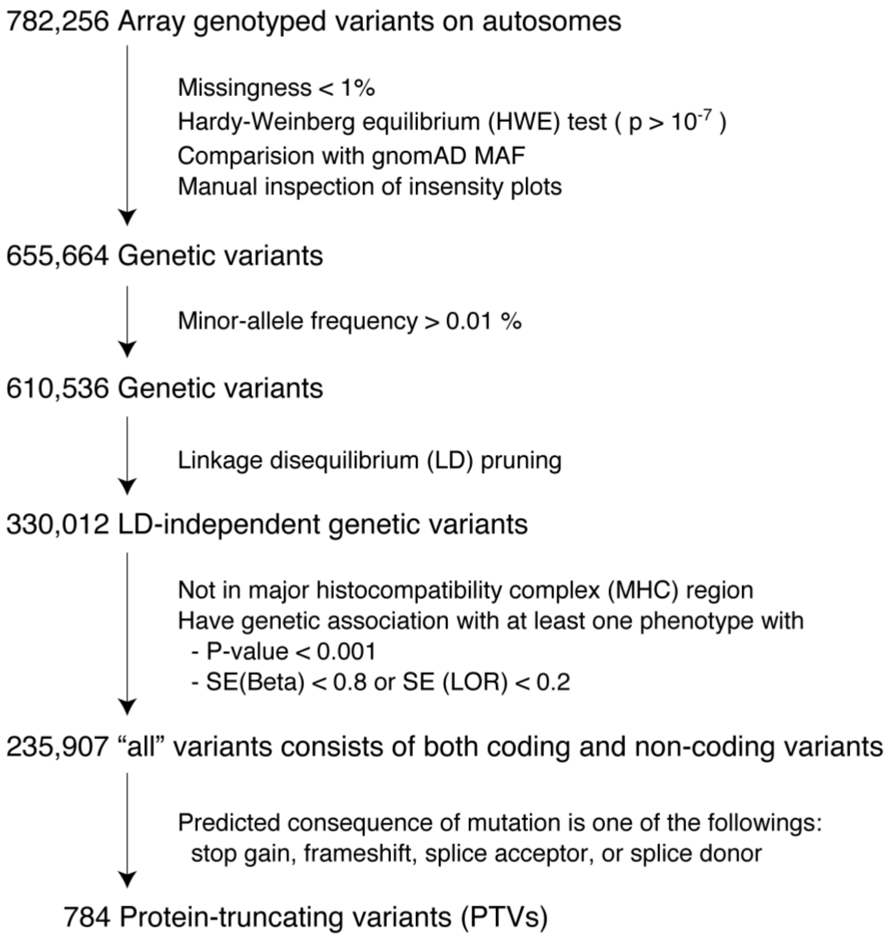

782,256 Array genotyped variants on autosomes
Missingness < 1%
Hardy-Weinberg equilibrium (HWE) test (p > 10-7)
Comparision with gnomAD MAF
Manual inspection of insensity plots
655,664 Genetic variants
Minor-allele frequency > 0.01 %
610,536 Genetic variants
Linkage disequilibrium (LD) pruning
330,012 LD-independent genetic variants
Not in major histocompatibility complex (MHC) region
Have genetic association with at least one phenotype with
- P-value < 0.001
- SE(Beta) < 0.8 or SE (LOR) < 0.2
235,907 “all” variants consists of both coding and non-coding variants
Predicted consequence of mutation is one of the followings:
stop gain, frameshift, splice acceptor, or splice donor
784 Protein-truncating variants (PTVs)

Fig. S1 Illustrative summary of the variant filters used in the study. The last two variant sets (“all” variants and PTVs) are used in the study. Abbreviations. SE: standard error. LOR: log odds ratio.

Fig. S2: Scree plot

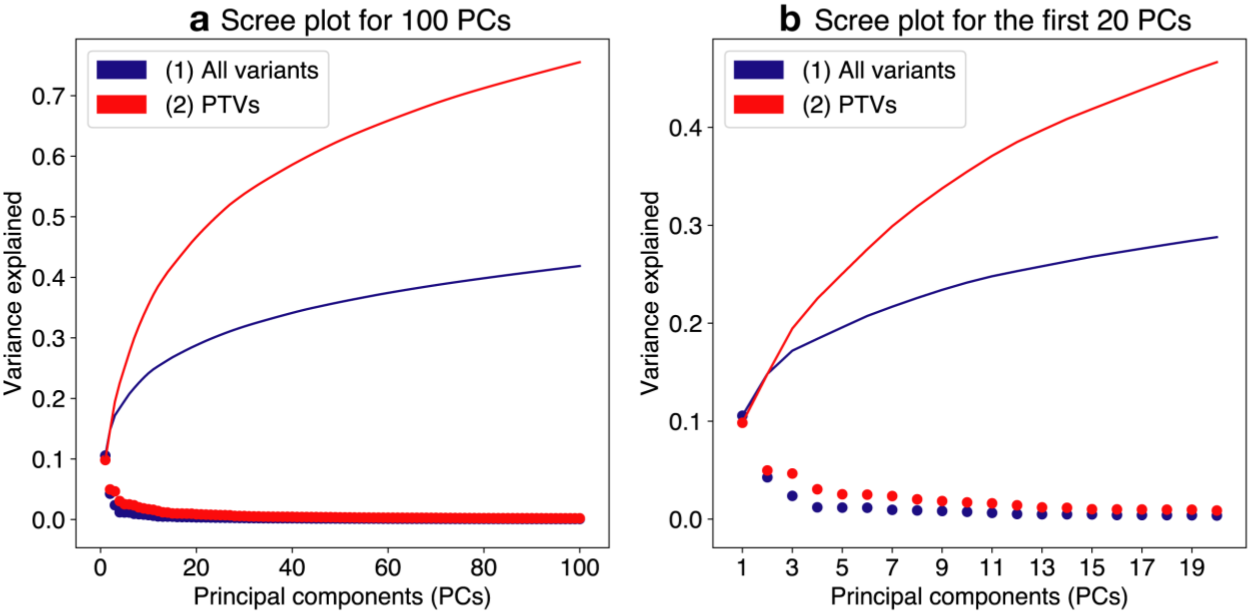


Fig. S2 Scree plot summarizes variance explained in each of the top 100 (**a**) and 20 (**b**) components. The scree plots are shown for two datasets consists of LD-pruned and QC-filtered sets of array-genotyped variants outside of MHC region: (1) all array-genotyped variants, which includes coding and non-coding variants (blue) and (2) protein-truncating variants (PTVs, red). For each component, we calculate the variance explained defined as squared eigenvalues divided by the total variance in the original matrix (Methods). We plotted those values as dots and cumulative values as lines.

Fig. S3: Squared cosine score (all variants dataset)

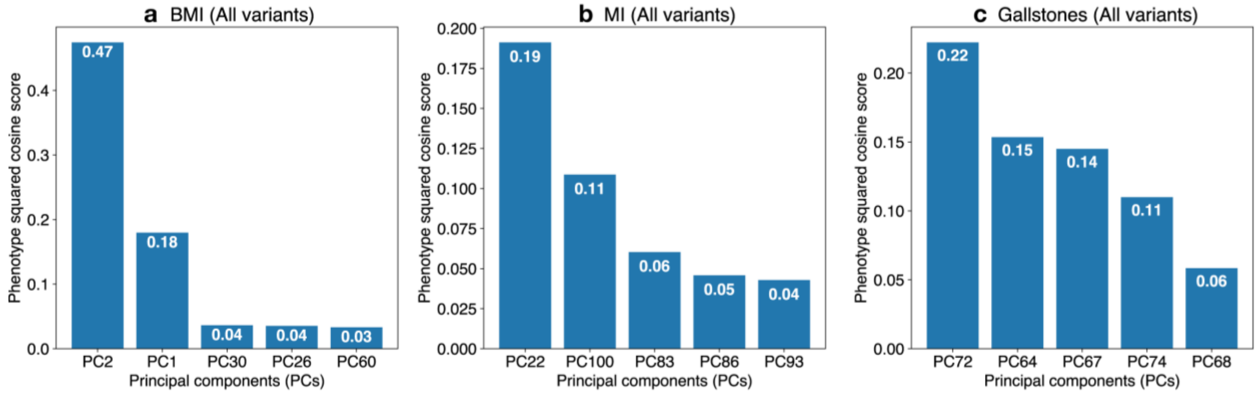


Fig. S3 Identification of the key components with phenotype squared cosine scores. Squared cosine score quantifies relative importance of the key components for a given phenotype. The top five key components are identified for all variant dataset that includes both coding and non-coding variants for three phenotypes: **a** body mass index (BMI), **b** myocardial infarction (MI), and **c** gallstones. The top five key components are shown on the horizontal axis and the corresponding squared cosine scores are shown on the vertical axis.

