WHOLE GENOME SEQUENCING ANALYSIS OF BODY MASS INDEX IDENTIFIES NOVEL AFRICAN ANCESTRY-SPECIFIC RISK ALLELE

AUTHORS

Xinruo Zhang[1]*, Jennifer A. Brody[2]*, Mariaelisa Graff[1]*, Heather M. Highland[1]*, Nathalie Chami[3]*, Hanfei Xu[4], Zhe Wang[3], Kendra Ferrier[5], Geetha Chittoor[6], Navya S. Josyula[6], Xihao Li[7], Zilin Li[8], Matthew A. Allison[9], Diane M. Becker[10,†], Lawrence F. Bielak[11], Joshua C. Bis[2], Meher Preethi Boorgula[12], Donald W. Bowden[13], Jai G. Broome[14,15], Erin J. Buth[14], Christopher S. Carlson[16], Kyong-Mi Chang[17,18], Sameer Chavan[12], Yen-Feng Chiu[19], Lee-Ming Chuang[20], Matthew P. Conomos[14], Dawn L. DeMeo[21], Margaret Du[22], Ravindranath Duggirala[23], Celeste Eng[24], Alison E. Fohner[25], Barry I. Freedman[26], Melanie E. Garrett[27], Xiuqing Guo[28], Chris Haiman[29], Benjamin D. Heavner[14], Bertha Hidalgo[30], James E. Hixson[31], Yuk-Lam Ho[32], Brian D. Hobbs[21,33], Donglei Hu[24], Qin Hui[34,35], Chii-Min Hwu[36], Rebecca D. Jackson[37,†], Deepti Jain[14], Rita R. Kalyani[38], Sharon L.R. Kardia[11], Tanika N. Kelly[39], Ethan M. Lange[5], Michael LeNoir[40], Changwei Li[39], Loic Le. Marchand[41], Merry-Lynn N. McDonald[42], Caitlin P. McHugh[14], Alanna C. Morrison[31], Take Naseri[43], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium[44], Jeffrey O'Connell[45], Christopher J. O'Donnell[32,46], Nicholette D. Palmer[13], James S. Pankow[47], James A. Perry[48], Ulrike Peters[49], Michael H. Preuss[3], D.C. Rao[50], Elizabeth A. Regan[51], Sefuiva M. Reupena[52], Dan M. Roden[53], Jose Rodriguez-Santana[54], Colleen M. Sitlani[2], Jennifer A. Smith[11,55], Hemant K. Tiwari[56], Ramachandran S. Vasan[57], Zeyuan Wang[34], Daniel E. Weeks[58,59], Jennifer Wessel[60,61,62], Kerri L. Wiggins[2], Lynne R. Wilkens[41], Peter W.F. Wilson[35,63], Lisa R. Yanek[10], Zachary T. Yoneda[64], Wei Zhao[11,55], Sebastian Zöllner[65], Donna K. Arnett[66], Allison E. Ashley-Koch[27], Kathleen C. Barnes[12], John Blangero[67], Eric Boerwinkle[31], Esteban G. Burchard[68], April P. Carson[69], Daniel I. Chasman[70,71], Yii-Der Ida Chen[72], Joanne E. Curran[73], Myriam Fornage[74,31], Victor R. Gordeuk[75], Jiang He[39], Susan R. Heckbert[76,2], Lifang Hou[77], Marguerite R. Irvin[78], Charles Kooperberg[16], Ryan L. Minster[58], Braxton D. Mitchell[79], Mehdi Nouraie[80], Bruce M. Psaty[2,76,81], Laura M. Raffield[82], Alexander P. Reiner[76], Stephen S. Rich[83], Jerome I. Rotter[28], M. Benjamin Shoemaker[64], Nicholas L.

Smith[84,85,86], Kent D. Taylor[28], Marilyn J. Telen[87], Scott T. Weiss[88], Yingze Zhang[80], Nancy Heard-Costa[89], Yan V. Sun[34,35], Xihong Lin[7,90], L. Adrienne Cupples[4,¥,†], Leslie A. Lange[5,¥], Ching-Ti Liu[4,¥], Ruth J.F. Loos[3,91,¥], Kari E. North[1,¥], Anne E. Justice[6,¥]

* These authors contributed equally to this work.

¥ These authors equally supervised this work.

† Published posthumously

AUTHOR INFORMATION

Affiliations

[1]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Department of Biostatistics, School of Public Health, Boston University, Boston, MA, USA. [5]Division of Biomedical Informatics and Personalized Medicine, School of Medicine University of Colorado, Anschutz Medical Campus, Aurora, CO, USA. [6]Population Health Sciences, Geisinger, Danville, PA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA. [9]Department of Family Medicine, Division of Preventive Medicine, The University of California San Diego, La Jolla, CA, USA. [10]Department of Medicine, General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [12]Department of Medicine, School of Medicine,

University of Colorado, Aurora, CO, USA. [13]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [14]Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA. [15]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA. [16]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [17]The Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA. [18]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [19]Institute of Population Health Sciences, National Health Research Institutes, Taipei, Taiwan. [20]Department of Internal Medicine, Division of Metabolism/Endocrinology, National Taiwan University Hospital, Taipei, Taiwan. [21]Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [22]Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [23]Life Sciences, College of Arts and Sciences, Texas A&M University-San Antonio, San Antonio, TX, USA. [24]Department of Medicine, Lung Biology Center, University of California, San Francisco, San Francisco, CA, USA. [25]Epidemiology, Institute of Public Health Genetics, School of Public Health, University of Washington, Seattle, WA, USA. [26]Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC, USA. [27]Department of Medicine, Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA. [28]Department of Pediatrics, Genomic Outcomes, The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [29]Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [30]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham School of Public

