

Supplementary Materials for

Sequencing of 640,000 exomes identifies GPR75 variants associated with protection against obesity

Parsa Akbari, Ankit Gilani, Olukayode Sosina, Jack A. Kosmicki, Lori Khrimian, Yi-Ya Fang, Trikaladarshi Persaud, Victor Garcia, Dylan Sun, Alexander Li, Joelle Mbatchou, Adam Locke, Christian Benner, Niek Verweij, Nan Lin, Sakib Hossain, Kevin Agostinucci, Jonathan V. Pascale, Ercument Dirice, Michael Dunn, Regeneron Genetics Center, DiscovEHR Collaboration, William E. Kraus, Svati H. Shah, Yii-Der I. Chen, Jerome I. Rotter, Daniel J. Rader, Olle Melander, Christopher D. Still, Tooraj Mirshahi, David J. Carey, Jaime Berumen-Campos, Pablo Kuri-Morales, Jesus Alegre-Díaz, Jason M. Torres, Jonathan R. Emberson, Rory Collins, Suganthi Balasubramanian, Alicia Hawes, Marcus Jones, Brian Zambrowicz, Andrew J. Murphy, Charles Paulding, Giovanni Coppola, John D. Overton, Jeffrey G. Reid, Alan Shuldiner, Michael Cantor, Hyun M. Kang, Goncalo R. Abecasis, Katia Karalis, Aris N. Economides, Jonathan Marchini, George D. Yancopoulos, Mark W. Sleeman, Judith Altarejos, Giusy Della Gatta, Roberto Tapia-Conyer, Michal L. Schwartzman, Aris Baras, Manuel A. R. Ferreira, Luca A. Lotta

Correspondence to: luca.lotta@regeneron.com and aris.baras@regeneron.com

This PDF file includes:

Banner Author List and Contributions Statements
Materials and Methods
Figs. S1 to S7

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S21 in an Excel document format

Regeneron Genetics Center Banner Author List and Contribution Statements

All authors/contributors are listed in alphabetical order.

RGC Management and Leadership Team

Goncalo Abecasis, Aris Baras, Michael Cantor, Giovanni Coppola, Andrew Deubler, Aris Economides, Katia Karalis, Luca A. Lotta, John D. Overton, Jeffrey G. Reid, Alan Shuldiner.
Contribution: All authors contributed to securing funding, study design and oversight. All authors reviewed the final version of the manuscript.

Sequencing and Lab Operations

Christina Beechert, Alex DeVito, Caitlin Forsythe, Erin D. Fuller, Zhenhua Gu, Michael Lattari, Joseph LaRosa, Alexander Lopez, M.S., Kia Manoochehri, Justin Marcovici, John D. Overton, Maria Sotiropoulos Padilla, Manasi Pradhan, Thomas D. Schleicher, Ricardo H. Ulloa, Emilia Weißenig, Louis Widom, Sarah E. Wolf.

Contribution: J.M., E.W., A.L., and J.D.O. performed and are responsible for sample genotyping. A.D., J.L., L.W., C.B., C.F., E.D.F., M.L., M.S.P., S.E.W., A.L., and J.D.O. performed and are responsible for exome sequencing. T.D.S., Z.G., A.L., and J.D.O. conceived and are responsible for laboratory automation. M.P., K.M., R.U., and J.D.O. are responsible for sample tracking and the library information management system.

Clinical Informatics

Nilanjana Banerjee, Michael Cantor, Dadong Li, Deepika Sharma.

Contribution: All authors contributed to the development and validation of clinical phenotypes used to identify study subjects and (when applicable) controls.

Genome Informatics

Xiaodong Bai, Suganthi Balasubramanian, Andrew Blumenfeld, Gisu Eom, Lukas Habegger, Alicia Hawes, Shareef Khalid, Jeffrey G. Reid, Evan K. Maxwell, William Salerno, Jeffrey C. Staples.

Contribution: X.B., A.H., W.S. and J.G.R. performed and are responsible for analysis needed to produce exome and genotype data. G.E. and J.G.R. provided compute infrastructure development and operational support. S.B., and J.G.R. provided variant and gene annotations and their functional interpretation of variants. E.M., J.S., A.B., L.H., J.G.R. conceived and are responsible for creating, developing, and deploying analysis platforms and computational methods for analyzing genomic data.

Analytical Genomics

Goncalo R. Abecasis, Josh Backman, Mathew Barber, Christian Benner, Shan Chen, Amy Damask, Lee Dobbyn, Manuel A. R. Ferreira, Arkopravo Ghosh, Lauren Gurski, Eric Jorgenson, Bindu Kalesan, Jack A. Kosmicki, Hyun Min Kang, Alexander Li, Nan Lin, Daren Liu, Adam Locke, Jonathan Marchini, Anthony Marcketta, Joelle Mbatchou, Arden Moscati, Colm O'Dushlaine, Charles Paulding, Jonathan Ross, Eli Stahl, Dylan Sun, Cristopher Van Hout, Kyoko Watanabe, Bin Ye, Andrey Ziyatdinov.

Contribution: G.R.A., M.A.R.F., E.J., B.K., H.M.K., J.Marchini, C.P., E.S. and C.V.H. are responsible for oversight and design of statistical analysis protocols and methods. G.R.A., J.B., M.B., C.B., M.A.R.F., A.G., L.G., J.A.K., H.M.K., A.L., A.Marcketta, J.Marchini, J.Mbatchou, J.R., D.S., K.W. and A.Z. are responsible for the design and implementation of key statistical methods, tools and pipelines to enable genetic association analyses. G.R.A., J.B., S.C., A.D., L.D., M.A.R.F., A.G., L.G., E.J., J.A.K., H.M.K., A.L., N.L., D.L., A.L., J.Marchini, A.Marcketta, J.Mbatchou, A.Moscati, C.O.D., E.S., D.S., C.V.H., B.Y. and A.Z. contributed key statistical genetic analysis and/or contributed to review and interpretation of statistical genetic findings.

Research Program Management

Marcus B. Jones, Michelle G. LeBlanc, Jason A. Mighty, Lyndon J. Mitnaul.

Contribution: All authors contributed to the management and coordination of all research activities, planning and execution. All authors contributed to the review process for the final version of the manuscript.

DiscovEHR Collaboration Banner Author List

Lance J. Adams, Jackie Blank, Dale Bodian, Derek Boris, Adam Buchanan, David J. Carey, Ryan D. Colonie, F. Daniel Davis, Dustin N. Hartzel, Melissa Kelly, H. Lester Kirchner, Joseph B. Leader, David H. Ledbetter, J. Neil Manus, Christa L. Martin, Michelle Meyer, Tooraj Mirshahi, Matthew Oetjens, Thomas Nate Person, Christopher D. Still, Natasha Strande, Amy Sturm, Jen Wagner, Marc Williams.

Materials and Methods

Participating cohorts

Discovery genetic association studies were performed in the United Kingdom (UK) Biobank (UKB) cohort (1), in the MyCode Community Health Initiative cohort from the Geisinger Health System (GHS) (2) and in the Mexico City Prospective Study (MCPS) (3). The UKB is a population-based cohort study of people aged between 40 and 69 years recruited through 22 testing centers in the UK between 2006-2010. A total of 428,719 European ancestry participants with available whole-exome sequencing and clinical phenotype data were included (**Table S1**). UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC; 11/NW/0382), which covers the UK. The GHS MyCode study is a health system-based cohort of patients from Central and Eastern Pennsylvania (USA) recruited in 2007-2019. A total of 121,061 European ancestry participants with available whole-exome sequencing and clinical phenotype data were included (**Table S1**). The GHS MyCode study was approved by the Geisinger Institutional Review Board (2006-0258). The MCPS is a cohort study of people aged ≥ 35 years recruited from two contiguous urban districts in Mexico City in 1998-2004. The study design and clinical characteristics of participants in MCPS has been described in detail in previous publications (3, 4). A total of 95,846 individuals of Admixed American ancestry with available whole-exome sequencing and clinical phenotype data were included (**Table S1**). The MCPS study was approved by the Mexican Ministry of Health, the Mexican National Council for Science and Technology, and the University of Oxford.

We further estimated the association with BMI of *GPR75* predicted loss-of-function (pLOF) variants in an additional 91,328 exomes not included in the discovery set. These included participants of non-European ancestries from the UK Biobank (UKB, N=12,321) (1), and participants in the Mount Sinai BioMe cohort (SINAI, N=21,143), the University of Pennsylvania Medicine BioBank (PMBB; N=7,519), the Duke Catheterization Genetics (CATHGEN) cohort (DUKE; N=8,171) (5), the Taiwanese Chinese from Taiwan Metabochip consortium (TAICHI; N=11,223) (6), the Dallas Heart Study (DHS; N=2,088) (7) and the Malmö Diet and Cancer Study (MALMO; N=28,863) (8). All participants provide informed consent for participation in these studies.