Fig. S4: Gene contribution score (all variants dataset)

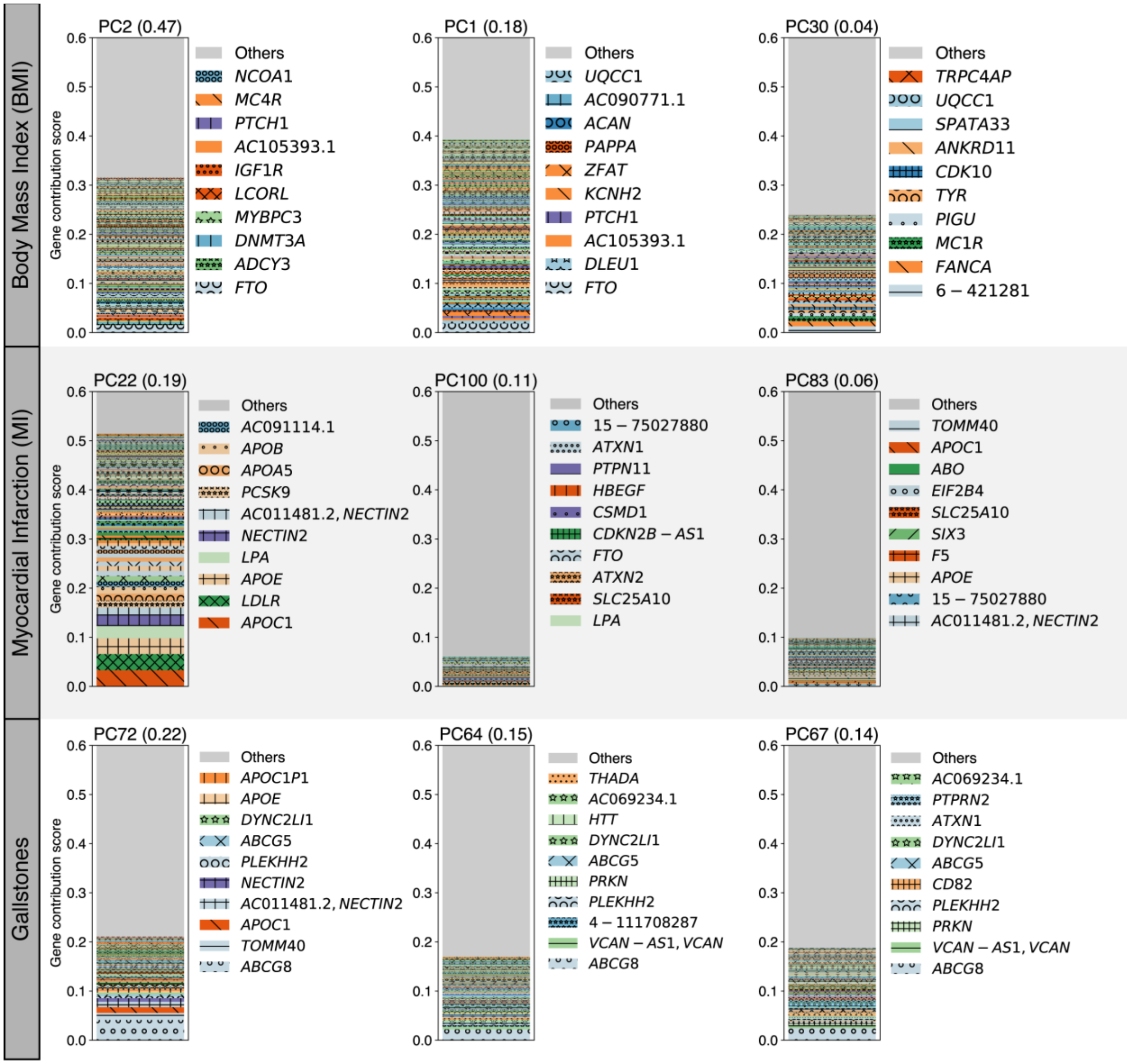


Fig. S4 Gene contribution scores for the top three key components for body mass index (BMI), myocardial infarction (MI), and gallstones using all variant dataset, which includes both coding and non-coding variants. For each phenotype, the top three key components with their phenotype squared cosine scores are shown on the top of the stacked bar plot and gene contribution scores for each of the components are shown as colored segments. Each colored segment represents a gene with at least 0.05% of contribution scores and the rest of the genes are aggregated as the gray bar at the top. For the visualization, the maximum value of the vertical axis is set to be 0.6. For each component, the labels for the top 10 driving genes are shown. For non-coding variants, we display their genomic coordinates.

Fig. S5: Variant PCA plot for myocardial infarction.

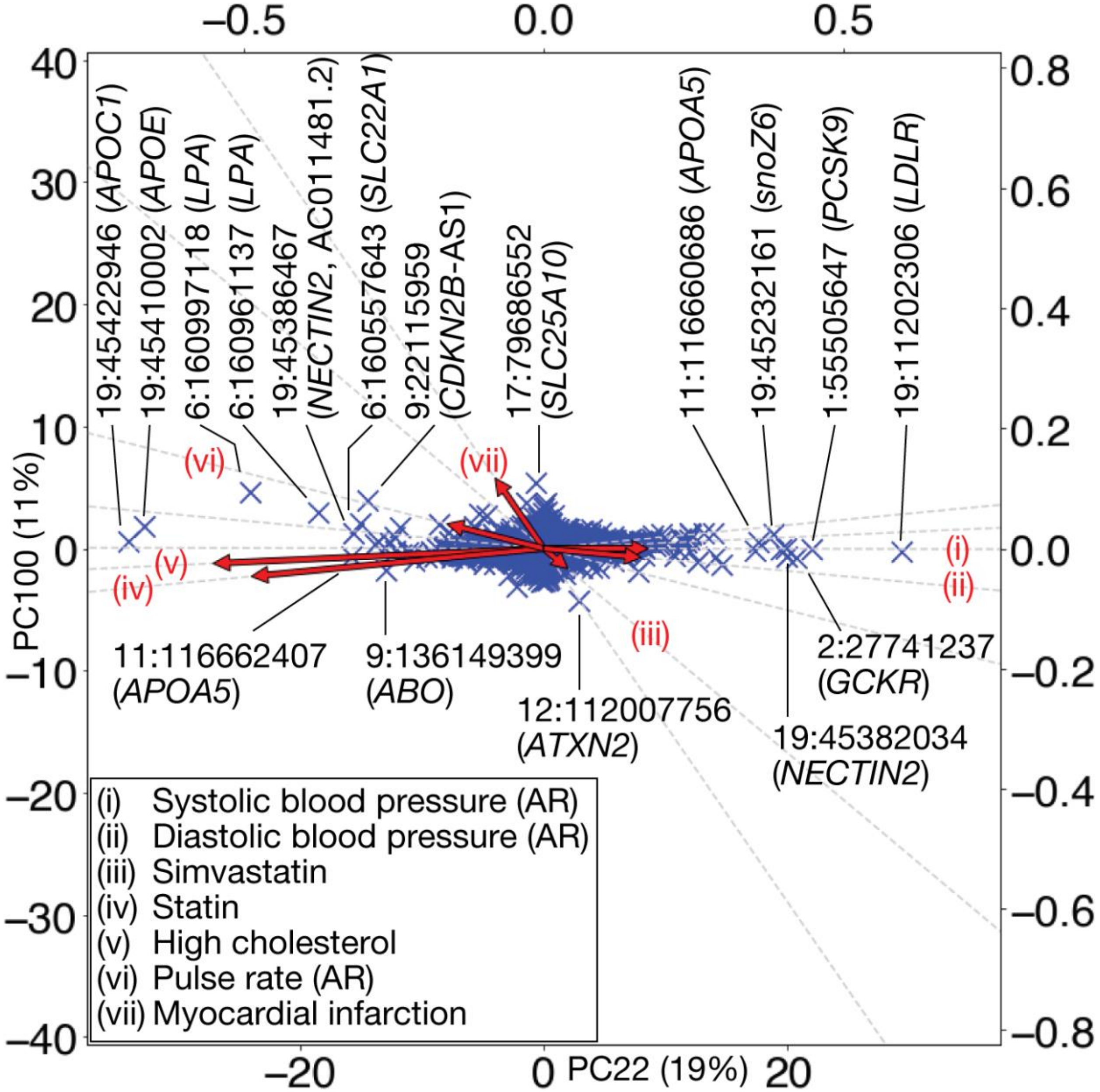


Fig. S5 Variant PCA plot with biplot annotation for the top two key components for myocardial infarction using "all" dataset. Genetic variants projected into the top two key components, PC22 (horizontal axis) and PC100 (vertical axis) are shown as scatter plot. Variants are annotated with gene symbols. Directions of genetic associations for relevant phenotypes are annotated as red arrows using the secondary axes (Methods). Abbreviations. AR: automated reading.