Health, Birmingham, AL, USA. [31]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [32]Veterans Affairs Boston Healthcare System, Boston, MA, USA. [33]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [34]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA, USA. [35]Atlanta VA Health Care System, Decatur, GA, USA. [36]Department of Medicine, Division of Endocrinology and Metabolism, Taipei Veterans General Hospital, Taipei, Taiwan, Taiwan. [37]Endocrinology, Ohio State University, Columbus, OH, USA. [38]Department of Medicine, Endocrinology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [39]Department of Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA. [40]Department of Pediatrics, Bay Area Pediatrics, Oakland, CA, USA. [41]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA. [42]Department of Medicine, Pulmonary, Allergy and Critical Care, University of Alabama at Birmingham, Birmingham, AL, USA. [43]Ministry of Health, Government of Samoa, Apia, Samoa. [44]A full list of study authors are provided in the acknowledgements within the Supplementary Note. [45]Department of Medicine, Program for Personalized and Genomic Medicine, University of Maryland, Baltimore, MD, USA. [46]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [47]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [48]Department of Medicine, School of Medicine, University of Maryland, Baltimore, MD, USA. [49]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA. [50]Division of Biostatistics, Washington University in St. Louis, St. Louis, MO, USA. [51]Department of Medicine, Rheumatology,

National Jewish Health, Denver, CO, USA. [52]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [53]Medicine, Pharmacology, and Biomedical Informatics, Clinical Pharmacology and Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [54]Centro de Neumologia Pediatrica, San Juan, PR, USA. [55]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. [56]Department of Biostatistics, University of Alabama at Birmingham School of Public Health, Birmingham, AL, USA. [57]Department of Medicine, School of Medicine, Boston University, Boston, MA, USA. [58]Department of Human Genetics, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [59]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [60]Department of Epidemiology, Indiana University, Indianapolis, IN, USA. [61]Department of Medicine, Indiana University, Indianapolis, IN, USA. [62]Diabaetes Translational Research Center, Indiana University, Indianapolis, IN, USA. [63]Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA. [64]Department of Medicine, Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [65]Department of Biostatistics, Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA. [66]Department of Epidemiology, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA. [67]Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, TX, USA. [68]Bioengineering and Therapeutic Sciences and Medicine, Lung Biology Center, University of California, San Francisco, San Francisco, CA, USA. [69]Department of Medicine, University of Mississippi, Jackson, MI, USA. [70]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [71]Harvard Medical School, Boston, MA, USA. [72]Department of Medical Genetics, Genomic Outcomes, Lundquist Institute for

Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [73]Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, TX, USA. [74]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA. [75]Department of Medicine, School of Medicine, University of Illinois at Chicago, Chicago, IL, USA. [76]Department of Epidemiology, University of Washington, Seattle, WA, USA. [77]Northwestern University, Chicago, IL, USA. [78]Department of Epidemiology, University of Alabama at Birmingham School of Public Health, Birmingham, AL, USA. [79]Department of Medicine, Division of Endocrinology, Diabetes and Nutrition, University of Maryland, Baltimore, MD, USA. [80]Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [81]Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA. [82]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [83]Public Health Science, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [84]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA. [85]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, WA, USA. [86]Seattle Epidemiologic Research and Information Center, Office of Research and Development, Department of Veterans Affairs, Seattle, WA, USA. [87]Department of Medicine, Hematology, Duke University Medical Center, Durham, NC, USA. [88]Department of Medicine, Channing Division of Network Medicine, Harvard Medical School, Boston, MA, USA. [89]Framingham Heart Study, School of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA. [90]Department of Statistics, Harvard University, Boston,

MA, USA. [91]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark

Corresponding Authors

Anne E. Justice: aejustice1@geisinger.edu; Xinruo Zhang: xinruo@email.unc.edu

Author Contributions

Conducted analyses or contributed to figures and tables: XZ, JAB, MG, NC, ZhW, KF, GC, NSJ, QH, AEJ; Supervised analyses: YVS, LAC, LAL, CTL, RJFL, KEN, AEJ. Contributed to the design of the current study: XZ, JAB, MG, HMH, NC, HX, LAC, LAL, CTL, RJFL, KEN, AEJ. Contributed to the conception or design of the TOPMed program and its operations (including organization and policies of TOPMed – e.g., exec committee, working group conveners, NHLBI staff, etc.): NC, ZhW, DLD, BDHe, JEH, SLRK, TNK, JSP, EAR, RSV, JW, DKA, KCB, JH, SRH, BMP, LMR, SSR, JIR, NLS, KDT, LAC, CTL, RJFL, KEN, AEJ. Provided phenotypic data and/or biosamples: MG, NC, MAA, LFB, MPB, SC, DLD, RD, XG, CHa , BH, JEH, YH, RDJ, SLRK, EML, LLM, TN , NDP, MHP, EAR, SMR, DMD, RSV, DEW, JW, LRW, LRY, ZTY, DKA, JB, EGB, JEC, MF, JH, SRH, CK, RM, BMP, LMR, APR, SSR, JIR, MBS, NLS, NH, LAC, LAL, RJFL, KEN, AEJ. Acquired WGS and/or other omics data: LFB, DLD, RD, JEH, SLRK, JAS, RSV, JW, WZ, DKA, JB, EGB, JEC, MF, JH, SRH, CK, BMP, APR, SSR, JIR, NLS, KDT, LAC, RJFL, KEN, AEJ. Created software, processed, and/or analyzed WGS or other study data for data summaries in this paper: XZ, JAB, MG, HMH, NC, JCB, JGB, EJB, MPC, BDHe, CL, CPM, KLW, AEJ. Drafted the manuscript and revised according to co-author suggestions: XZ, JAB, MG, HMH, NC, KEN, AEJ. All authors critically reviewed the manuscript, suggested revisions as needed, and approved the final version.

Disclosures/Competing Interests

DLD received grants from Bayer and honoraria from Novartis. BDHo receives grant support from Bayer and has received an honorarium from AstraZeneca for an educational lecture. CJO is employed by Novartis Institute of Biomedical Research, Cambridge MA KCB is an employee of Tempus. BMPs serve on the TOPMed Steering Committee. LMR is a consultant for the TOPMed Administrative Coordinating Center (through Westat). XLin is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences.