Phenotype definitions

Body mass index was calculated as weight in kilograms divided by the square of height in meters on the basis of anthropometric measurements taken at one of the study visits. BMI measured at the baseline visit was the outcome variable in UKB and MCPS, while median BMI from clinical encounters present in the GHS database was the outcome variable for GHS consistent with previous studies (9). BMI categories were defined on the basis of the World Health Organization classification (10). BMI values were transformed by the inverse standard normal function, applied within each ancestry group and separately in men and women. Body weight differences were calculated for a person 170 cm tall. Overall and regional body lean and fat masses, percentages and body-surface normalized indices were measured by bioelectrical impedance in the UKB cohort. At the baseline visit, UKB also collected self-reported information comparative body size at age 10 by asking the multiple choice question: "When you were 10 years old, compared to average would you describe yourself as: thinner, plumper, about average, do not know, prefer not to answer?".

Genotype data

High coverage whole exome sequencing was performed at the Regeneron Genetics Center as previously described in detail (9, 11) and as summarized below. NimbleGen probes (VCRome; for part of the GHS cohort) or a modified version of the xGen design available from Integrated DNA Technologies (IDT; for the rest of GHS and other cohorts) were used for target sequence capture of the exome. A unique 6 base pair (bp) barcode (VCRome) or 10 bp barcode (IDT) was added to each DNA fragment during library preparation to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled prior to exome capture. Sequencing was performed using 75 bp paired-end reads on Illumina v4 HiSeq 2500 (for part of the GHS cohort) or NovaSeq (for the rest of GHS and other cohorts) instruments. Sequencing had a coverage depth (ie, number of sequence-reads covering each nucleotide in the target areas of the genome) sufficient to provide greater than 20x coverage over 85% of targeted bases in 96% of VCRome samples and 20x coverage over 90% of targeted bases in 99% of IDT samples. Data processing steps included sample de-multiplexing using Illumina software, alignment to the GRCh38 Human Genome reference sequence including generation of binary alignment and mapping files (BAM), processing of BAM files (eg, marking of duplicate reads and other read mapping evaluations). Variant calling and annotation were based on the GRCh38 Human Genome reference sequence and Ensembl v85 gene definitions using the snpEff software. The snpEff predictions that involve protein-coding transcripts with an annotated start and stop were then combined into a single functional impact prediction by selecting the most deleterious functional effect class for each gene. The hierarchy (from most to least deleterious) for these annotations was frameshift, stop-gain, stop-loss, splice acceptor, splice donor, stop-lost, in-frame indel, missense, other annotations. Predicted LOF genetic variants included (a) insertions or deletions resulting in a frameshift, (b) insertions, deletions or single nucleotide variants resulting in the introduction of a premature stop codon or in the loss of the transcription start site or stop site, and (c) variants in donor or acceptor splice sites. Missense variants were classified for likely functional impact according to the number of *in silico* prediction algorithms that predicted deleteriousness using SIFT (12), Polyphen2_HDIV (13) and Polyphen2_HVAR (13), LRT (14) and MutationTaster (15). For each gene, the alternative allele frequency (AAF) and functional annotation of each variant determined inclusion into these 7 gene burden exposures: (1) pLOF variants with AAF < 1%; (2) pLOF or missense variants predicted deleterious by 5/5 algorithms with AAF < 1%; (3) pLOF or missense variants predicted deleterious by 5/5 algorithms with AAF < 0.1%; (4) pLOF or missense variants predicted deleterious by at least 1/5 algorithms with AAF < 1%; (5) pLOF or missense variants predicted deleterious by at least 1/5 algorithms with AAF < 0.1%; (6) pLOF or any missense with AAF < 1%; (7) pLOF or any missense variants with AAF < 0.1%.

SNP array genotyping was performed in the UKB as previously described (16). In GHS, genotyping was performed using the Human Omni Express Exome array (OMNI) and the Global Screening array (GSA). In MCPS, genotyping was performed using the GSA array.

In vitro studies of GPR75 variants

In vitro validation studies were performed for two *GPR75* pLOF genetic variants (Ala110fs and Gln234*) that were (a) individually associated with lower BMI ($p < 0.05$) and (b) had at least 10 heterozygous carriers. Briefly, pcDNA 3.1 plasmids encoding for N-terminally HA-tagged wild-type, Ala110fs and Gln234* *GPR75* were transiently transfected using Eugene 6 (Promega) in HEK293 cells. HEK293 and HEK293T cell lines were purchased from ATCC and maintained in the Regeneron Tissue Culture Core. Their identity was confirmed by STR profiling. *In vitro*

assays included mRNA and protein analysis by Taqman and Western Blotting, and protein localization by fluorescence-activated cell sorting and immunofluorescence.

Cell culture, plasmids and cell transfection: HEK293 cells were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, and antibiotics (50 units/mL penicillin and 50 µg/mL streptomycin; Thermo Fisher Scientific). pcDNA 3.1 plasmids encoding for N-terminus HA-tagged GPR75 wild-type, Ala110fs and Gln234* were synthesized by GenScript (USA). Cells at approximately 60-70% confluence were transiently transfected with plasmid containing cDNA encoding HA-tagged *GPR75* wild-type, Ala110fs and Gln234* and green fluorescent protein control plasmid using FuGENE 6 (Promega) according to the manufacturer's protocol (Promega Literature: # TM350), at a ratio of 1µg DNA:5µl FuGENE transfection reagent. After 48 hours, cells were washed with 1x DPBS (Thermo Fisher Scientific) and collected for downstream analysis.

Western blotting: Transfected HEK293 were collected in RIPA buffer for cell lysis and 5-10 µg of protein was loaded per sample. The following primary antibodies were used: HA (mouse monoclonal, Sigma cat. Cat# H3663) and GAPDH 14C10 (Rabbit mAb, Cell Signaling Cat# 2118). The appropriate LI-COR secondary IRDye antibodies (anti-rabbit [926-32211] and anti-mouse [926-32210]) were used to detect and quantify immunoblots using a LI-COR Odyssey Infrared Imaging System (LI-COR, Lincoln, NE).

Flow cytometry: Cells were washed once with 1x DPBS (Cat# 14190144). Cell Dissociation Buffer (Cat# 13150016) was added and cells were incubated at 37°C for 3 minutes. Cells were re-suspended in culture media and centrifuged at 200xg for 5 minutes. Cells were washed twice with DPBS, re-suspended in DPBS, aliquoted and stained with Live/Dead Blue Fixable Viability Dye (Thermo Fisher Scientific) at room temperature for 15 minutes with no light. Cells were washed twice with DPBS - all washes centrifuged at 400xg for 5 mins and all staining in the dark. Cells were treated with human Fc Block (BD Biosciences) in MACS buffer (Miltenyi Biotec) for 15 minutes at 4°C and stained with alexa fluor anti-HA.11 epitope tag antibody (Cat# 682404) at 1:100 dilution in MACS buffer for 30 minutes at 4°C. Cells were washed with MACS buffer and fixed with CytoFix (BD Biosciences) for 15 min at 4°C. Cells were washed twice with MACS buffer, filtered and FACS was performed on a CytoFLEX (Beckman Coulter). Data was analyzed using FlowJo 10.6.2 (Becton Dickinson & Company).

Immunofluorescence assays: For immunofluorescence assays, cells were seeded onto open 8-well µ-Slides (chamber slide) with a glass bottom (Ibidi, cat# 80827) at a density of 14,000 cells/well. At 48h post-transfection, cells were fixed in ice-cold 4% PFA for 10min at RT and washed 3x with ice-cold 1x DPBS (all subsequent wash steps were done 3 times ice-cold 1x DPBS for 5 min per wash). Cells that were not permeabilized were blocked for 1h using 10% normal donkey serum (NDS) (Jackson Immunoresearch Laboratories, # 017-000-121), while permeabilized cells were blocked in 10% NDS with 0.1% Triton X-100; these were subsequently used as staining buffers for non-permeabilized and permeabilized cells, respectively. Cells were incubated with 1:500 (non-permeabilized) or 1:3000 (permeabilized) anti-HA antibody (Sigma, Cat# H3663) for 1h at RT, washed and then incubated for 1h with 1:1000 alexa fluor 594-conjugated anti-mouse secondary antibody (Thermo Fisher Scientific, Cat# A-21203). Wells were then washed, and slides were mounted with ProLong® Gold Antifade Reagent with DAPI (Cell Signaling, #8961). Slides were imaged using Zeiss confocal LSM880.