Fig. S6: Variant PCA plot for Gallstones

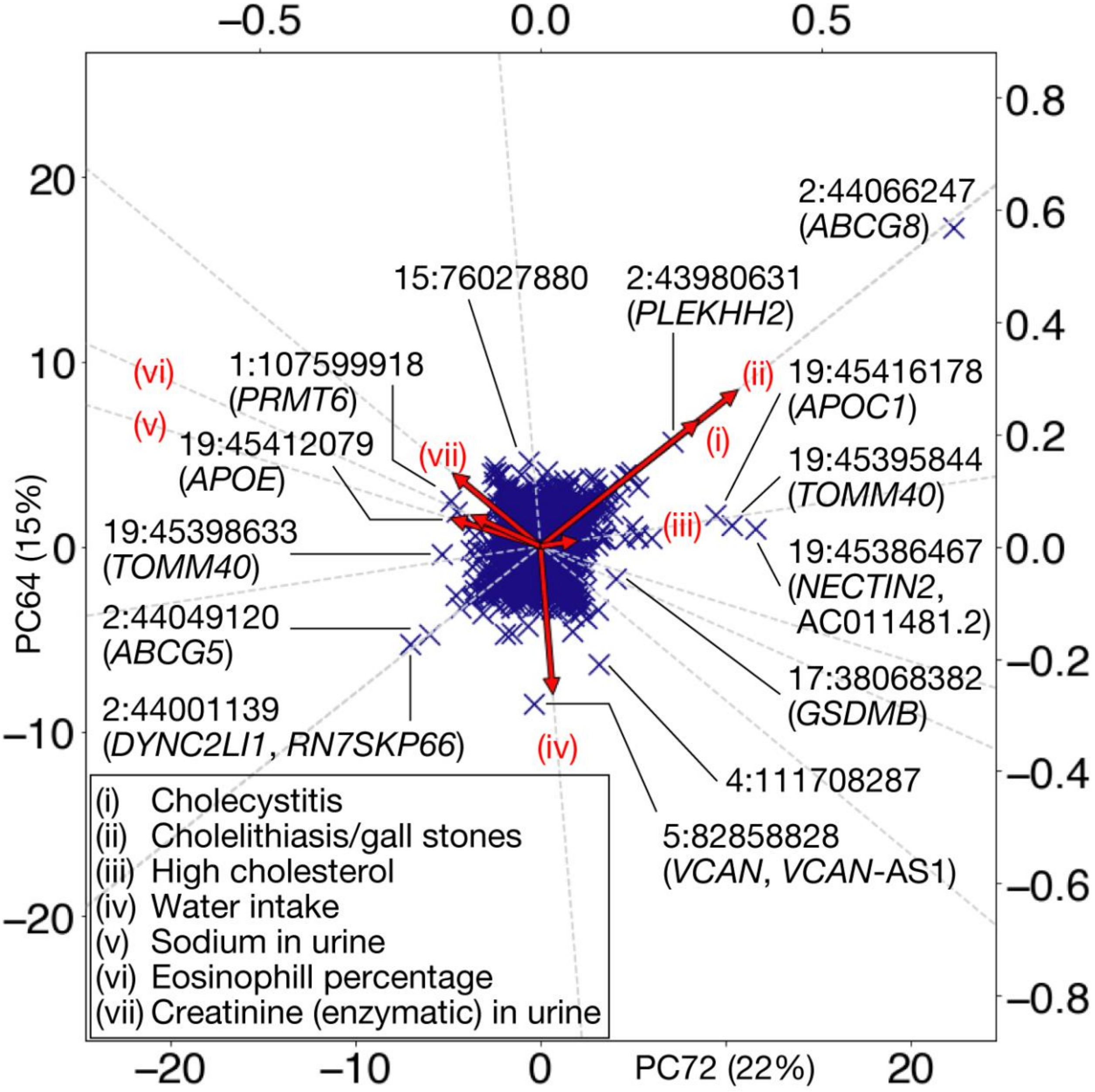


Fig. S6 Variant PCA plot with biplot annotation for the top two key components for gallstones using "all" dataset. Genetic variants projected into the top two key components, PC72 (horizontal axis) and PC64 (vertical axis). Variants are annotated with gene symbols. Directions of genetic associations for relevant phenotypes are annotated as red arrows using the secondary axes (Methods).

Fig. S7: GREAT enrichment analysis for MI

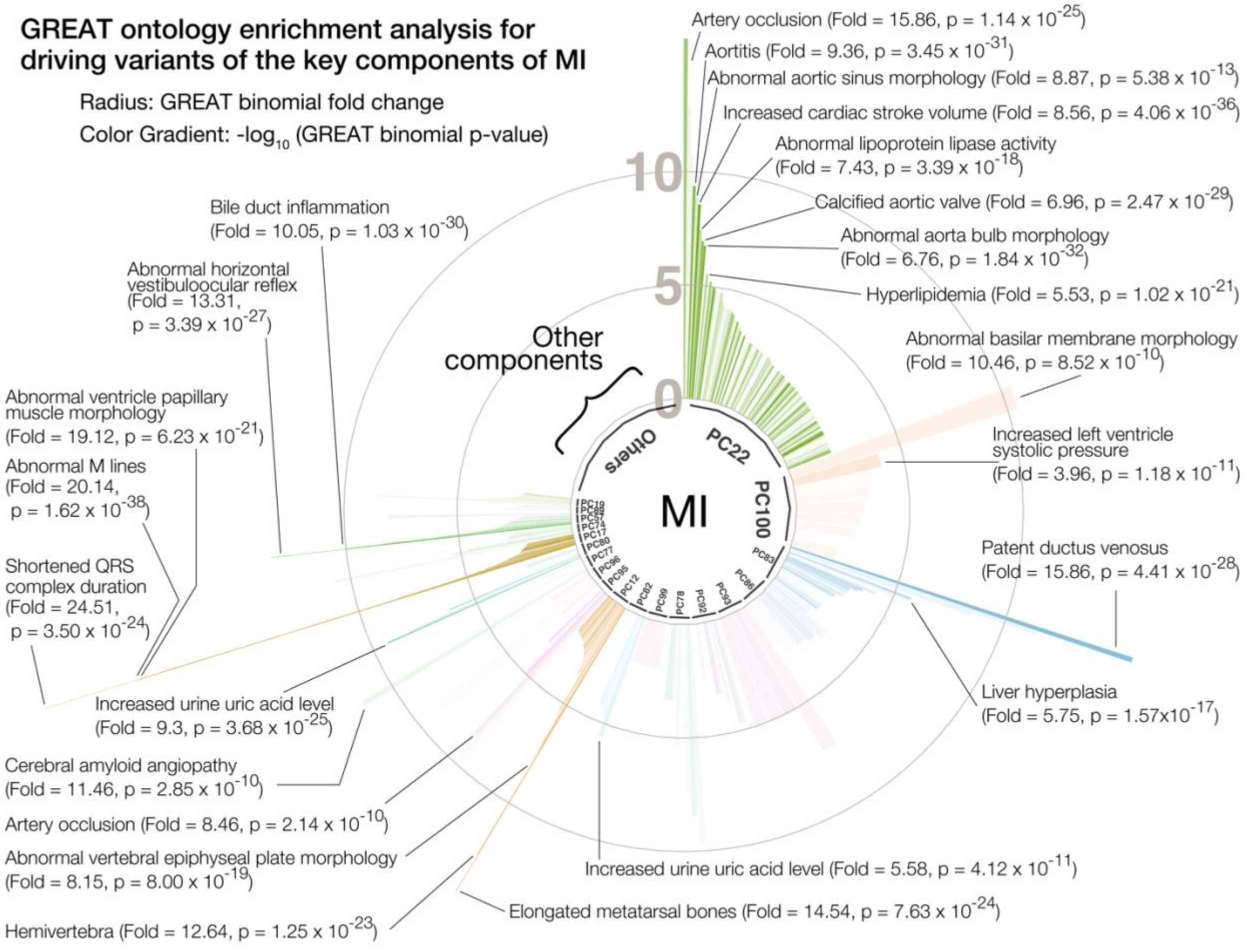


Fig. S7 Biological characterization of driving non-coding and coding variants of the key components for myocardial infarction (MI) with the genomic region enrichment analysis tool (GREAT) using the all variants dataset. The key components are shown proportional to their squared cosine score along with significantly enriched terms in mouse genome informatics (MGI) phenotype ontology. The radius represents binomial fold change and the color gradient represents p-value from GREAT ontology enrichment analysis.