ACKNOWLEDGEMENTS

This work was funded in part by National Institutes of Health (NIH) grants (R01 DK122503, T32 HL007055, T32 HL129982, R01 HL142825, I01-BX003362, U01 HL120393, R01 HL68959, U01 HL072507, K08 HL136928, R01 HL119443, R01 HL055673-18S1, R01 HL92301, R01 HL67348, R01 NS058700, R0 AR48797, R01 DK071891, R01 AG058921, F32 HL085989, U01 HL089897, U01 HL089856, R01 HL093093, R01 HL133040, I01 BX003340, I01 BX004821, U01 HL072524, R01 HL104135-04S1, U01 HL054472, U01 HL054473, U01 HL054495, U01 HL054509, R01 HL055673 with supplement -18S1, R01 HL104608, R01 AI132476, R01 AI114555, R01 HL104608-S1, U01 HL072507, P20 GM109036, KL2 TR002490 and T32 HL129982, P01 HL132825, R35 CA197449, P01 CA134294, U19 CA203654, R01 HL113338, U01 HG009088, R01 HL142302, R01 DK124097, R01 DK110113, R01 DK107786, X01 HL134588, R01 HG010297, U01 HG007416, R01 HL105756) and contracts (HHSN268201800001I, HHSN268201500014C,) American Diabetes Association (ADA) grant #1-19-PDF-045, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122), and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484). A full list of study acknowledgements is detailed in the **Supplementary Note**.

ABSTRACT

Obesity is a major public health crisis associated with high mortality rates. Previous genome-wide association studies (GWAS) investigating body mass index (BMI) have largely relied on imputed data from European individuals. This study leveraged whole-genome sequencing (WGS) data from 88,873 participants from the Trans-Omics for Precision Medicine (TOPMed) Program, of which 51% were of non-European population groups. We discovered 18 BMI-associated signals ($P < 5 \times 10^{-9}$). Notably, we identified and replicated a novel low frequency single nucleotide polymorphism (SNP) in *MTMR3* that was common in individuals of African descent. Using a diverse study population, we further identified two novel secondary signals in known BMI loci and pinpointed two likely causal variants in the *POC5* and *DMD* loci. Our work demonstrates the benefits of combining WGS and diverse cohorts in expanding current catalog of variants and genes confer risk for obesity, bringing us one step closer to personalized medicine.

INTRODUCTION

In 2015, approximately 12% of adults worldwide had obesity [1], and four years later, the global obesity-related deaths amounted to five million, translating to an age-standardized mortality rate of 62.6 per 100,000 individuals in 2019 [2]. Previous genome-wide association studies (GWAS) have identified hundreds of loci associated with obesity-related traits, primarily with body mass index (BMI) – a practical and widely used proxy of overall adiposity. However, most of these genome-wide screens relied on meta-analyses of imputed data, predominantly from individuals of European ancestry [3,4].

Despite making some advancements toward improving ancestral diversity in GWAS, ancestry-stratified analyses and multi-ancestry analyses leveraged for discovery and fine-mapping are uncommon and largely underpowered by comparison. Furthermore, follow-up investigations for known BMI loci identified in European ancestry populations are insufficiently conducted to evaluate the generalizability of these loci. As such, the majority of BMI risk variants are common variants (minor allele frequency [MAF] > 5%) in primarily European ancestry populations, most of which exhibit small effect sizes. While

these index variants collectively explain less than 5% of the total phenotypic variation in BMI [5], it is estimated that as much as 1/5 of the phenotypic variance can be captured by common variants across the entire genome [5], leaving low and rare variants (MAF ≤ 5%) with potentially large effects to be explored [6].

Whole-genome sequencing (WGS) outperforms genotyping arrays in capturing low and rare frequency variants, as demonstrated in a recent study where researchers revealed that the heritability of BMI estimated using WGS data was comparable to the pedigree-based estimates, $h^2 \approx 0.40$ [7]. Thus, the discrepancy between phenotypic variance explained by genetic variations in GWAS compared to the expected heritability may be due to the use of imputed genotypes rather than directly sequenced variations. Causal variants or SNPs in known loci may not be represented on 1000 Genomes panels, or not well imputed from reference data because of differences in linkage disequilibrium (LD) across populations. To address this limitation, we conducted WGS association analyses to identify rare, low-frequency, and ancestry-specific genetic variants associated with BMI, using data from the Trans-Omics for Precision Medicine (TOPMed) Program [8], which is the most racially and ethnically diverse WGS program to date, as well as the Centers for Common Disease Genomics (CCDG) Program [9].

METHODS

Study Population and Phenotype

Our study population was racially, ethnically, geographically, and ancestrally diverse. We analyzed a multi-population sample of 88,873 adults from 36 studies in the freeze 8 TOPMed and CCDG programs (**Figure 1**, **Supplementary Data 1)**. They belonged to 15 population groups, reflecting the way participants self-identified in each study. For individuals who had unreported or non-specific population memberships (e.g., “Multiple” or “Other”), we applied the Harmonized Ancestry and Race/Ethnicity (HARE) method [10] to infer their group memberships using genetic data. This imputation was applied to 8,015 participants (9% of the overall population), assigning each to one of the existing population groups. In this way, our study population groups were defined based on a combination of self-reported identity

and the first nine genetic principal components (PCs) **(Figure 1, Supplementary Fig 1,** and **Supplementary Data 1)**.

The 15 population groups were labeled by their self-identified or primary inferred population group (e.g., predominantly African ancestry/admixed African/Black were labeled as "African"). Sample sizes for these groups ranged from 341 to over 43,000 as follows: African (N = 22,488), Amish (N = 1,106), Asian (N = 1,241), Barbadian (N = 248), Central American (N = 776), Costa Rican (N = 341), Cuban (N = 2,128), Dominican (N = 2,046), European (N = 43,434), Han Chinese (N = 1,787), Mexican (N = 4,265), Puerto Rican (N = 4,991), Samoan (N = 1,274), South American (N = 695), and Taiwanese (N = 2,053). We refer to analyses involving all 15 population groups as multi-population analysis and group-specific analyses by their primary population group.