Quantitative real-time polymerase chain reaction: RNA was extracted from transfected HEK293 using TRIzol reagent and following the manufacturer's instructions (Thermo Fisher Scientific). Genomic DNA was removed using MagMAX™Turbo™DNase Buffer and TURBO DNase (Ambion by Life Technologies). mRNA (up to 2 µg) was reverse-transcribed into cDNA

using SuperScript® VILO™ Master Mix (ThermoFisher Scientific). *GPR75* cDNA was amplified with the PowerUp SYBR Green Master Mix (Thermo Fisher Scientific) using the QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). *ACTINB* housekeeping gene was used as the internal control gene to normalize cDNA input differences. Expression of *GPR75* was calculated relative to *ACTINB* housekeeping gene.

Primer sequences were as follows:

GPR75-forward: 5'-GCTTGTGGCCCAAGTCATTC-3'

GPR75-reverse: 5'-GAGTGTTGATGGGGGTCGAG-3'

ACTINB-forward: 5'-CACCATTGGCAATGAGCGGTTC-3'

ACTINB-reverse: 5'-AGGTCTTTGCGGATGTCCACGT-3'

In vitro studies of *MC4R* and *GIPR* variants

GIPR cDNA or *MC4R* cDNA constructs containing N-terminal MYC tag (Genscript) in pRG980 vector (Regeneron) were used throughout the study. Site-directed mutagenesis was performed using QuikChange II XL kit (Agilent Technologies, Catalog #200516) according to the manufacturer's protocols and verified by DNA sequencing. HEK293T cells were transiently transfected with wild-type or mutant *MC4R* or *GIPR*. Ligand-induced Gs signaling was measured by cyclic adenosine monophosphate response element (CRE) dependent luciferase assay, while Gq signaling was measured by nuclear factor of activated T-cells (NFAT) dependent luciferase assay. A total of 20,000 HEK293T cells were seeded in 96 well black poly-D-lysine plates and transfected the next day. On the transfection day, cells were then transfected with pGL4.29[luc2P/CRE/Hygro] plasmid (Promega, E8471) or pGL4.30[luc2P/NFAT-RE/Hygro] (Promega, E8481), and plasmid encoding either wild-type or mutant *MC4R* or *GIPR* by using Fugene6 (Promega). After 48 hours, the transfection media was replaced with assay media (Opti-Mem 1% BSA, 0.1% FBS). Cells were then stimulated with alpha-melanocortin stimulating hormone (alpha-MSH) or glucose-dependent insulintropic polypeptide (GIP), at various concentrations. After 6 hours, luciferase activity was quantified using One-Glo (Promega) reagent, and luminescence was measured with an EnVision plate reader. Percentage Max relative lights units of wild-type = 100*(Max relative light units [RLU] for the genetic variant / Max RLU wild-type) were calculated, and normalized data were merged and presented as sum curves \pm standard error of the mean.

Beta-arrestin 1/2 recruitment for wild-type or mutant *MC4R* or *GIPR* was assayed using a NanoBiT protein-protein interaction assay (Promega, M2015). Wild-type or mutant *MC4R* or *GIPR* containing C-terminal LgBiT, and beta-arrestin 1/2 were cloned into a SmBiT TK-Neo Flexi® vector. HEK293T cells were seeded in poly-D-lysine-coated, black 96-well plates, and transiently transfected with each of the two constructs using Fugene 6. Following 48hr transfection, medium was replaced with assay media (Opti-Mem 1% BSA, 0.1% FBS). Nano-Glo Live Cell Assay System (Promega, N2013) was added and cells were equilibrated while basal luciferase activity was measured for 5 min (1 min intervals). Subsequently, cells were stimulated with alpha-MSH, GIP, or the *MC4R* antagonist agouti-related peptide (AgRP). AgRP experiments were carried out in the presence of a fixed concentration of alpha-MSH. Chemiluminescent signal was quantified for 30 min (1 min intervals). The area under the curve (AUC) was calculated for each wild-type or mutant protein. For data normalization, the Max AUC from wild-type was set as 100%. Results are from at least 3 independent experiments. Normalized data were merged and presented as sum curves \pm standard error of the mean.

HEK293T cells were seeded in 6-well plates (Corning, 3506) and transfected with wild-type or mutant *MC4R* or *GIPR* constructs. Forty-eight hours after transfection, cells were suspended and washed three times with PBS with 1% BSA. Cells were incubated on ice for 30 min with mouse anti-Myc mAb (Cell Signaling Technology, #2276). After two wash steps, cells were incubated with Goat Anti-Mouse IgG APC (Jackson) for 30 min. Following two washes with PBS with 1% BSA, cells analyzed on FACS Accuri C6 (BD Biosciences). Flow cytometry data analysis and mean fluorescence intensity (MFI) values were calculated by FlowJo analysis software (Tree Star) on live-gated cells (>80% live cells). Fold change is calculated by mean MFI of the stained cell type / same cell type unstained.

Mouse models

The genetically engineered *Gpr75*^{-/-} mouse strain was created using Regeneron's VelociGene® technology (16, 17). Briefly, C57Bl/6NTac embryonic stem cells were targeted for ablation of the entire *Gpr75* locus, beginning immediately after the endogenous ATG and ending at the *Gpr75* stop codon. Ablation was achieved using a modified bacterial artificial chromosome (BAC) targeting construct such that BAC *Gpr75* sequence was replaced with a self-deleting, floxed lacZ reporter cassette containing a neomycin resistance gene under the control of the human *UBC* (ubiquitin) promoter. The deletion was engineered such that the *lacZ* reporter was inserted in frame immediately after the endogenous ATG. This construct was electroporated into C57Bl/6NTac embryonic stem cells. Following selection with neomycin, correctly targeted clones were identified by TaqMan analysis and microinjected into 8-cell Swiss Webster embryos (Charles River Laboratories), resulting in F0 VelociMouse® fully derived from the injected modified embryonic stem cells (17).

Heterozygous *Gpr75*^{+/-} mice were bred to generate age-matched wild type *Gpr75*^{+/+}, heterozygous *Gpr75*^{+/-} and knock-out *Gpr75*^{-/-} littermates that were used for experimentation. Male and female mice (N=56; 27 males, 29 females; age range, 10-12 weeks; mean body weight, 20.9 g, standard deviation, 2.1 g) were housed in static cages (≤4 mice per cage) with free access to food and water and fed either control chow diet or a high-fat diet (HFD; Envigo, #TD.03584, Huntingdon, UK) for 14 weeks. The control diet consisted of the following components in amounts represented by percent kilocalories (kcal): fat: 13.4%, carbohydrate: 58.0%, and protein: 28.7%. HFD consisted of the following components in percent kilocalories: fat: 58.4%, carbohydrate: 26.6%, and protein: 15.0%.

All animals were monitored for changes in body weight on a weekly basis. No adverse reactions or signs of discomfort were observed during the course of the experiment. No significant differences were identified between male and female mice in the measured parameters. The data presented in the manuscript are those of male and female mice combined.

Fasting blood glucose was measured after overnight fasting before and at the end of the diet-feeding period. An intra-peritoneal glucose tolerance tests were performed at the end of the experiment. Followed by an overnight fasting period, glucose (2 g/kg) was administered to each mouse by intra-peritoneal injection. The tip of the tail of each mouse was scratched to draw blood. Blood samples were collected at 0, 30, 60, 90, and 120 min, and glucose was measured using Contour blood glucose monitoring system (Bayer, Whippany, NJ). After these measurements, blood was collected in capillary tubes and used for insulin measurements. Blood was centrifuged at 2,000 rpm for 15 min to separate the plasma. Ultra-Sensitive Mouse Insulin ELISA kit (Crystal Chem. #90080, Elk Grove Village, IL) was used to quantify plasma insulin levels as per manufacturer's instructions. Plasma levels of leptin and adiponectin were measured by ELISA according to the manufacturer's instructions (Abcam, Cambridge, MA; #ab100718

and #ab108785 for leptin and adiponectin, respectively). All protocols were approved by the Institutional Animal Care and Use Committee in accordance with the National Institutes of Health Guidelines for the Care and Use of Laboratory Animals.

Statistics: The Graph Pad Prism version 9 software was used for statistical analysis. Significance of difference in mean values was estimated using repeated measures two-way ANOVA followed by Tukey's *post hoc* multiple comparison test.

Statistical analysis

Overview: We estimated the association with BMI of genetic variants or their gene burden by fitting mixed-effects regression models using BOLT-LMM v2.3.4 (19) or REGENIE v1.0 (20). These approaches account for relatedness and population structure by estimating a polygenic score using genotypes from across the genome. Then, the association of genetic variants or their burden is estimated conditional upon that polygenic score along with other covariates. To ensure that burden associations were statistically independent of BMI-associated common genetic variants, we further adjusted the exome association analyses for sentinel common variants ($AAF \geq 1\%$) identified by fine-mapping genome-wide associations of common alleles with BMI as described below. Results across cohorts were pooled using inverse-variance weighted meta-analysis.