Fig. S8: GREAT enrichment analysis for gallstones

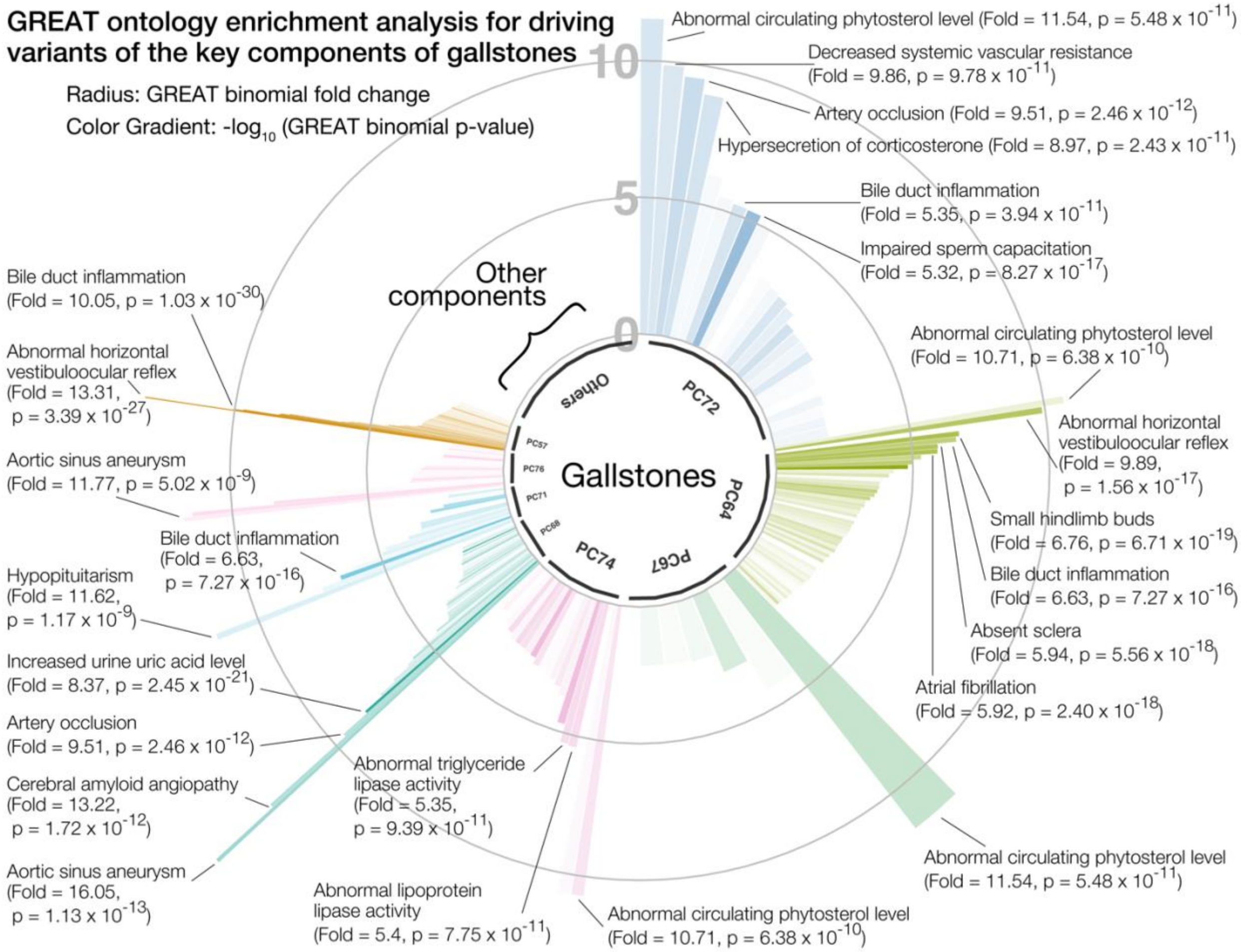
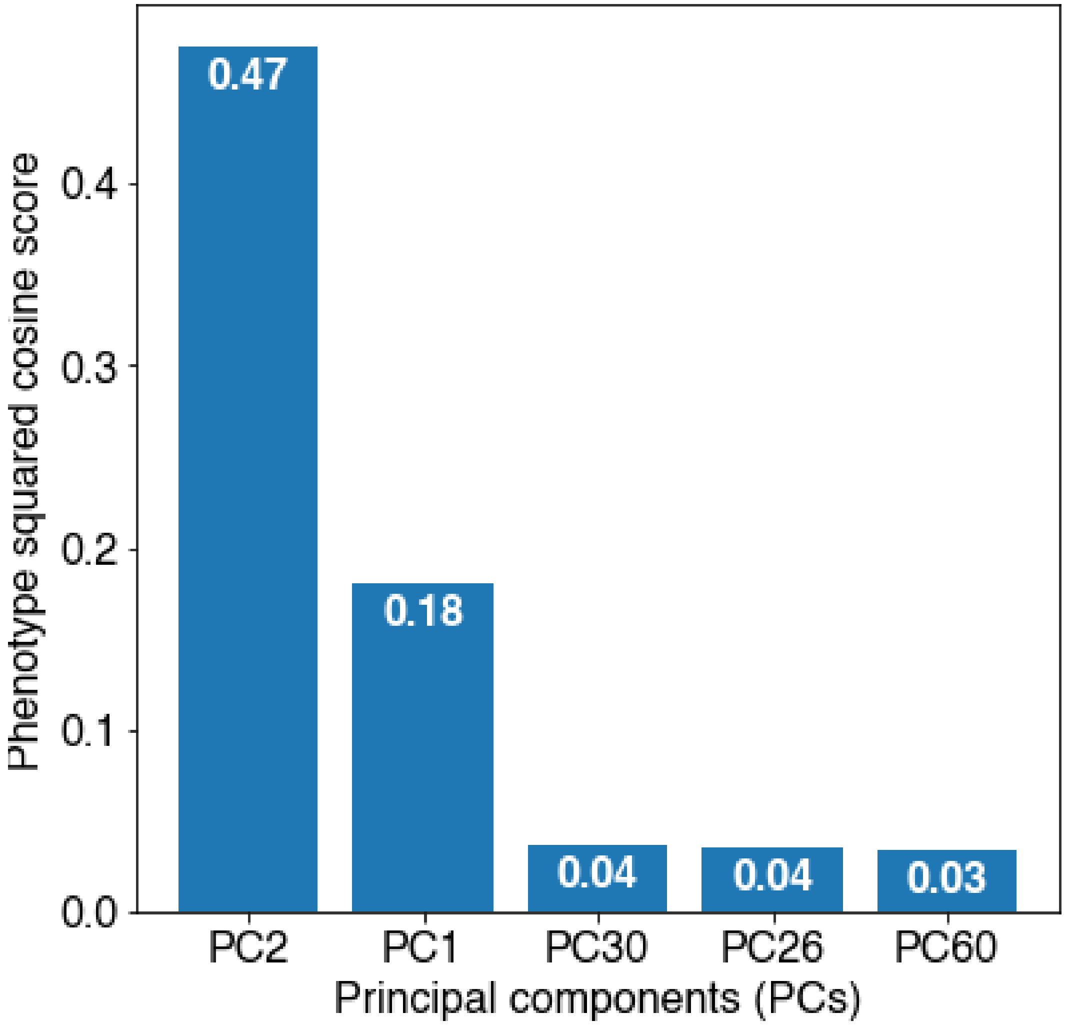


Fig. S8 Biological characterization of driving non-coding and coding variants of the key components for gallstones with the genomic region enrichment analysis tool (GREAT) using the all variants dataset. The key components are shown proportional to their squared cosine score along with significantly enriched terms in mouse genome informatics (MGI) phenotype ontology. The radius represents binomial fold change and the color gradient represents p-value from GREAT ontology enrichment analysis.