Among the 88,873 participants, 53,109 (60%) were female and 45,439 (51%) were non-European. The mean (SD) age of the participants was 53.5 (15.1) years. Additional descriptive tables of the participants are presented in **Supplementary Data 2 – 4**. BMI was calculated by dividing weight in kilograms by the square of height in meters. Participants were excluded from analyses if less than 18 years of age, had known pregnancy at the time of BMI measurement, had implausible BMI values (above 100 kg/m^2 without corroborating evidence), or did not provide appropriate consent. The mean (SD) of BMI varied by study, ranging from 23.4 (3.1) in GenSALT to 32.7 (6.8) in VAFAR (**Supplementary Data 2**), and by population group, ranging from 23.4 (3.1) in Han Chinese to 33.7 (6.8) in Samoans (**Supplementary Data 3**).

TOPMed WGS

A detailed description of WGS methods has been reported previously [11]. Details regarding the laboratory methods, data processing, and quality control are described on the TOPMed website (https://www.nhlbiwgs.org/methods). Briefly, ~30X WGS was conducted using Illumina HiSeq X Ten instruments at six sequencing centers. At the Center for Statistical Genetics at University of Michigan, TOPMed sequence data were mapped to the GRCh38 human genome reference sequence in a manner

consistent with the joint CCDG/TOPMed functionally equivalent read mapping pipeline [12]. Joint genotype calling on all samples in Freeze 8 used the GotCloud pipeline [13]. Variants were filtered using a Support Vector Machine (SVM) implemented in the libsvm software package. Sample-level quality assurance steps included concordance between annotated and genetic sex, between prior SNP array genotyping and WGS-derived genotypes, and between observed and expected relatedness and pedigree information from cleaned sequence data.

Common Variant Association Analysis

We performed multi-population WGS association analysis of BMI using GENESIS [14] on the Analysis Commons (http://analysiscommons.com) [15] computation platform. Analyses were performed using linear mixed models (LMM). To improve power and control for false positives with a non-normal phenotype distribution, we implemented a fully adjusted two-stage procedure for rank-normalization when fitting the null model [16]. The first model was fit by adjusting BMI for age, age squared, sex, study, population group, ancestry-representative PCs generated using PC-AiR[17], sequencing center, sequencing phase and project. A 4th degree sparse empirical kinship matrix (KM) computed with PC-Relate [18] was included to account for genetic relatedness among participants. We also allowed for heterogeneous residual variances across sex by population group (e.g., female European), as it has previously been shown to improve control of genomic inflation [19]. Residuals from the first model were rank normal transformed within population group and sex strata. The resulting transformed residuals were used to fit the second stage null model allowing for heterogeneous variances by the population group and sex strata and accounting for relatedness using the kinship matrix. Variants with a MAF of at least 0.5% were then tested individually.

Due to the large number of variants tested (N = 90,142,062) in the multi-population analysis, we adopted a significance threshold of 5×10^{-9} as has been used previously [20]. Group-specific analyses were conducted in the two largest population groups, European and African.

Replication Analyses

For the novel single-variant association identified in our discovery analyses, we requested replication from five independent cohorts of similar race, ethnicity, and continental ancestry to our discovery populations (N_{total} = 109,748): Multiethnic Cohort (MEC) [21], Million Veteran Program (MVP) [22,23], BioMe BioBank Program [24], UK Biobank (UKBB) [25,26], and the REasons for Geographic And Racial Differences in Stroke (REGARDS) study [27]. Phenotypes were developed and analyses were carried out under the same protocol as outlined above. We subsequently conducted inverse variance weighted fixed effects meta-analysis in METAL[28], using study-specific summary results. Additional details on the parent study design for each replication study are included in the **Supplementary Note**.

Conditional Analysis

To identify loci harboring multiple independent signals, we performed stepwise conditional analyses on the most significant signal within 500kb of our index variant. The significance threshold for secondary signals was determined by Bonferroni correction for the number of variants across all regions tested, $P = 5.96 \times 10^{-7}$ ($P < 0.05/83{,}928$ SNPs with MAF > 0.5% within 500kb of the 16 index SNPs). Variants passing the significance threshold after the first round were further conditioned on the top variant in the locus after the first round of conditioning, to identify potential third signals within each locus using the same threshold.

To determine whether association signals in known loci were independent of known signals, we performed conditional analyses using previously published index variants [5,29-48]. Specifically, we analyzed all genome-wide significant variants that were not the previously reported index variants but located within 500 kb of a known GWAS SNP. Given that these are potential new signals in regions known to influence BMI, index variants were considered independent if the estimated effect (β) value remained ≥ 90% of the unconditioned β value and $P < 6.25 \times 10^{-3}$ (0.05/8 loci tested). LDlink was used to calculate pairwise LD between potentially independent signals in known loci and produce LD heatmaps using the 1000 Genomes Global reference panel [49].

Rare Variant Aggregate Association Analysis

Rare variants with a MAF ≤ 1% were tested in aggregate by gene unit across studies in the multi-population analysis. Variants were grouped into gene units in reference to GENCODE v28, including both coding variants and variants falling within gene-associated non-coding elements. Coding variants included high-confidence loss of function variants (Ensembl Variant Effect Predictor [VEP] LoF = HC), missense variants (MetaSVM score > 0) and in-frame insertion/deletions or synonymous variants (FATHNMM-XF coding score > 0.5). In addition to coding variants, we included variants falling within the promoter of each transcript tested. Promoter regions were defined as falling in the 5 kb region 5' of the transcript and also overlaying a FANTOM5 Cage Peak [50]. In order to identify regulatory elements likely to be acting through the tested gene, we leveraged the GeneHancer database [51]. GeneHancer identifies enhancer regions and associates them with the specific genes they are likely to regulate, allowing us to aggregate regulatory regions by the likely target gene. GeneHancer regions were limited to the top 50% scored regions and variants falling in these regulatory elements were further filtered to those most likely to have a functional impact (FATHMM-MKL noncoding score > 0.75). Variants aggregated to gene units were tested using variant set mixed model association tests (SMMAT) [52]. Variants were weighted inversely to their MAF using a beta distribution density function with parameters 1 and 25. Genes were considered significantly associated after Bonferroni correction for the number of genes analyzed ($P < 5 \times 10^{-7}$).