Association with BMI of the burden of rare nonsynonymous variants identified by exome-sequencing: In the primary analysis of this study, we estimated the association with BMI of the burden of rare nonsynonymous variants in each gene by fitting mixed-effects regression models adjusted for a polygenic score that approximates a genomic kinship matrix using BOLT-LMM v2.3.4 (19) or REGENIE v1.0 (20). Analyses were further adjusted for age, age², sex, an age-by-sex interaction term, experimental batch-related covariates, and genetic principal components. We adjusted for 10 common-variants derived principal components in the UKB and GHS, while we used 10 common-variants derived principal components as well as 10 rare-variants derived principal components in the admixed MCPS study. Ensuring that rare variants associations are independent of nearby trait-associated common alleles is essential for the correct causal variant and gene attribution in studies focused on exome variation (21). To ensure that burden associations were statistically independent of BMI-associated common genetic variants, we adjusted the exome-wide gene burden association analyses for common variants identified by fine-mapping genome-wide associations of common alleles with BMI (listed in **Table S18**). In line with previous similar studies (22, 23), the exome-wide level of statistical significance for the gene burden analysis was defined as $p < 3.6 \times 10^{-07}$, a Bonferroni correction for 20,000 genes and seven variant selection models.

Rare nonsynonymous single variant analysis: In a secondary analysis, we estimated the association with BMI of individual rare nonsynonymous variants (minor allele frequency < 1% and minor allele count > 25) identified by exome sequencing. We used the same analytical approach as with the gene burden analysis, including adjustment for BMI-associated common variants identified by fine-mapping. This step is essential to confirm the conditionally-independent nature of the association of these rare variants (21). In this analysis, we used a statistical threshold for association of $p < 5 \times 10^{-08}$, a Bonferroni correction for ~1,000,000 rare nonsynonymous variants tested in this analysis which is also the conventional threshold for genome-wide significance used in GWAS (24).

GWAS of common variants and fine-mapping: We identified BMI-associated common variants by performing a genome-wide association study including over 12 million common-to-low-frequency genetic variants imputed using the Haplotype Reference Consortium panel (25).

In the GHS study, imputation was performed separately in samples genotyped with the Illumina Human Omni Express Exome array (OMNI set) and the Global Screening array (GSA set). Dosage data from imputed variants were then merged across the two GHS sets, to obtain a combined dataset for association analysis. Genome-wide association analyses were performed in the GHS, UKB and MCPS cohorts separately by fitting mixed-effects linear regression models using BOLT-LMM (19) or REGENIE (20). Results from the UKB and GHS analyses were then combined by inverse variance-weighted meta-analysis to obtain a genome-wide meta-analysis in the European subset of the discovery cohorts. To identify conditionally-independent genetic association signals driven by common variants, we performed fine-mapping at genomic regions harboring genetic variants associated with BMI at the genome-wide significance threshold of $p < 5 \times 10^{-08}$ using the FINEMAP software (26). Linkage disequilibrium was estimated using genetic data from the exact set of individuals included in the genome-wide association analyses. Fine-mapping was performed separately in the meta-analysis of the European ancestry GHS and UKB cohorts and in the Admixed American ancestry analysis in the MCPS cohort. Fine-mapping identifies independent common variant signals and assigns a posterior probability of causal association for variants linked to a given independent signal. For each locus that was fine-mapped, we identified the 95% credible variant set, i.e. the minimal set of variants that capture the 95% posterior probability of causal association. We also defined the sentinel variant as the variant with the highest posterior probability of causal association at each given independent signal.

Transethnic meta-analysis of GWAS with MANTRA: To estimate the strength of association across ancestries for fine-mapped variants identified in ancestry specific analyses, transethnic meta-analysis was performed using GWAS summary statistics from the UKB, GHS, and MCPS cohorts using the MANTRA (Meta-ANalysis of Transethnic Association studies) software (27).

Prioritization of likely effector genes at fine-mapped loci: we used physical proximity, common nonsynonymous variants, and expression quantitative trait loci (eQTLs) data to prioritize likely effector genes at GWAS fine-mapped loci. For the physical proximity criterion, we prioritized the gene nearest to the sentinel variant of a fine-mapped signal. For the common nonsynonymous variant criterion, we considered whether the sentinel variant or one of its proxies ($R^2 > 0.8$) was a nonsynonymous variant in a gene. For the eQTL data, we performed colocalization analyses using summary association statistics from the v8 release of GTEx and summary association statistics from our common variants GWAS of BMI. We separately used GTEx results based on all individuals (ALL) or GTEx results based only on individuals of European ancestry (EUR). First, for each BMI sentinel variant identified in our fine-mapping analysis, we looked for associations with gene expression in any one of 49 tissues in GTEx (p-value thresholds for gene expression analysis were based on gene level thresholds derived in GTEx). For each variant, gene and tissue combination, we performed colocalization using the coloc Bayesian framework (28). We defined a 500kb region around the sentinel variant and considered only variants present in both datasets (BMI GWAS and GTEx) and with minor allele frequency above 1% in both datasets for the colocalization analysis. Furthermore, we used the default priors in the coloc package and conducted sensitivity analyses (29) for each colocalized result to see how robust they were to the specification of the priors. We defined each gene-region combination as being colocalized if its posterior probability of having a shared single causal variant for both BMI and gene expression, ie. PP4, was greater than 0.8 (80%).

Generation of a genome wide-polygenic score for BMI in the UKB study: A polygenic score capturing predisposition to higher BMI due to over 2.5 million common variants was generated

using the LDpred software (30) with a rho parameter value of 1, from the results of a previous large genome-wide association study in an independent dataset (31).

Phenome-wide analysis for GPR75 predicted loss-of-function variants: We undertook a phenome-wide analysis of the association of pLOF variants in *GPR75* with hundreds of continuous traits or disease outcomes in the GHS and UKB studies. To increase power, we performed inverse-variance weighted meta-analysis using the METAL software (32) to combine association results across GHS and UKB for disease outcomes available in both studies. To minimize the risk of false positive associations due to the small number of variant carriers, we excluded outcomes with ≤ 25 individuals carrying *GPR75* pLOF genetic variants, determined based on individuals with a non-missing phenotype for continuous traits, or based on affected individuals for binary disease outcomes. After these exclusions, results were available for 2,173 outcomes. To control for the number of statistical tests performed, associations were considered statistically significant if the association p-value met a Bonferroni correction for 2,173 tests, that is $p < 2.3 \times 10^{-5}$ (corresponding to a p-value threshold of 0.05 divided by 2,173 statistical tests).

Continuous traits and disease outcomes were defined as described below. In the UKB study, for continuous traits, the values of biomarker, imaging variables or other continuous traits measured during one of the UKB visits or their averages within a given study visit or across study visits were used as outcomes. For binary disease outcomes, case status definition required one or more of the following criteria to apply (a) self-reported disease status or use of medication at digital questionnaire or interview with a trained nurse or (b) EHR of inpatient encounters from the UK National Health Service Hospital Episode Statistics database coded using the ICD-10 coding system. For each binary outcome, controls were individuals without any of the criteria for case definition. In the GHS study, for binary disease outcomes, case status definition required one or more of the following criteria to apply: (1) a problem-list entry of the ICD-10 diagnosis code, (2) an inpatient hospitalization-discharge ICD-10 diagnosis code, or (3) an encounter ICD-10 diagnosis code entered for 2 separate outpatient visits on separate calendar days. Controls were individuals without any of the criteria for case definition. Individuals were excluded if they had the relevant ICD-10 code associated with only one outpatient encounter. For continuous traits, data cleaning was performed by removing non-physiologic lab values, invalid or contaminated specimens, and those that were over 5x upper limit of normal. Then the minimum, median, and maximum laboratory result values over the duration of follow-up were derived for each patient and used as outcomes.

Tissue Enrichment analysis: Tissue enrichment analyses were performed using annotation data based on gene expression values from the V8 data freeze from GTEx (<https://www.gtexportal.org/home/faq#citePortal>) (33). Annotation for each gene is obtained as follows. For each tissue in the GTEx V8 Freeze, we used the same QC filters as GTEx to remove outlier samples and null genes. With the filtered dataset, we performed a between sample normalization using the TMM approach (34) and obtained the median expression level across individuals for each gene within a given tissue. With these we calculated $z\text{-scores}_{\text{Tissue}}$ within each tissue for each gene, where the median and the median absolute deviation (adjusted for asymptotically normal consistency) were used to center and scale the normalized gene-expression values. Hence, each $z\text{-score}_{\text{Tissue}}$ accounts for the deviation of each gene expression level from the median expression value, scaled by the variability seen in the tissue. Using these tissue specific $z\text{-scores}_{\text{Tissue}}$, we then calculated another $z\text{-score}_{\text{Gene}}$ per gene across tissues. For a given gene, tissues where the newly obtained $z\text{-scores}_{\text{Gene}}$ are at least 6 standard deviations away from the median $z\text{-scores}_{\text{Gene}}$ seen for that gene across tissues are then identified as tissues where we see an enhanced expression for the gene (ie. tissue-enhanced gene). Using this tissue-

enhanced gene definition, we grouped our gene burden association results by tissue and then transformed the gene-burden p-values to z-scores so that small p-values correspond to large z-scores. With the z-score we estimated and tested (one-sided) the average change in z-scores comparing genes in tissue of interest to genes not in the tissue of interest using the generalized estimating equation approach (35) to account for correlated z-scores across gene-masks.