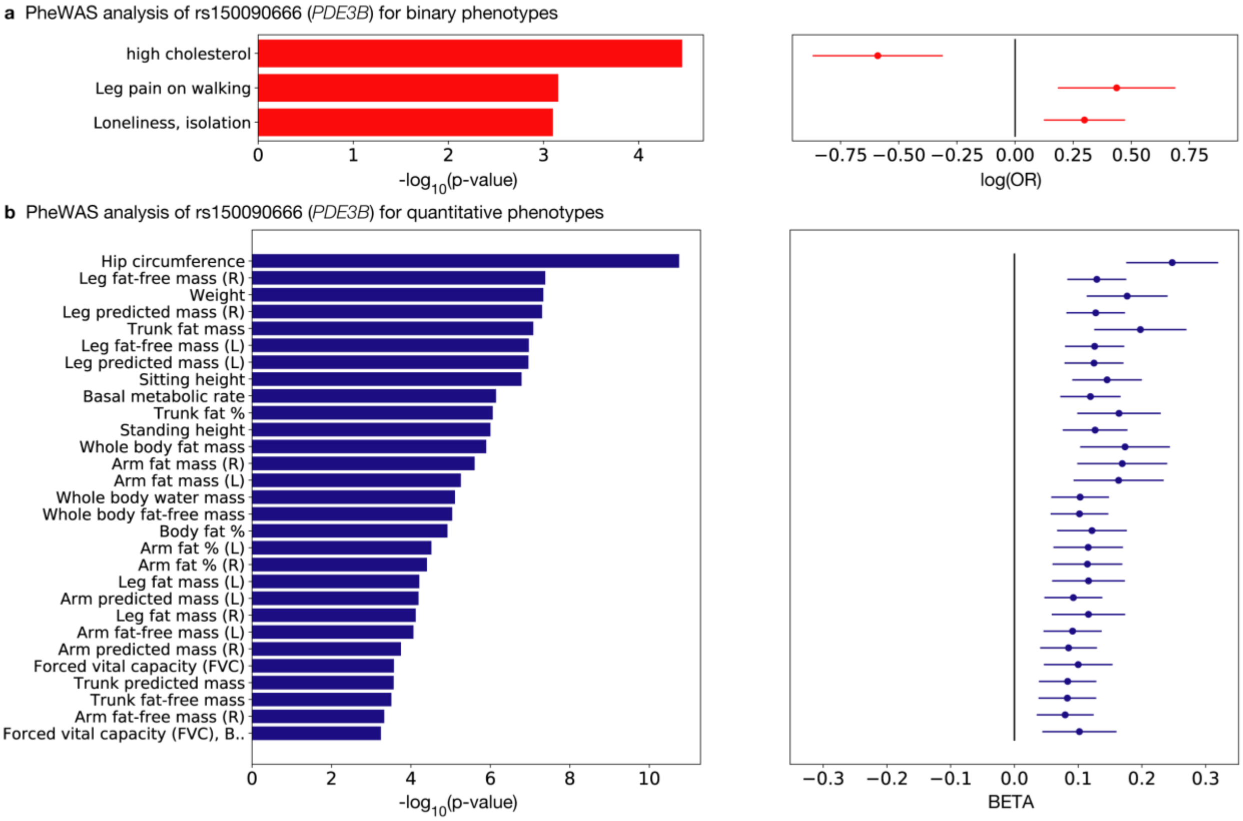
953 Fig. S9: Squared cosine score of BMI (PTVs dataset)



954

955 **Fig. S9** Identification of the key components for BMI with phenotype squared cosine scores 956 using the PTVs dataset. The top five key components are shown on the horizontal axis and the 957 corresponding squared cosine scores are shown on the vertical axis.