Fine-Mapping

In order to identify candidate functional variants underlying association regions, we performed fine-mapping analyses in our multi-population GWAS single variant association summary statistics, using the program PAINTOR [53] which integrates the association strength and genomic functional annotation. We used the annotation file from aggregate-based testing described above under "Rare Variant Aggregate Association Analysis" to identify deleterious coding variants, variants within GeneHancer regions, and

variants within gene promoter regions. We restricted this analysis to variants located within 100 kb of the locus index variants. We calculated LD using our analysis subset of the TOPMed data, assuming one causal variant per locus, unless evidence of independent secondary signals was identified, in which case we allowed for two causal variants per locus.

PheWAS

To identify potential novel phenotypic associations with newly discovered variants, we performed a phenome-wide association (PheWAS) in the MyCode Community Health Initiative Study (MyCode), a hospital-based population study in central and northeastern Pennsylvania [54], and in the Charles Bronfman Institute for Personalized Medicine's BioMe BioBank Program (BioMe) located in New York City[55]; both studies had genetic data linked to electronic health records (EHR). ICD-10-CM and ICD-9-CM codes were mapped to unique PheCodes using the Phecode Map v1.2 [56] from the EHR. Cases were defined if individuals had two or more PheCodes on separate dates, while controls had zero instances of the relevant PheCode. We performed association analyses on PheCodes with $N \geq 20$ cases and 20 controls using logistic regressions, adjusting for current age, sex (for non-sex-specific PheCodes), and the first 15 PCs calculated from genome-wide data, and assuming an additive genetic model using the PheWAS package [57] in R. Analyses were conducted in the overall population as well as in individuals of genetically-informed African ancestry alone (as inferred from k-means clustering of the PCs [58]), given the potential population-specific association of our novel locus. We restricted our analyses to unrelated individuals up to 2nd degree. Association analyses were conducted within each study, followed by inverse variance weighted fixed effects meta-analysis in METAL[28]. PheCodes were deemed statistically significant after Bonferroni correction for the number of PheCodes analyzed ($P < 0.05/538 = 9.3 \times 10^{-5}$).

RESULTS

Single-variant Analyses

Among the 90 million SNPs included in the multi-population analysis, 86% (N = 77,178,487) were rare SNPs with a study-wide MAF of 0.5% < MAF ≤ 1%, and 6% (N = 5,542,150) were low-frequency (1% < MAF ≤ 5%) SNPs. In the multi-population unconditional analysis, we identified 16 loci that reached the prespecified genome-wide significance threshold of $P < 5 \times 10^{-9}$ (**Table 1**, **Figure 2, Supplementary Figs. 2 – 3**), including one low-frequency (MAF = 4%) and 15 common (MAF 14% – 50%) tag SNPs. In general, the low-frequency variant in our primary discovery showed a stronger effect than the common variants, with an estimated effect 2.14 times larger than the average common variants (0.078 *vs.* 0.037 on average). Of these 16 loci, 15 were in known BMI-associated regions, and one novel locus was identified on chromosome 22 harboring a low-frequency index SNP near *MTMR3* (rs111490516; MAF = 4%, β = 0.078, SE = 0.013, $P = 4.52 \times 10^{-9}$; **Table 1**). The MAF of this *MTMR3* locus varied widely across population groups, with the highest MAF observed in the African (13%) and Barbadian (13%) population groups, while it ranged from 0% to 5% in other population groups (**Supplementary Data 5**).

In the two population-specific analyses, 10 association signals reached genome-wide significance (**Supplementary Data 6, Supplementary Figs 4 – 7**). Two of these signals were also detected in the multi-population analysis. For two loci, *SEC16B* and *FTO*, each population-specific analysis revealed a distinct lead variant compared to the multi-population analysis; however, they were in high LD with (R^2 = 0.95 and R^2 = 1.00, according to TOP-LD [59]; **Supplementary Data 6**) and within 30 kb of the multi-population lead SNPs. Notably, the novel locus in *MTMR3* achieved significance exclusively in the African group. While the most significant SNP in the African population group (rs73396827) differed from that in the multi-population analysis (rs111490516), the two were in strong LD in the TOPMed African population (R^2 = 1.00). Both of these SNPs were fixed in the European group (MAF = 0%). In the European group analysis, one SNP in the *ALKAL2* locus on chromosome 2 (rs62107261, β = -0.102, SE =

0.016, $P = 2.08 \times 10^{-10}$, MAF = 5%) was not in LD with the corresponding lead variant in the multi-population analysis ($R^2 = 0.00$, as calculated in the analysis subset), but was a known independent secondary signal at this locus [40]. The remaining SNPs were in the proximity to the index SNPs in the corresponding loci from the multi-population analysis.

Replication

The replication sample sizes ranged from 4,413 in BioMe to 79,889 in MVP (**Supplementary Data 7**). In the five replication studies of Blacks, Africans, and African Americans, the MAF of rs111490516 in *MTMR3* ranged from 11% to 13%, aligning with the 13% observed in our African and Barbadian groups and contrasting to the 0% to 5% range in our non-African discovery groups (**Supplementary Data 7**). We replicated the novel variant rs111490516, demonstrating directionally consistent associations with BMI across the replication studies and a 68% reduction in the estimated effect when meta-analyzing across replication studies ($\beta = 0.025$, SE = 0.007, $P = 4.76 \times 10^{-4}$, MAF = 11%) compared to the discovery analysis (**Figure 3**, **Supplementary Data 7**). In the meta-analysis of 198,621 individuals from both discovery and replication studies, the estimated effect was 0.037 with a SE of 0.006 and a *P*-value of 4.19×10^{-9} (**Figure 3**, **Supplementary Data 7**).