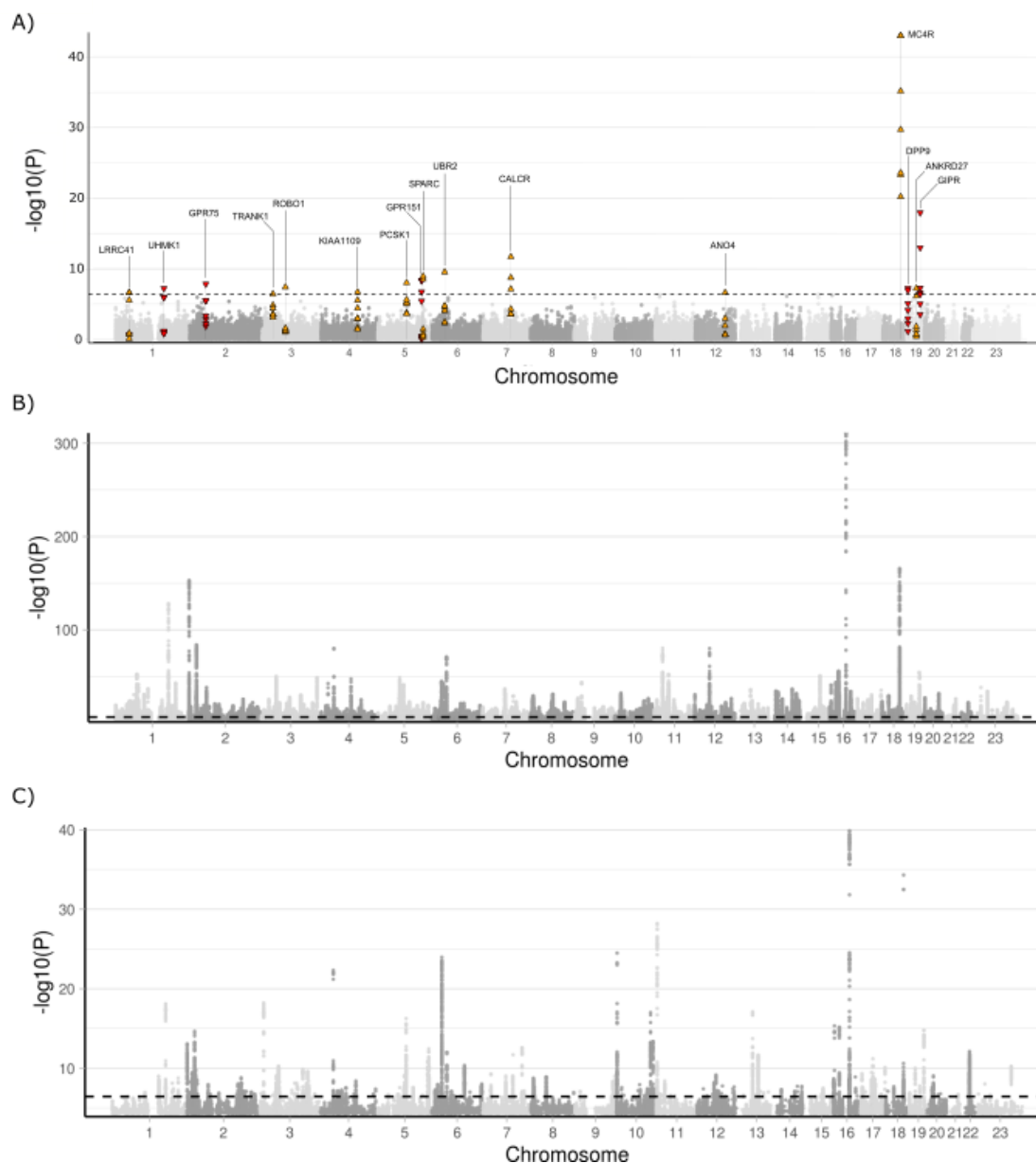


Fig. S1. Manhattan plots for the association with BMI of gene burden (exome sequencing) and common variants (GWAS of imputed variants) in the discovery cohorts. *A* Association of the gene burden of rare nonsynonymous alleles with body mass index in the discovery exome-sequencing analysis. Orange triangles pointing upwards indicate gene burden associations with higher body mass index, while red triangles pointing downwards indicate gene burden associations with lower body mass index. *B* Association of common variants in a GWAS of European ancestry individuals from the UKB and GHS cohorts. *C* Association of common variants in a GWAS of admixed American individuals from the MCPS cohort.

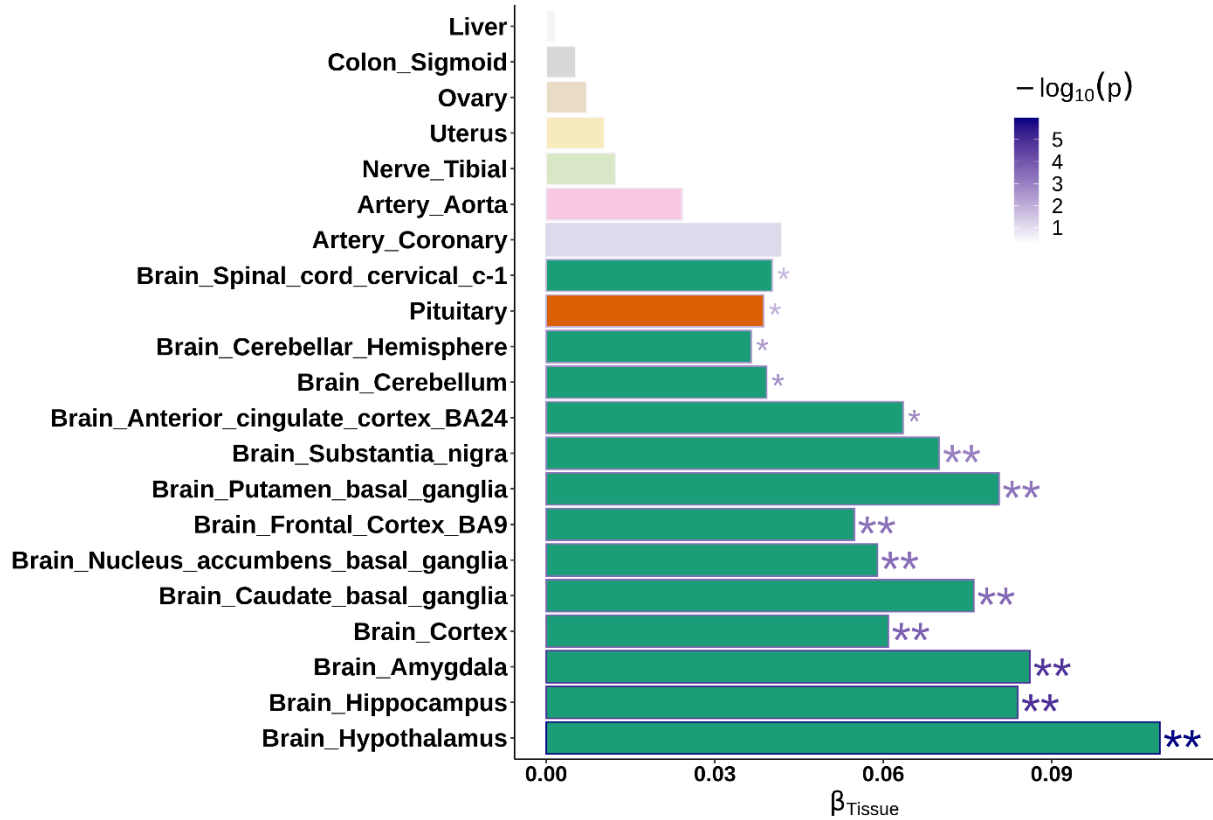


Fig. S2. Tissue enrichment analysis for body mass index gene burden associations.

Results from tissue enrichment analyses, based on gene burden tests from the BMI exome analysis, using GTEx V8 data. Shown on the plot are tissues (y-axis) whose enhanced genes, on average, have a stronger association (x-axis) with BMI.