Fig. S10: PheWAS analysis for *PDE3B*

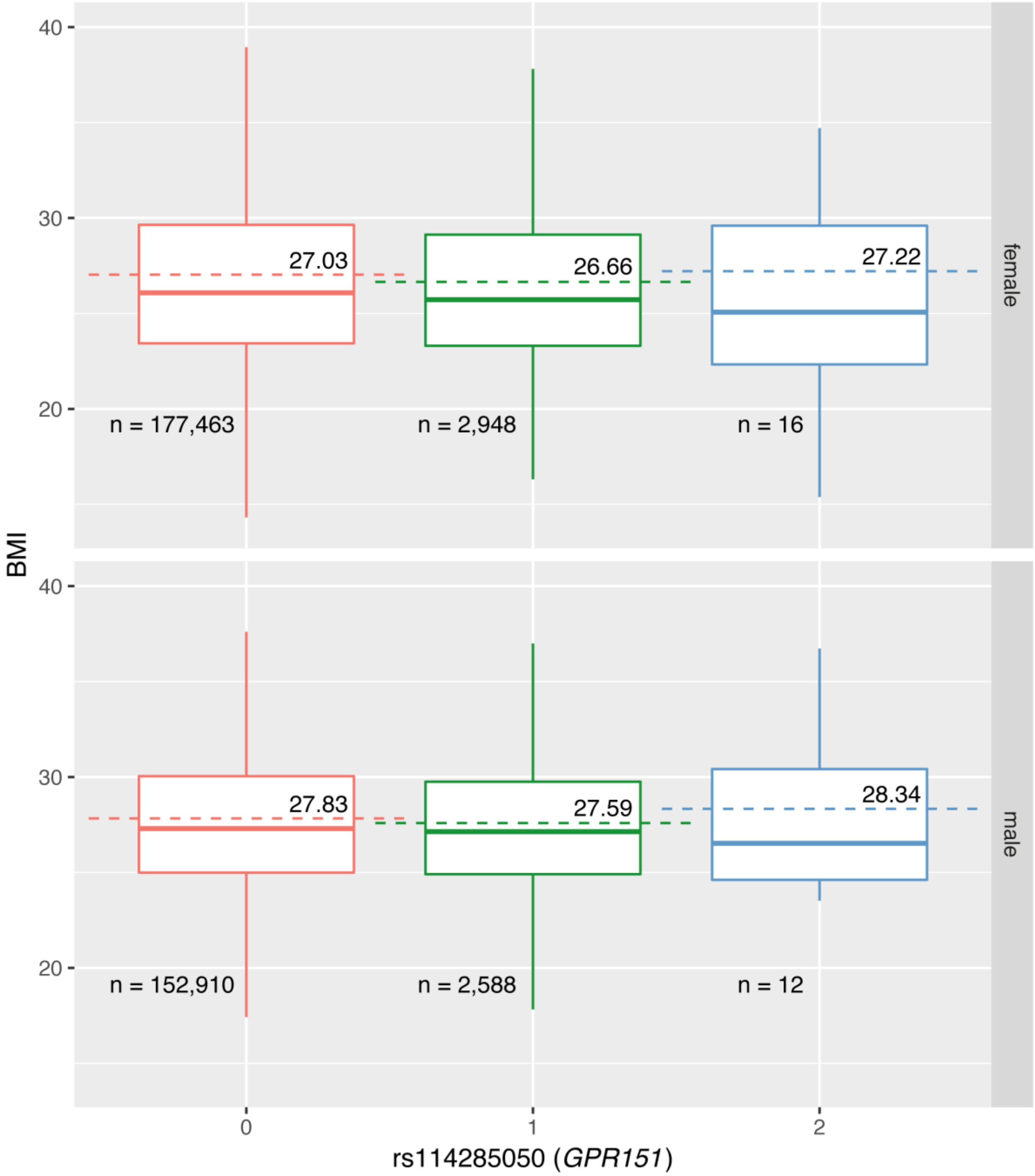


a PheWAS analysis of rs150090666 (*PDE3B*) for binary phenotypes

b PheWAS analysis of rs150090666 (*PDE3B*) for quantitative phenotypes

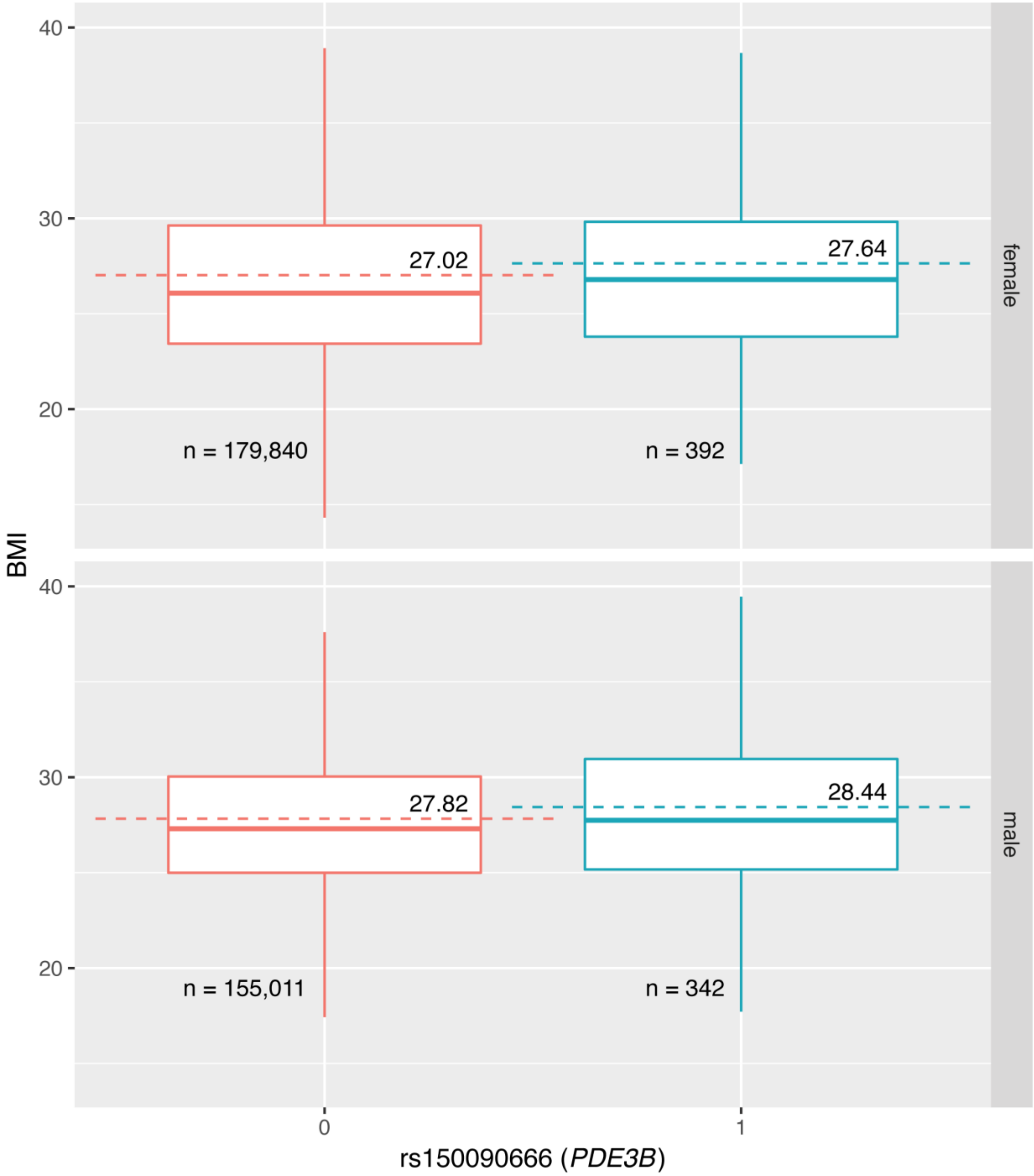
Fig. S10 Phenome-wide association (PheWAS) analysis for rs150090666, a stop-gain variant in *PDE3B*. The p-values (left) and log odds ratio (binary phenotypes, shown as red) or beta (quantitative phenotypes, shown as blue) (right) along with 95% confidence interval are shown for the phenotypes with minimum case count of 1,000 (binary phenotypes, **a**) or 1,000 individuals with non-missing values (quantitative phenotypes, **b**) and strong association ($p \leq 0.001$) and with this variants among all the phenotypes used in the study.

966 Fig. S11: Univariate regression analysis for *GPR151*



967
968
969 **Fig. S11** Distribution of BMI stratified by sex and genotype of rs114285050, a stop-gain variant
970 in *GPR151*. The outliers are removed from the plot and the mean values are annotated and
971 shown as dashed lines. The number of carriers of the variants are shown at the bottom.
972

973 Fig. S12: Univariate regression analysis for *PDE3B*



974

975 **Fig. S12** Distribution of BMI stratified by sex and genotype of rs150090666, a stop-gain variant
976 in *PDE3B*. The outliers are removed from the plot and the mean values are annotated and
977 shown as dashed lines. The number of carriers of the variants are shown at the bottom.

978

Fig. S13: *GPR151* overexpression

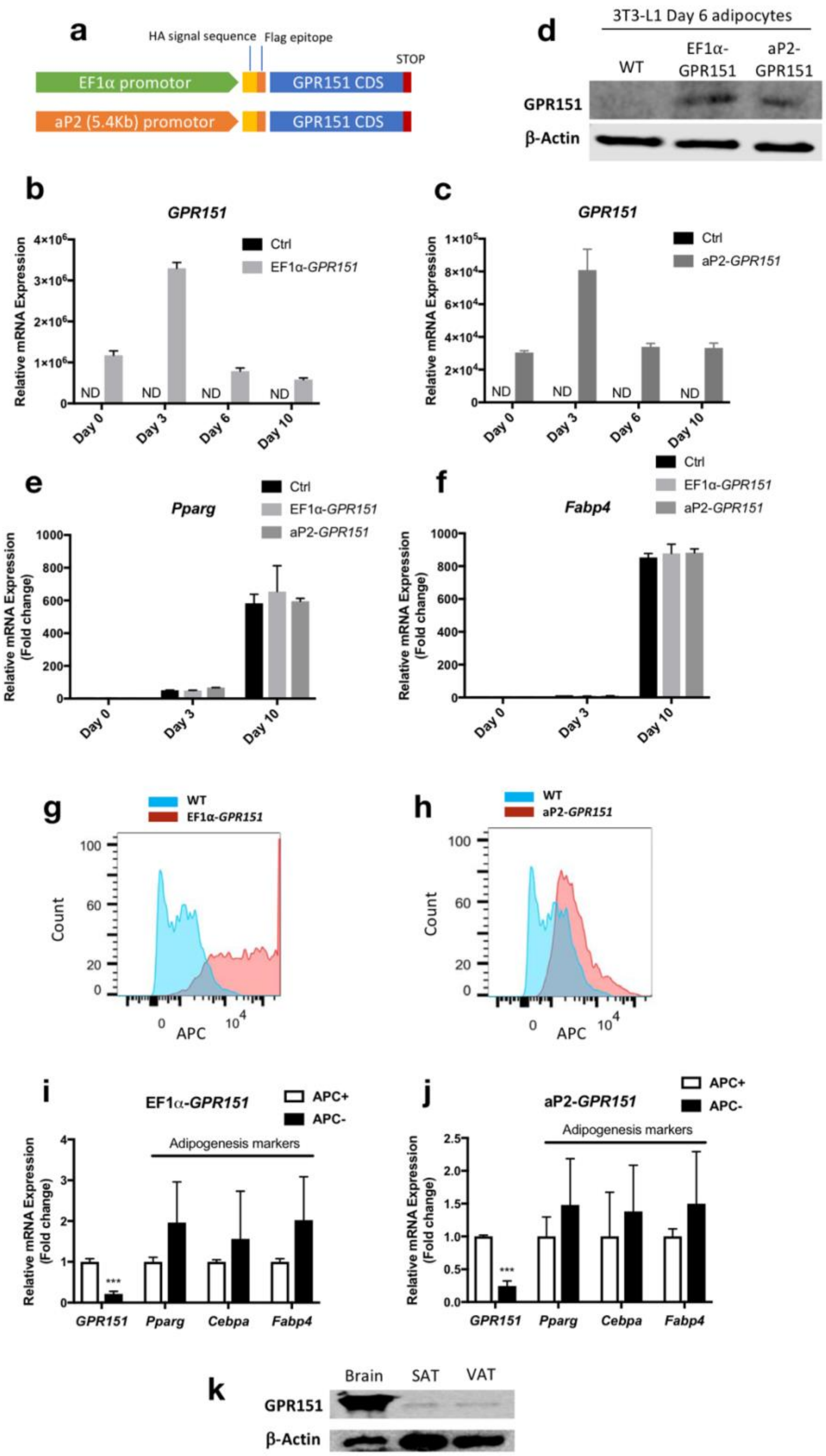


Fig. S13 Effects of *GPR151* overexpression on 3T3-L1 adipogenesis. **a** Structure of *GPR151* overexpression construct driven by either EF1α or aP2 promotor. **b**-**d** Confirmation of *GPR151*

overexpression at both mRNA (**b**-**c**) and protein levels (**d**) in 3T3-L1 cells during adipogenesis. **e**-**f** qPCR analysis of the effect of *GPR151* overexpression on adipogenesis markers, *Pparg* (**e**) and *Fabp4* (**f**). **g**-**h** FACS analysis of APC fluorescence in Day 6 3T3-L1 adipocytes infected with either EF1α-*GPR151* (**g**) or aP2-*GPR151* (**h**) (shown in red), in comparison to wild-type (WT) cells (shown in blue). **i**-**j** Relative mRNA levels of *GPR151* and adipogenic markers (*Pparg, Cebpa, Fabp4*) in purified APC+ and APC- cells from Day 6 3T3-L1 adipocytes infected by either EF1α-*GPR151* (**i**) or aP2-*GPR151* (**j**). **k** Comparison of protein levels of GPR151 in mouse brain, subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT). ND: not-detectable.