To gain a better understanding of the potential functional consequence of the *MTMR3* locus, we used Ensembl VEP [60] to annotate all variants in high LD with our top SNP ($R^2 > 0.8$ in the African population group using TOP-LD [59]). Of the 54 variants in high LD, most were intronic or nearby *MTMR3* (**Supplementary Data 8**). Of these, four variants had a moderate CADD (combined annotation dependent depletion) score (scaled CADD > 10) with rs73394881 having the highest relative CADD score[61], three of which lay within a possible enhancer (rs73396896, $R^2 = 0.884$; rs73394881, $R^2 = 0.889$; rs74832232, $R^2 = 0.889$).

Conditional Analyses

Conditional analysis using the most associated variant at each locus revealed two significant secondary signals after multiple testing correction (**Tables 1, Supplementary Data 9, Supplementary Figure 8**). These included a known BMI-associated index variant on chromosome 2 (rs62107261, β = -0.097, SE = 0.014, $P = 2.06 \times 10^{-12}$, near *ALKAL2*) [40], which was also the most significant variant at this locus in the European group analysis (**Supplementary Data 6**). We further identified rs78769612 on chromosome 18 (β = -0.100, SE = 0.019, $P = 2.17 \times 10^{-7}$, near *MC4R*). Although both secondary SNPs were in known BMI-associated loci, rs78769612 near *MC4R* was a new index variant.

We additionally assessed independence for the top variants in known loci, by conditioning on all previously-reported index variants [5,29-48]. Two SNPs, rs2206277 in *TFAP2B* and rs3838785 in *BDNF*, remained significant after multiple test correction, indicating potentially novel signals in known loci (**Supplementary Data 10)**. The novel index variant from the internal conditional analysis, rs78769612 near *MC4R*, was not robust to this treatment, suggesting that this novel variant was not independent of known BMI variants. The LD matrix plots highlighted low LD (R^2 range 0.018 – 0.342) between our top SNP at the *BDNF* locus, rs3838785, and previously published lead variants within 500 kb (**Supplementary Fig. 9**). Although our top SNP, rs2206277, in the *TFAP2B* locus was conditionally independent of previously published BMI-risk SNPs ($\beta \geq 90\%$ of the unconditioned β and $P < 6.25 \times 10^{-3}$), this SNP was in high LD with two nearby published SNPs ($R^2 = 0.822$ for rs987237 and $R^2 = 0.793$ for rs72892910).

Aggregate-based testing

We did not identify any novel gene regions through association tests at the genome-wide level ($P < 5 \times 10^{-7}$) when aggregating variants with MAF ≤ 1%. Nevertheless, we successfully replicated previous gene-based associations with the well-known melanocortin 4 receptor (*MC4R*) gene ($P = 8.47 \times 10^{-8}$), with 111 alleles across 37 sites within coding regions, enhancers, and promoters for *MC4R*. The *MC4R* locus was also identified in single-variant analyses.

Fine-mapping

To pinpoint the most probable causal variant(s) underlying each of the 16 loci, we subsequently performed fine-mapping using PAINTOR [53]. Assuming one causal variant per locus, the index variants were the most likely causal variants in 14 loci, with posterior probabilities (PP) ranging from 0.02 and 1.00, and seven of them had a PP above 0.50 (**Supplementary Data 11, Supplementary Fig 10**). Two intronic index variants, rs2307111 in *POC5* and rs1379871 in *DMD*, were particularly noteworthy with PP exceeding 0.98. In contrast, variants with the highest PP in *ADCY3* and *ZC3H4* were not the reported index variants, although the highest PP for the *ADCY3* locus was below 0.50, and thus not likely the causal variant underlying this signal. In the *ZC3H4* locus, the highest PP variant (rs55731973, PP = 0.77) was intronic, located in the 5' UTR or upstream of *ZC3H4* depending on alternative transcripts, and resided in probable enhancer regions. Additionally, this variant was a significant cis-eQTL for *SAE1*[62], another nearby downstream gene.

PheWAS

To explore potential novel pleiotropy, we conducted association tests between the tag variant from our novel locus, rs111490516, and 538 PheCodes available in the MyCode and BioMe studies. No PheCode was significantly associated with rs111490516 following multiple test correction ($P < 9.3 \times 10^{-5}$). However, PheCode 327.3 (Sleep Apnea) and 327.32 (Obstructive Sleep Apnea) ranked among the top associated PheCodes ($P < 0.001$) (**Supplementary Data 12**, **Supplementary Fig 11**). Perhaps not coincidentally, obesity is one of the strongest risk factors for sleep apnea [63].

DISCUSSION

By leveraging WGS data from a large multi-population study, we identified and replicated one novel low-frequency BMI variant in *MTMR3*, specific to the diversity of our sample. We also identified two common secondary signals in known BMI loci, supported gene-based associations for *MC4R*, and

refined resolution in multiple loci by prioritizing candidate SNPs with high PP. Our discovery of the novel BMI-associated variant emphasizes the importance of studying diverse populations, which could further refine and expand the catalog of genes and variants that confer risk for obesity and potentially other disease traits.

The novel *MTMR3* variant, rs111490516, was most common in our African and Barbadian population groups (MAF = 13%) and of moderate frequency in our Dominican population group (MAF = 5%). We further replicated this association in study samples of similar population background. Yet, previous GWAS of BMI focusing on African ancestry individuals failed to identify a significant association in this region. It is not available for lookup in the most recent MVP BMI GWAS [23], although included in our replication. In one of the largest GWAS meta-analyses of imputed genotype data in African ancestry individuals with summary data available publicly, which was conducted by the African Ancestry Anthropometry Genetics Consortium (AAAGC, N up to 42,751)[37], this variant was directionally consistent and suggestively associated ($\beta = 0.042$, $P = 1.80 \times 10^{-4}$, MAF = 12%)[37]. Similarly, in our replication analysis of 109,748 individuals with imputed genotypes, *MTMR3* (rs111490516) was suggestively significant ($\beta = 0.025$, $P = 4.76 \times 10^{-4}$, MAF = 11%). Therefore, the lack of discovery in prior publications is likely not due to insufficient power. As indicated by our fine-mapping results for this novel locus, our index SNP is likely not causal but could be in LD with a causal SNP and also poorly captured in studies relying on imputation. In other words, the causal variant underlying this locus may be nearby, less frequent, and on an LD block more frequent in a population poorly represented in other imputation reference panels, but well represented in our WGS and highly diverse sample (e.g., Caribbean admixed individuals). In this case, one would require sequencing data in a large sample size with the relevant haplotype to detect a significant association that was not able to be identified with imputation in a similar number of people.