* $p < 0.05$

** $p < 0.001$ (Bonferroni correction for number of tests)

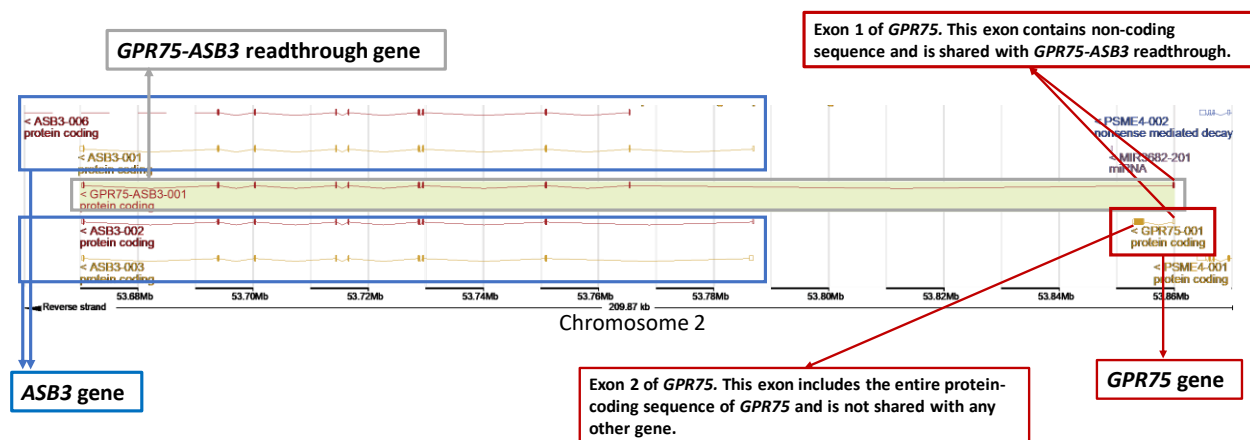


Fig. S3. *GPR75*, *ASB3* and *GPR75-ASB3* genes. The Figure shows the gene model and chromosomal locations for the *GPR75*, *GPR75-ASB3* and *ASB3* genes. *GPR75* shares exon 1, containing non-coding sequence, with the *GPR75-ASB3* readthrough gene. Exon 2 of *GPR75*, containing its entire coding sequence, is exclusive to the *GPR75* gene and is not shared with any other gene. *ASB3* and *GPR75-ASB3* share several exons with each other but not with *GPR75*.

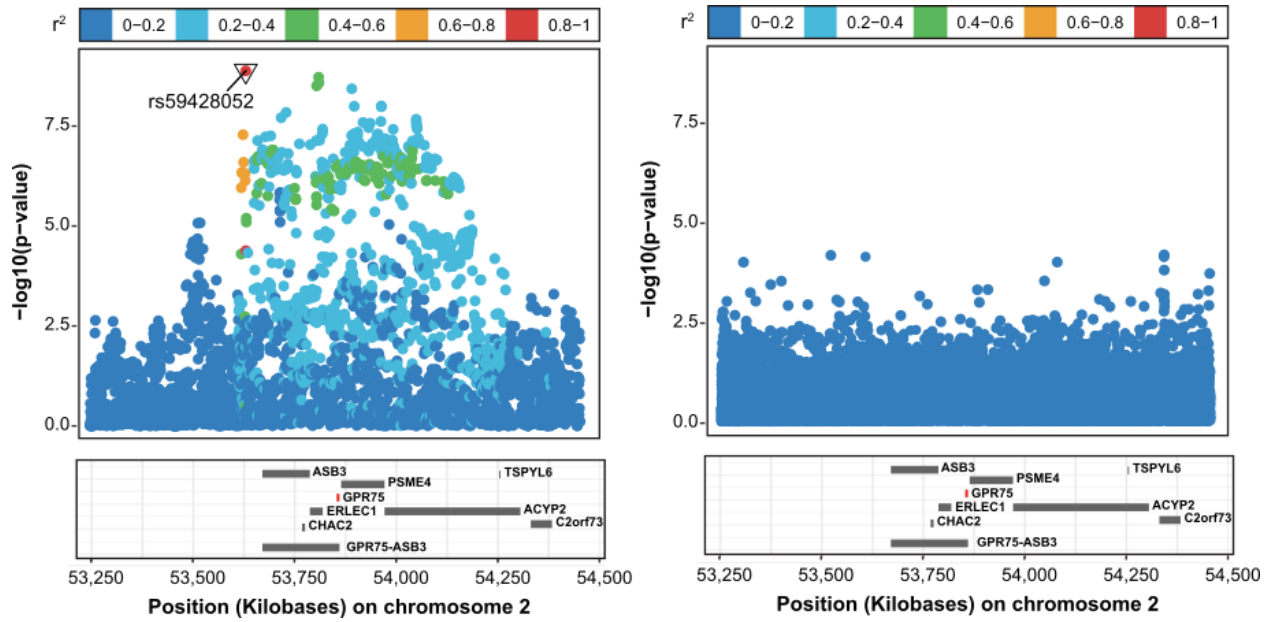


Fig. S4. Associations with BMI for common variants at the *GPR75* locus. Results from GWAS analyses in European ancestry individuals from UKB and GHS are shown on the left panel and those from GWAS analyses in admixed Americans from the MCPS cohort on the right. The sentinel variant in the GWAS of European individuals is highlighted (rs59428052), there were no genome-wide significant associations in Admixed Americans ($p < 5 \times 10^{-8}$).

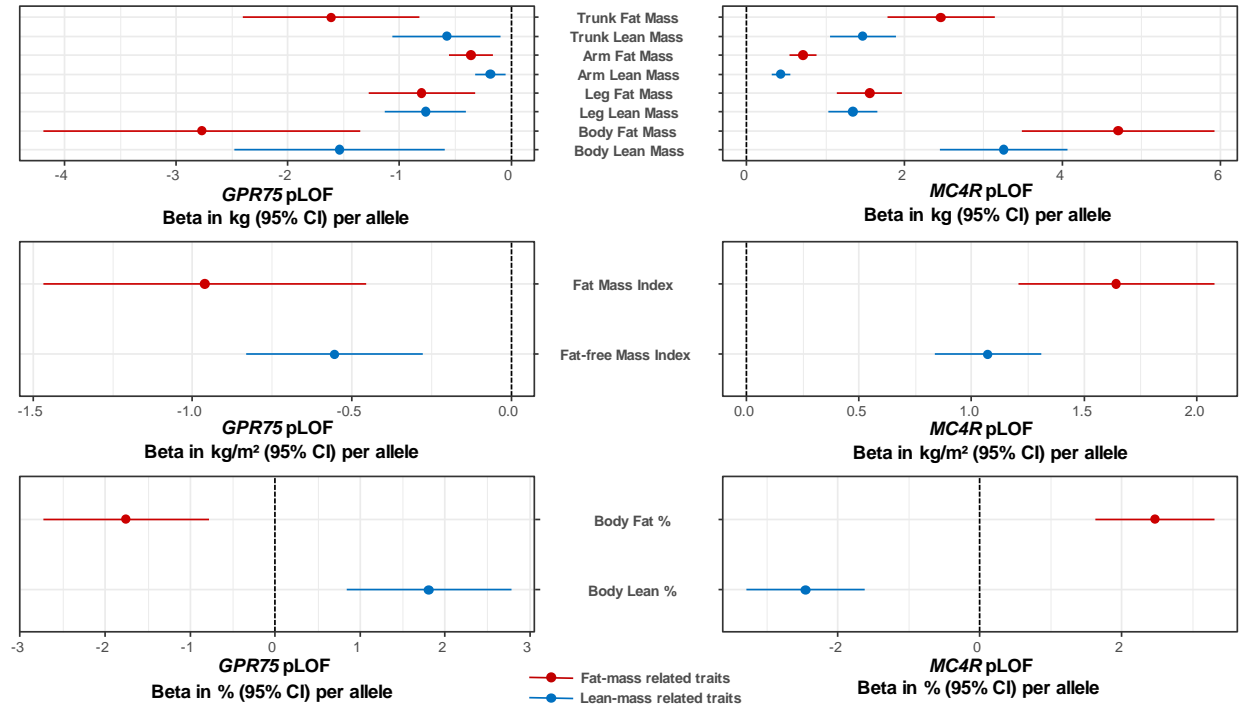


Fig. S5. Association of pLOF variants in *GPR75* and *MC4R* with body fat and lean mass indices estimated by bioelectrical impedance. Association analyses we performed in 423,418 participants of the UK Biobank study who underwent whole exome sequencing and bioelectrical impedance measurements.

Abbreviations: pLOF, predicted loss of function; kg, kilograms; CI, confidence interval.