Fig. S14: *Pde3b* knockdown

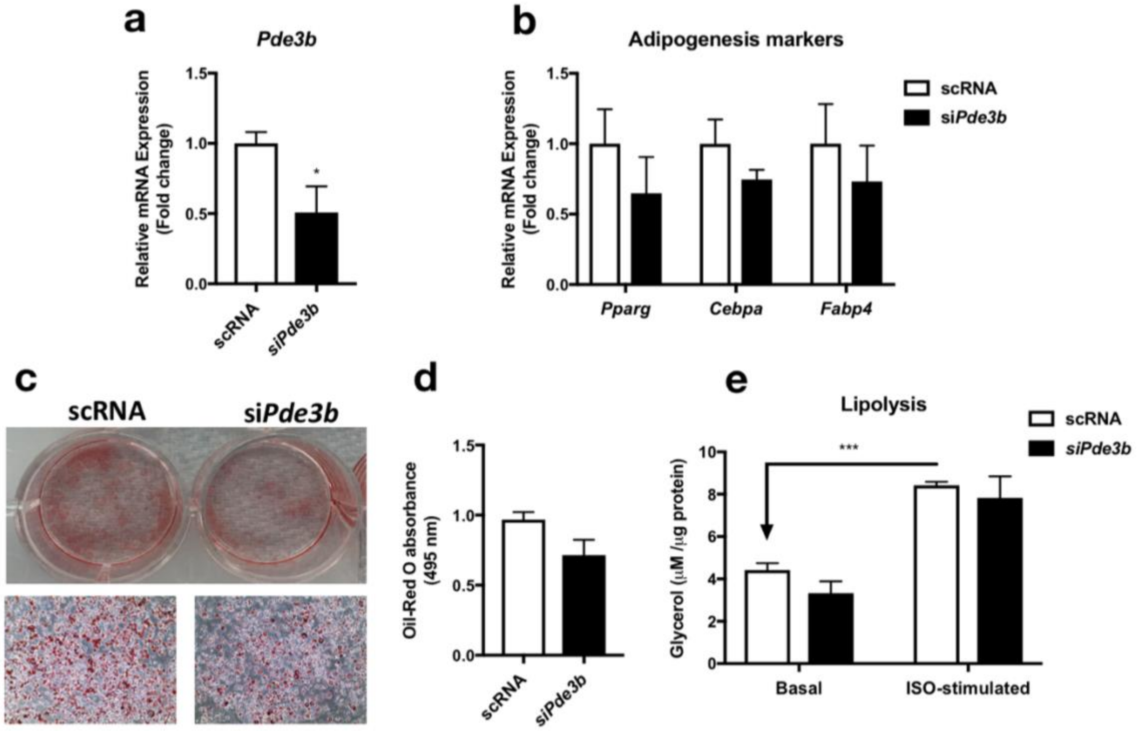


Fig. S14 Effects of *Pde3b* knockdown in 3T3-L1 adipogenesis. **a** qPCR analysis of *Pde3b* mRNA knockdown in 3T3-L1 preadipocytes. **b** qPCR analysis of the effect of si*Pde3b* knockdown on adipogenesis markers, *Pparg*, *Cebpa* and *Fabp4*. **c**-**d** Oil-Red O staining (**c**) and quantification (**d**) of lipid droplets in scRNA- or si*Pde3b*-tansfected adipocytes. **e** lipolysis assays of scRNA- or si*Pde3b*-tansfected adipocytes. Means ± SEM are shown (***p-value<0.001, *p-value<0.05). scRNA: scrambled siRNA. ISO: isoproterenol.

Table S1 List of phenotype categories

List of phenotype categories used in our study and their data source are shown with one example phenotype per category. Abbreviation in the type column. B: binary, Q: quantitative, P: described in previously published literature, F: the UK Biobank data field ID, and C: the UK Biobank data category ID.

Phenotype group name	Type	Number of phenotypes	Example	Data source
Disease outcome	B	363	Hypertension	P[12]
Cancer	B	46	Skin cancer	P[12]
Family History	B	10	High blood pressure	P[12]
Medication	B	709	Aspirin intake	F:20003
Questionnaire (binary)	Q	49	Wears glasses or contact lenses	C:100025
Imaging	Q	683	Volume of white matter	C:100003
Physical Measurement	Q	122	Standing height	C:100006
Assay	Q	34	Red blood cell (erythrocyte) count	C:100079
Questionnaire (quantitative)	Q	62	Sleep duration	C:100079
Miscellaneous (binary)	B	19	Ever attempted suicide	
Miscellaneous (quantitative)	Q	42	Number of medications taken	

Table S2 List of phenotypes

The list of phenotypes considered in the study. The table is sorted by category, number cases (for binary phenotypes), and the number of non-missing values (for quantitative phenotypes). The two columns, "All" and "PTVs" indicates whether the phenotype is used in each of the dataset after imposing the filters on the genome-and phenome-wide summary statistics matrix. One can browse the summary statistics from genome-wide association studies on the Global Biobank Engine with the URL in the table.

I am showing the first five lines of the table here. The full table is in Excel file.

Category	Phenotype name	Number of cases	All	PTVs	Global Biobank Engine phenotype page (URL)
Disease outcome	hypertension	107407	Y	Y	https://biobankengine.stanford.edu/coding/HC215
Disease outcome	essential hypertension	64234	Y	Y	https://biobankengine.stanford.edu/coding/HC273
Disease outcome	asthma	43626	Y	Y	https://biobankengine.stanford.edu/coding/HC382
Disease outcome	high cholesterol	43054	Y	Y	https://biobankengine.stanford.edu/coding/HC269

Table S3: Phenotype groupings for visualization

The list of phenotype groups used in the phenotype contribution score plots are summarized.

Phenotype groups	List of phenotypes in the group
fat-free	Arm fat-free mass (left)
	Arm fat-free mass (right)
	Leg fat-free mass (left)
	Leg fat-free mass (right)
	Total fat-free mass
	Trunk fat-free mass
	Whole body fat-free mass
fat	Android fat mass
	Android tissue fat percentage
	Arm fat mass (left)

	Arm fat mass (right)
	Arm fat percentage (left)
	Arm fat percentage (right)
	Arm tissue fat percentage (left)
	Arm tissue fat percentage (right)
	Arms fat mass
	Arms tissue fat percentage
	Body fat percentage
	Gynoid fat mass
	Gynoid tissue fat percentage
	Leg fat mass (left)
	Leg fat mass (right)
	Leg fat percentage (left)
	Leg fat percentage (right)
	Leg tissue fat percentage (left)
	Leg tissue fat percentage (right)
	Legs fat mass
	Legs tissue fat percentage
	Total fat mass
	Total tissue fat percentage
	Trunk fat mass
	Trunk fat percentage
	Trunk tissue fat percentage
	Whole body fat mass
impedance	Impedance of arm (left)
	Impedance of arm (right)

	Impedance of leg (left)
	Impedance of leg (right)
	Impedance of whole body
reticulocyte	High light scatter reticulocyte count
	High light scatter reticulocyte percentage
	Immature reticulocyte fraction
	Mean reticulocyte volume
	Reticulocyte count
	Reticulocyte percentage
meridian	3mm strong meridian (left)
	3mm strong meridian (right)
	3mm weak meridian (left)
	3mm weak meridian (right)
	6mm strong meridian (left)
	6mm strong meridian (right)
	6mm weak meridian (left)
	6mm weak meridian (right)
spirometry	Forced expiratory volume in 1-second (FEV1)
	Forced expiratory volume in 1-second (FEV1), Best measure
	Forced expiratory volume in 1-second (FEV1), predicted
	Forced expiratory volume in 1-second (FEV1), predicted percentage
	Forced vital capacity (FVC)
	Forced vital capacity (FVC), Best measure
	Peak expiratory flow (PEF)

Table S4: Summary of contribution scores for the key components

The list of top 20 driving phenotypes, genes, and variants for the first five principal components and the top three key components for the phenotypes highlighted in the study are summarized in the table.

I am showing the first four lines of the table here. The full table is in Excel file.