The SNP rs111490516 lies in an intron of the *MTMR3* (myotubularin related protein 3) gene, with limited evidence of involvement in regulatory or functional protein activity. Other variants mapped to *MTMR3* have been associated with obesity-related traits in GWAS. In a study of 155,961 healthy and

499 medication-free UKBB participants, rs5752989 near *MTMR3* was associated with fat-free mass (β =
500 0.115, $P = 8.00 \times 10^{-9}$, allele G frequency = 43%)[64]. In a meta-analysis of up to 628,000 BioBank Japan
501 (BBJ), UKBB, and FinnGen (FG) participants, the same SNP was associated with body weight (β = -
502 0.010, $P = 3.86 \times 10^{-8}$, allele A frequency ranged from 51% in FG to 86% in BBJ) [65].

503 The primary cellular function of *MTMR3* relates to regulation of autophagy [66]. Although there is
504 no direct evidence linking *MTMR3* to obesity, previous studies have established a connection between
505 *MTMR3* and related cardiometabolic traits. *MTMR3* was associated with LDL cholesterol ($P = 1 \times 10^{-8}$) in
506 a GWAS meta-analysis of European, East Asian, South Asian, and African ancestry individuals [67]. A
507 potential mechanism was proposed later suggesting *MTMR3* may mediate the association between
508 miRNA-4513 and total cholesterol [68]. Furthermore, pyruvate dehydrogenase complex-specific knockout
509 mice with high-fat diet induced obesity also exhibited increased blood glucose and higher expression
510 levels of *MTMR3* [69]. We utilized the Ensembl VEP database to explore predicted functional consequences
511 of our novel locus. While there is limited knowledge on the biological implications of the lead variant and
512 those in high LD, there are multiple lines of evidence supporting a role in obesity at this locus.

513 The use of WGS coupled with inclusion of non-European populations improved fine-mapping
514 resolution, as has been shown previously [47]. While there have been multiple attempts to fine-map
515 previously identified BMI loci [5,33,48], no previous study has successfully identified BMI risk variants of
516 high confidence at the *POC5* and *DMD* loci. By applying a Bayesian fine-mapping approach, we reduced
517 associated signals to 95% credible sets of two likely causal SNPs. Functional annotation revealed that one
518 of them, rs2307111, was a benign missense variant in *POC5* (NP_001092741.1:p.His36Arg) according to
519 ClinVar [70,71], while the other is an intron variant in the promotor region of *DMD*. These two variants were
520 also considered high-confidence causal variants ($PP_{rs2307111} = 0.96$, $PP_{rs1379871} = 0.99$) in a recent joint
521 analysis of three biobanks (UKBB, FG, BBJ) [72]. Notably, unlike in Kanai et al. where the PP appeared to
522 be driven by the Europeans (for rs2307111: $PP_{UKBB} = 0.96$, $PP_{BBJ} = 0.12$, $PP_{FG} = 0.01$; for rs1379871:

523 PP_{UKBB} = 1.00, PP_{FG} = 1.00), the effect alleles in our study were observed in high proportions across many
524 non-European population groups (**Supplementary Data 5**).

525 In addition to our novel findings, 17 of the 18 identified variants reside in previously reported
526 BMI-associated loci, highlighting the generalizability of the genes underlying BMI across populations,
527 including *SEC16B*, *TMEM18*, *ETV5*, *GNPDA2*, *BDFN*, and *MC4R* [5,34,48,73]. Three of the loci harbor genes
528 implicated in severe and early-onset obesity – *ADCY3*, *BDNF*, and *MC4R* [4]. We also consistently
529 identified multiple association signals of high effect in *MC4R*, which is a well-established monogenic
530 obesity gene, through our discovery analysis, internal conditional analysis, and rare variant aggregate
531 analysis.

532 While our study included a large sample size of diverse populations and leveraged high quality
533 WGS data from well-characterized and harmonized cohorts, our results should also be interpreted with
534 the following limitations. First, although our study is large compared to other harmonized and sequenced
535 data samples, the total study size is relatively modest compared to existing GWAS meta-analyses of
536 common variants using imputed genotype data. Moreover, rare variants, such as those analyzed in our
537 study, may require even larger sample sizes for novel discoveries. Even though our study is among the
538 most racially, ethnically, and ancestrally diverse yet conducted, the European population group still
539 represented 49% of our participants. On the other hand, diversity can contribute to added heterogeneity of
540 effect sizes, potentially limiting discovery in the multi-population analysis. We sought to overcome this
541 limitation by allowing for heterogeneous residual variances across population groups and examining
542 population stratified results when samples sizes were adequate. Notably, all our genome-wide significant
543 loci from population stratified analyses were also captured in the multi-population analysis, likely owing
544 to our considerations of heterogenous effects, self-identity (population groups), and ancestry (genotype
545 PCs). As has been shown by others [47], this underscores the importance of conducting multi-population
546 analysis using appropriate methods that account for heterogeneity and minimize the risk of inflation or
547 missed detection of loci that may vary in MAF or phenotypic effects across populations.

In summary, our study demonstrates the power of leveraging WGS data from diverse populations for new discoveries associated with BMI. As we enter the era of incorporating GWAS-based risk models in clinical practice, it is critical that we continue to diversify the data collected and analyzed in genomic research. Failure to do so risks further exacerbating health disparities for public health crisis such as obesity. Ultimately, our study brings us one step closer to understanding the complex genetic underpinnings of obesity, translating these leads into mechanistic insights, and developing targeted preventions and interventions to address this global public health challenge.