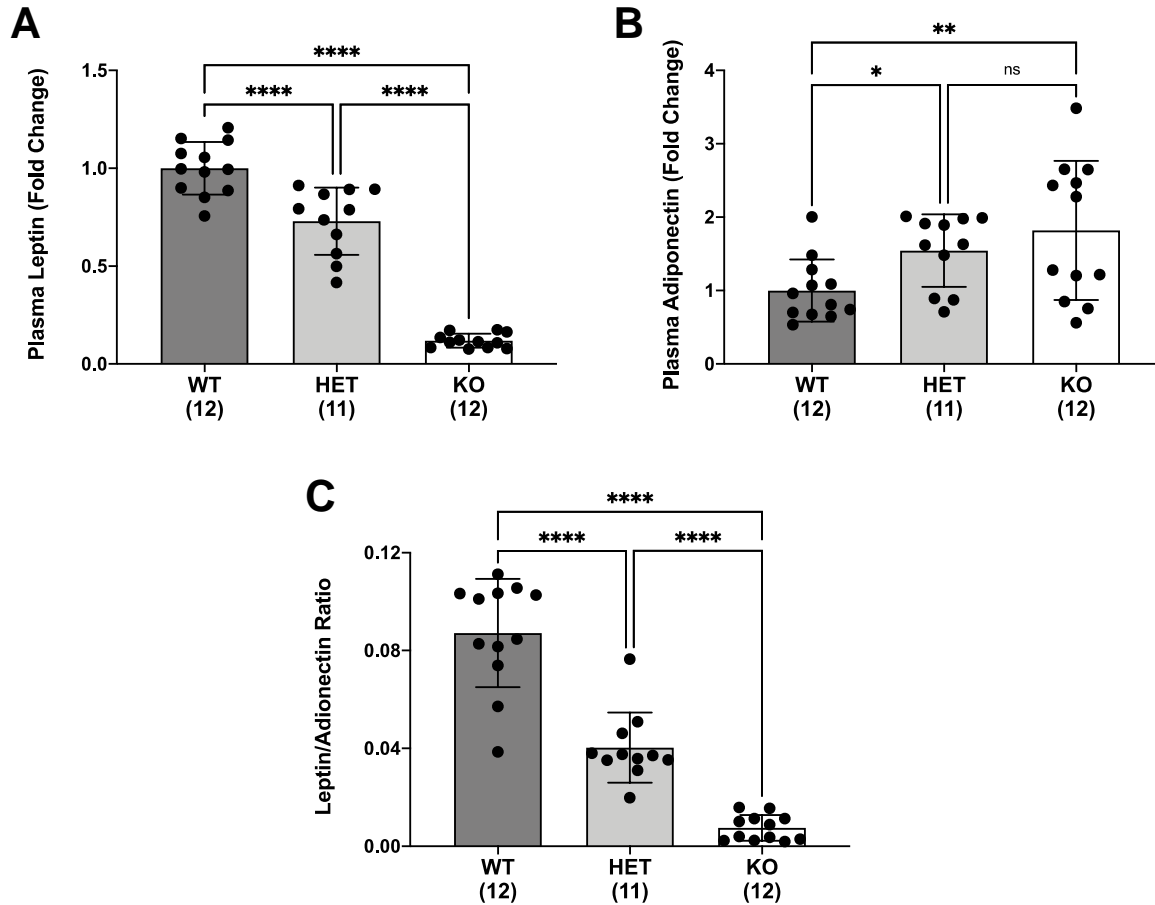


Fig. S6. Plasma leptin, adiponectin and leptin-adiponectin ratio in mouse experiments. *A* Plasma leptin levels in $Gpr75^{+/+}$ (WT), $Gpr75^{+/-}$ (HET), and $Gpr75^{-/-}$ (KO) mice after the high-fat diet challenge expressed as fold difference compared to wild-type (set as 1). Absolute levels (mean \pm standard deviation) for wild-type mice were 208 ± 42 pg/mL. *B* Plasma adiponectin levels in $Gpr75^{+/+}$ (WT), $Gpr75^{+/-}$ (HET), and $Gpr75^{-/-}$ (KO) mice after the high-fat diet challenge expressed as fold difference compared to wild-type (set as 1). Absolute levels (mean \pm standard deviation) for wild-type mice were $3,911 \pm 1,656$ ng/mL. *C* ratio of leptin to adiponectin in $Gpr75^{+/+}$ (WT), $Gpr75^{+/-}$ (HET), and $Gpr75^{-/-}$ (KO) expressed in ratio units. Number of mice included in each group and analysis are in parenthesis in the x-axis labels. Results are presented as mean \pm standard deviation. ns, not statistically-significant; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$ by two-way ANOVA with Tukey's multiple comparisons test.

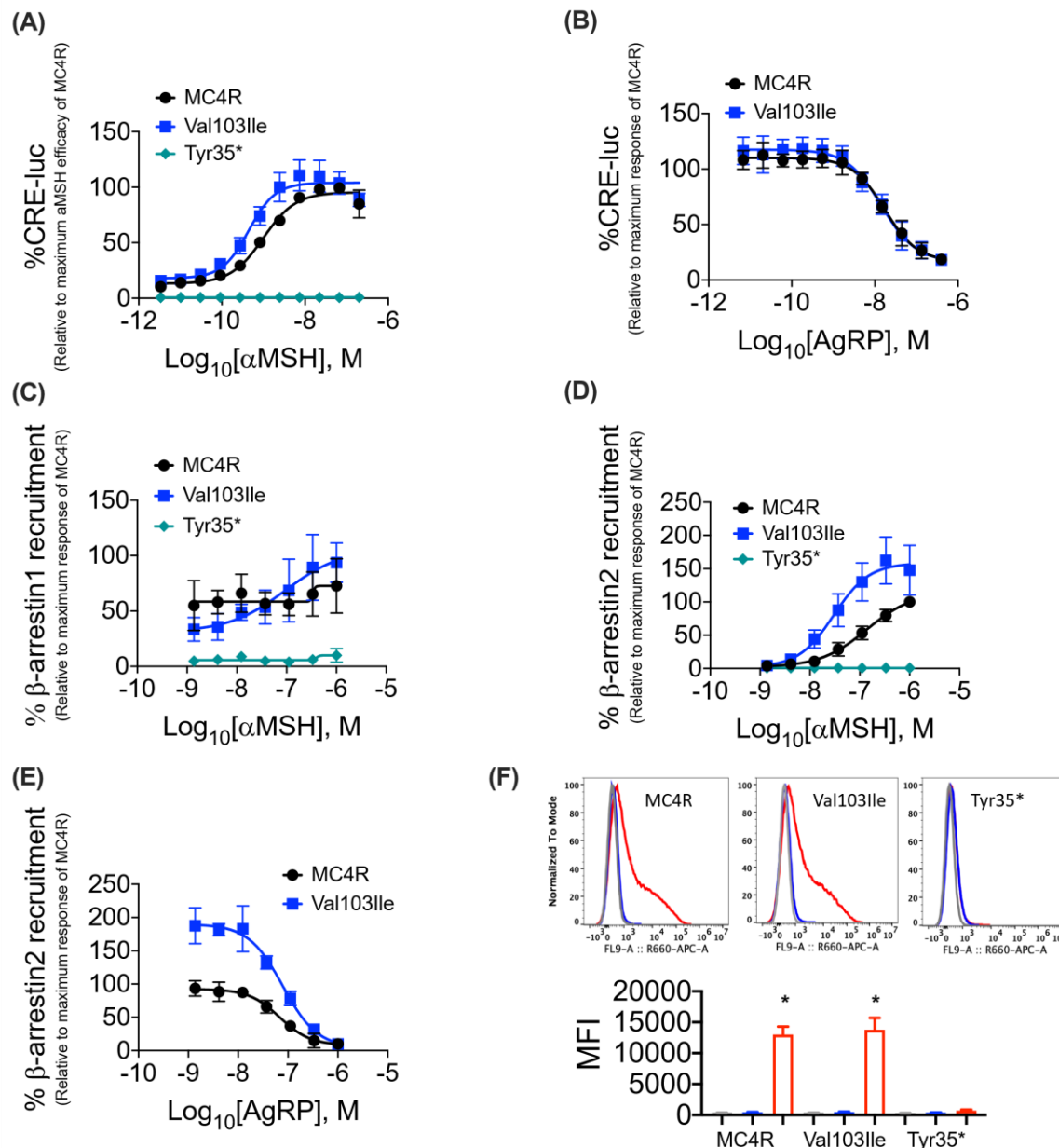


Fig. S7. *In vitro* experiments on the MC4R Val103Ile gain-of-function variant. *A* results of wild-type, Val103Ile, and Tyr35* MC4R Gs activation in response to stimulation with α -MSH assayed by CRE-luc reporter. *B* results of wild-type and Val103Ile mutant MC4R Gs in response to AgRP assayed by CRE-luc reporter. *C* and *D* β -arrestin1 and β -arrestin2 recruitment respectively in response to α -MSH stimulation of wild-type, Val103Ile, or Tyr35* MC4R assayed by NanoBiT protein interaction assay. *E* β -arrestin2 recruitment by wild-type or Val103Ile mutant MC4R in response to AgRP assayed by NanoBiT protein interaction assay. *F* flow-cytometry assay to quantify wild-type, Val103Ile, or Tyr35* MC4R localization to the plasma membrane. Addition of no antibody, only secondary antibody, and primary and secondary antibody are represented by the gray, blue, and red bars respectively. Abbreviations: α MSH, alpha-Melanocyte-stimulating hormone; AgRP, agouti-related peptide; MFI, mean fluorescence intensity.