Dataset	Phenotype of interest	PC	Squared cosine score	Rank	Phenotype	Phenotype contribution score	Gene	Gene contribution score	Variant	Variant contribution score	rsid	GBE
All_variants	BMI	2	47.44%	1	Standing height	9.51%	FTO	1.52%	16-53813367	0.97%	rs17817449	https://biobankengine.stanford.edu/variant/16-53813367
All_variants	BMI	2	47.44%	2	Arm fat percentage (left)	5.76%	ADCY3	0.31%	16-53826034	0.28%	rs7187961	https://biobankengine.stanford.edu/variant/16-53826034
All_variants	BMI	2	47.44%	3	Body fat percentage	5.64%	DNMT3A	0.30%	2-417167	0.27%	rs62106258	https://biobankengine.stanford.edu/variant/2-417167

Table S5: GREAT enrichment analysis for BMI

Biological characterization of driving non-coding and coding variants of the key components for BMI with the genomic region enrichment analysis tool (GREAT) using the all variants dataset. The results of the enrichment analysis for MGI phenotype ontology, a manually curated genotype-phenotype relationship knowledgebase for mouse, is summarized by the key components. The two major summary statistics from GREAT, binomial fold and binomial p-value, are shown. Abbreviation. BFold: binomial fold, BPval: binomial p-value.

Here, I'm showing the first 3 lines of the table. The full table is in Excel file.

PC rank	PC	Term	BFold	BPval
1	PC2	brachypodia	9.05	1.40E-23
1	PC2	abnormal pancreas topology	8.13	8.80E-12
1	PC2	abnormal urine catecholamine level	7.32	5.19E-18

Table S6: GREAT enrichment analysis for MI

Biological characterization of driving non-coding and coding variants of the key components for MI with the genomic region enrichment analysis tool (GREAT) using the all variants dataset. The results of the enrichment analysis for MGI phenotype ontology, a manually curated genotype-phenotype relationship knowledgebase for mouse, is summarized by the key components. The two major summary statistics from GREAT, binomial fold and binomial p-value, are shown. Abbreviation. BFold: binomial fold, BPval: binomial p-value.

Here, I’m showing the first 3 lines of the table. The full table is in Excel file.

PC rank	PC	Term	BFold	BPval
1	PC22	artery occlusion	1.59E+01	1.14E-25
1	PC22	aortic sinus aneurysm	1.28E+01	3.88E-10
1	PC22	abnormal circulating phytosterol level	1.07E+01	6.38E-10

Table S7: GREAT enrichment analysis for gallstones

Biological characterization of driving non-coding and coding variants of the key components for gallstones with the genomic region enrichment analysis tool (GREAT) using the all variants dataset. The results of the enrichment analysis for MGI phenotype ontology, a manually curated genotype-phenotype relationship knowledgebase for mouse, is summarized by the key components. The two major summary statistics from GREAT, binomial fold and binomial p-value, are shown. Abbreviation. BFold: binomial fold, BPval: binomial p-value.

Here, I’m showing the first 3 lines of the table. The full table is in Excel file.

PC rank	PC	Term	BFold	BPval
1	PC72	abnormal circulating phytosterol level	1.15E+01	5.48E-11
1	PC72	decreased systemic vascular resistance	9.86E+00	9.78E-11
1	PC72	artery occlusion	9.51E+00	2.46E-12

Table S8: PheWAS analysis for rs114285050 (*GPR151*)

Phenome-wide association (PheWAS) analysis for rs114285050, a stop-gain variant in *GPR151*.

GBE phenotype code	Name	Case	-log_10 p-value	log(OR) or Beta	1.96 * SE(log(OR)) or 1.96 * SE(beta)
BIN1960	Fed-up feelings	136434	3.041	-0.09304	0.054978
INI48	Waist circumference	336659	7.599	-0.06544	0.02301
INI23100	Whole body fat mass	330970	6.87	-0.06872	0.025539
INI23128	Trunk fat mass	331295	6.835	-0.07053	0.026284
INI23120	Arm fat mass (right)	331422	6.816	-0.06863	0.025617
INI23099	Body fat percentage	331318	6.816	-0.05306	0.019816

INI23127	Trunk fat percentage	331314	6.79	-0.06356	0.023775
INI21002	Weight	336260	6.654	-0.06087	0.02303
INI23116	Leg fat mass (left)	331470	6.649	-0.05468	0.020698
INI23112	Leg fat mass (right)	331488	6.62	-0.05517	0.020933
INI21001	Body mass index (BMI)	336144	6.498	-0.06789	0.026029
INI23111	Leg fat percentage (right)	331491	6.341	-0.04201	0.016327
INI23124	Arm fat mass (left)	331362	6.317	-0.06587	0.025656
INI23115	Leg fat percentage (left)	331473	6.17	-0.04087	0.016123
INI23119	Arm fat percentage (right)	331445	5.424	-0.04689	0.019874
INI23123	Arm fat percentage (left)	331395	5.048	-0.04485	0.019796
INI49	Hip circumference	336620	4.649	-0.05669	0.026205
INI23126	Arm predicted mass (left)	331345	4.211	-0.03373	0.016499
INI23125	Arm fat-free mass (left)	331358	3.929	-0.03257	0.01658
INI23105	Basal metabolic rate	331502	3.923	-0.03368	0.017154
INI23117	Leg fat-free mass (left)	331454	3.423	-0.03063	0.016887
INI23118	Leg predicted mass (left)	331449	3.336	-0.02998	0.016776
INI23121	Arm fat-free mass (right)	331418	3.32	-0.02894	0.016241
INI23122	Arm predicted mass (right)	331413	3.176	-0.02808	0.016174
INI23102	Whole body water mass	331510	3.044	-0.02784	0.01644
INI23114	Leg predicted mass (right)	331480	3.019	-0.02812	0.016689

Table S9: PheWAS analysis for rs150090666 (*PDE3B*)

Phenome-wide association (PheWAS) analysis for rs150090666, a stop-gain variant in *PDE3B*.

GBE phenotype code	**Name**	**Case**	**-log_10 p-value**	**log(OR) or Beta**	**1.96 * SE(log(OR)) or 1.96 * SE(beta)**
HC269	high cholesterol	43054	4.457	-0.5904	0.279692
BIN4728	Leg pain on walking	28151	3.154	0.4366	0.252448
BIN2020	Loneliness, isolation	60153	3.098	0.2983	0.174322
INI49	Hip circumference	336620	10.75	0.2476	0.072167

INI23113	Leg fat-free mass (right)	331480	7.381	0.1293	0.046197
INI21002	Weight	336260	7.333	0.1769	0.063445
INI23114	Leg predicted mass (right)	331480	7.3	0.1276	0.045884
INI23128	Trunk fat mass	331295	7.079	0.1977	0.072304
INI23117	Leg fat-free mass (left)	331454	6.965	0.1259	0.046432
INI23118	Leg predicted mass (left)	331449	6.958	0.1249	0.046119
INI20015	Sitting height	336513	6.783	0.1454	0.054449
INI23105	Basal metabolic rate	331502	6.141	0.1193	0.047177
INI23127	Trunk fat percentage	331314	6.059	0.1641	0.065405
INI50	Standing height	336500	6	0.1266	0.050725
INI23100	Whole body fat mass	330970	5.895	0.1736	0.070227
INI23120	Arm fat mass (right)	331422	5.601	0.1692	0.070462
INI23124	Arm fat mass (left)	331362	5.255	0.1635	0.070521
INI23102	Whole body water mass	331510	5.107	0.1031	0.045198
INI23101	Whole body fat-free mass	331486	5.039	0.1021	0.045119
INI23099	Body fat percentage	331318	4.919	0.1217	0.054508
INI23123	Arm fat percentage (left)	331395	4.516	0.1158	0.054429
INI23119	Arm fat percentage (right)	331445	4.401	0.1146	0.054645
INI23116	Leg fat mass (left)	331470	4.208	0.1163	0.056918
INI23126	Arm predicted mass (left)	331345	4.189	0.09246	0.045374
INI23112	Leg fat mass (right)	331488	4.119	0.1162	0.057565
INI23125	Arm fat-free mass (left)	331358	4.061	0.09128	0.04559
INI23122	Arm predicted mass (right)	331413	3.746	0.085	0.044472
INI3062	Forced vital capacity (FVC)	309028	3.572	0.1001	0.053841
INI23130	Trunk predicted mass	331203	3.565	0.08357	0.045002
INI23129	Trunk fat-free mass	331234	3.508	0.08307	0.045158
INI23121	Arm fat-free mass (right)	331418	3.326	0.07965	0.044649
INI20151	Forced vital capacity (FVC), Best measure	255494	3.243	0.102	0.058016