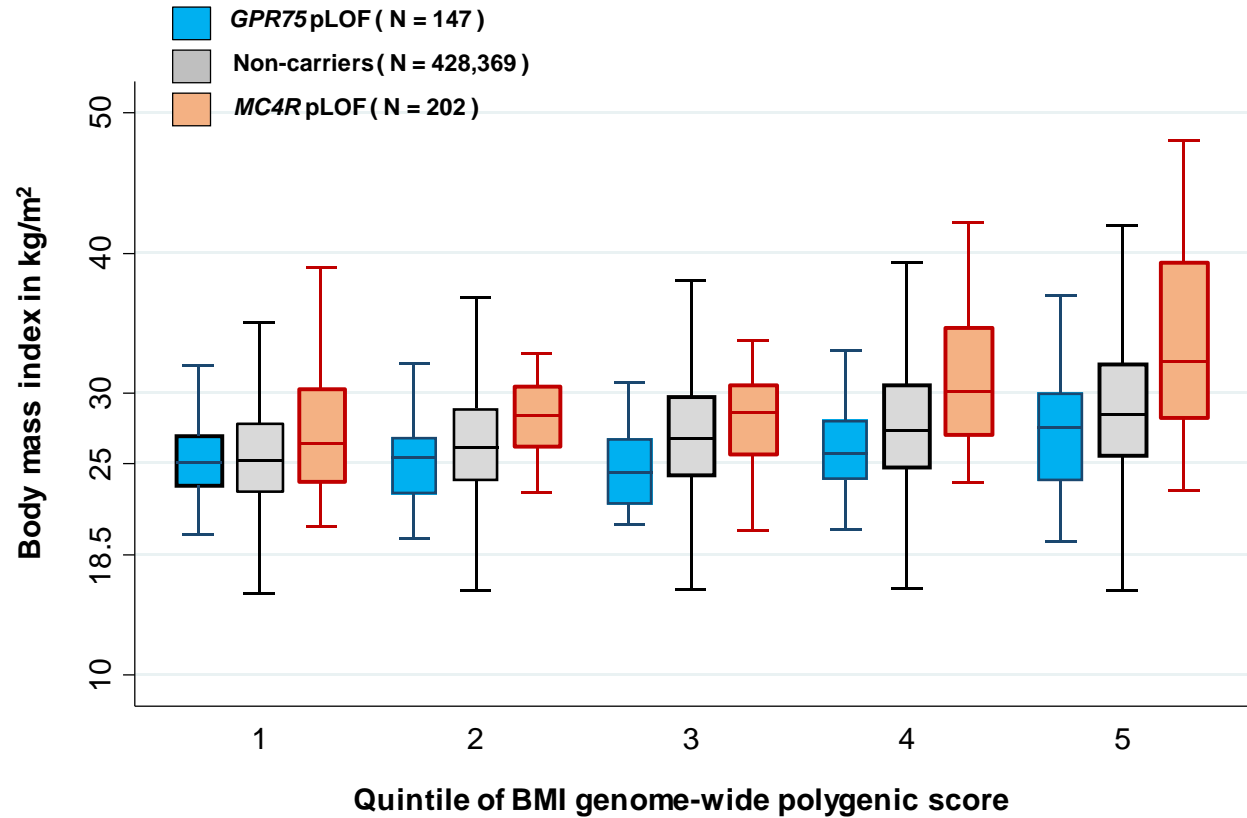


Fig. S8. Rare predicted loss-of-function genetic variants in *GPR75* and *MC4R*, polygenic predisposition and body mass index distribution. The Figure shows the distribution of body mass index in heterozygous carriers of predicted loss of function variants in *GPR75*, non-carriers, or heterozygous carriers of predicted loss of function variants in *MC4R* within quintiles of a genome-wide polygenic score for higher BMI. Boxes display the median, 25th and 75th percentiles, while whiskers display the upper and lower adjacent values for each group. Data are from European ancestry participants in the UK Biobank study who underwent exome sequencing. P-values for interaction between the polygenic score and rare pLOF variants on body mass index were 0.36 and 0.82 for *GPR75* and *MC4R*, respectively. Abbreviations: pLOF, potential loss of function; BMI, body mass index.

1. C. Sudlow *et al.*, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
2. D. J. Carey *et al.*, The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* **18**, 906-913 (2016).
3. R. Tapia-Conyer *et al.*, Cohort profile: the Mexico City Prospective Study. *Int J Epidemiol* **35**, 243-249 (2006).
4. J. Alegre-Diaz *et al.*, Diabetes and Cause-Specific Mortality in Mexico City. *N Engl J Med* **375**, 1961-1971 (2016).
5. W. E. Kraus *et al.*, A Guide for a Cardiovascular Genomics Biorepository: the CATHGEN Experience. *J Cardiovasc Transl Res* **8**, 449-457 (2015).
6. T. L. Assimes *et al.*, Genetics of Coronary Artery Disease in Taiwan: A Cardiometabochip Study by the Taichi Consortium. *PLoS One* **11**, e0138014 (2016).
7. R. G. Victor *et al.*, The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol* **93**, 1473-1480 (2004).
8. G. Berglund, S. Elmstahl, L. Janzon, S. A. Larsson, The Malmo Diet and Cancer Study. Design and feasibility. *J Intern Med* **233**, 45-51 (1993).
9. F. E. Dewey *et al.*, Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
10. WHO, Obesity and overweight. (2020).
11. I. T. Cristopher V. Van Hout, Joshua D. Backman, Joshua D. Hoffman, Daren Liu, Ashutosh K. Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, Alexander H. Li, Colm O'Dushlaine, Anthony Marcketta, Jeffrey Staples, Claudia Schurmann,, Alicia Hawes, Evan Maxwell, Leland Barnard, Alexander Lopez, John Penn,, Lukas Habegger, Andrew L. Blumenfeld, Xiaodong Bai, Sean O'Keeffe, Ashish Yadav, Kavita Praveen, Marcus Jones, William J. Salerno, Wendy K. Chung, Ida Surakka, Cristen J. Willer, Kristian Hveem, Joseph B. Leader, David J. Carey, David H. Ledbetter, Geisinger-Regeneron DiscovEHR Collaboration, Lon Cardon, George D. Yancopoulos, Aris Economides, Giovanni Coppola, Alan R. Shuldiner, Suganthi Balasubramanian, Michael Cantor, Regeneron Genetics Center, Matthew R. Nelson, John Whittaker, Jeffrey G. Reid, Jonathan Marchini, John D. Overton, Robert A. Scott, Gonalo R. Abecasis, Laura Yerges-Armstrong, Aris Baras. , Exome sequencing and characterization of 49,960 individuals in UK Biobank. *Nature*, (2020).
12. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009).
13. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).
14. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-1561 (2009).
15. J. M. Schwarz, C. Rodelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-576 (2010).
16. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
17. D. M. Valenzuela *et al.*, High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol* **21**, 652-659 (2003).
18. W. T. Poueymirou *et al.*, F0 generation mice fully derived from gene-targeted embryonic stem cells allowing immediate phenotypic analyses. *Nat Biotechnol* **25**, 91-99 (2007).
19. P. R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, A. L. Price, Mixed-model association for biobank-scale datasets. *Nat Genet* **50**, 906-908 (2018).
20. J. Mbatchou *et al.*, Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv*, 2020.2006.2019.162354 (2020).

21. A. Mahajan *et al.*, Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* **50**, 559-571 (2018).
22. J. Flannick *et al.*, Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71-76 (2019).
23. R. Do *et al.*, Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102-106 (2015).
24. C. Wellcome Trust Case Control, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
25. S. McCarthy *et al.*, A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016).
26. C. Benner *et al.*, FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
27. A. P. Morris, Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-822 (2011).
28. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
29. C. Wallace, Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet* **16**, e1008720 (2020).
30. B. J. Vilhjalmsen *et al.*, Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-592 (2015).
31. A. E. Locke *et al.*, Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
32. C. J. Willer, Y. Li, G. R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010).
33. M. Mele *et al.*, Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).
34. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).
35. S. L. Zeger, K. Y. Liang, P. S. Albert, Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060 (1988).