REFERENCES

1. Collaborators, G.B.D.O. *et al.* Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N Engl J Med* **377**, 13-27 (2017).
2. Chong, B. *et al.* Trends and predictions of malnutrition and obesity in 204 countries and territories: an analysis of the Global Burden of Disease Study 2019. *EClinicalMedicine* **57**, 101850 (2023).
3. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243-250 (2022).
4. Loos, R.J.F. & Yeo, G.S.H. The genetics of obesity: from discovery to biology. *Nat Rev Genet* **23**, 120-133 (2022).
5. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
6. Speakman, J.R., Loos, R.J.F., O'Rahilly, S., Hirschhorn, J.N. & Allison, D.B. GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. *Int J Obes (Lond)* **42**, 1524-1531 (2018).

7. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv*, 588020 (2019).

8. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*, 563866 (2019).

9. Abel, H.J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83-89 (2020).

10. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763-772 (2019).

11. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).

12. Regier, A.A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun* **9**, 4038 (2018).

13. Jun, G., Wing, M.K., Abecasis, G.R. & Kang, H.M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25**, 918-25 (2015).

14. Gogarten, S.M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346-5348 (2019).

15. Brody, J.A. *et al.* Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* **49**, 1560-1563 (2017).

16. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol* **43**, 263-275 (2019).

17. Conomos, M.P., Miller, M.B. & Thornton, T.A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276-93 (2015).

18. Conomos, M.P., Reiner, A.P., Weir, B.S. & Thornton, T.A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-48 (2016).

19. Conomos, M.P. *et al.* Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am J Hum Genet* **98**, 165-84 (2016).

20. Lin, D.Y. A simple and accurate method to determine genomewide significance for association tests in sequencing studies. *Genet Epidemiol* **43**, 365-372 (2019).

21. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* **151**, 346-57 (2000).

22. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).

23. Huang, J. *et al.* Genomics and phenomics of body mass index reveals a complex disease network. *Nat Commun* **13**, 7973 (2022).

24. Abul-Husn, N.S. *et al.* Implementing genomic screening in diverse populations. *Genome Med* **13**, 17 (2021).

25. Sudlow, C. *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).

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

27. Howard, V.J. *et al.* The reasons for geographic and racial differences in stroke study: objectives and design. *Neuroepidemiology* **25**, 135-43 (2005).

28. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

29. Wen, W. *et al.* Meta-analysis identifies common variants associated with body mass index in east Asians. *Nat Genet* **44**, 307-11 (2012).

30. Okada, Y. *et al.* Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nat Genet* **44**, 302-6 (2012).

31. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**, 501-12 (2013).

32. Monda, K.L. *et al.* A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat Genet* **45**, 690-6 (2013).

33. Gong, J. *et al.* Fine Mapping and Identification of BMI Loci in African Americans. *Am J Hum Genet* **93**, 661-71 (2013).

34. Wen, W. *et al.* Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum Mol Genet* **23**, 5492-504 (2014).

35. Winkler, T.W. *et al.* The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet* **11**, e1005378 (2015).

36. Fernandez-Rhodes, L. *et al.* Trans-ethnic fine-mapping of genetic loci for body mass index in the diverse ancestral populations of the Population Architecture using Genomics and Epidemiology (PAGE) Study reveals evidence for multiple signals at established loci. *Hum Genet* **136**, 771-800 (2017).

37. Ng, M.C.Y. *et al.* Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. *PLoS Genet* **13**, e1006719 (2017).

38. Justice, A.E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat Commun* **8**, 14977 (2017).

39. Graff, M. *et al.* Genome-wide physical activity interactions in adiposity - A meta-analysis of 200,452 adults. *PLoS Genet* **13**, e1006528 (2017).

40. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am J Hum Genet* **100**, 865-884 (2017).

41. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet* **49**, 1458-1467 (2017).

649 42. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways
650 that control energy intake and expenditure in obesity. *Nat Genet* **50**, 26-41 (2018).
651 43. Gong, J. *et al.* Trans-ethnic analysis of metabochip data identifies two new loci associated with
652 BMI. *Int J Obes (Lond)* **42**, 384-390 (2018).
653 44. Hoffmann, T.J. *et al.* A Large Multiethnic Genome-Wide Association Study of Adult Body Mass
654 Index Identifies Novel Loci. *Genetics* **210**, 499-515 (2018).
655 45. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass
656 index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649
657 (2018).
658 46. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J*
659 *Hum Genet* **104**, 65-75 (2019).
660 47. Wojcik, G.L. *et al.* Genetic analyses of diverse populations improves discovery for complex
661 traits. *Nature* **570**, 514-518 (2019).
662 48. Fernandez-Rhodes, L. *et al.* Ancestral diversity improves discovery and fine-mapping of genetic
663 loci for anthropometric traits-The Hispanic/Latino Anthropometry Consortium. *HGG Adv* **3**,
664 100099 (2022).
665 49. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-
666 specific haplotype structure and linking correlated alleles of possible functional variants.
667 *Bioinformatics* **31**, 3555-7 (2015).
668 50. Consortium, F. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
669 51. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in
670 GeneCards. *Database (Oxford)* **2017**(2017).
671 52. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary
672 Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum Genet* **104**, 260-274
673 (2019).

53. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).

54. Carey, D.J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* **18**, 906-13 (2016).

55. Abul-Husn, N.S. & Kenny, E.E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **177**, 58-69 (2019).

56. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).

57. Carroll, R.J., Bastarache, L. & Denny, J.C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-6 (2014).

58. Dewey, F.E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**(2016).

59. Huang, L. *et al.* TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. *Am J Hum Genet* **109**, 1175-1181 (2022).

60. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

61. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

62. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).

63. Jehan, S. *et al.* Obstructive Sleep Apnea and Obesity: Implications for Public Health. *Sleep Med Disord* **1**(2017).

64. Hubel, C. *et al.* Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am J Med Genet B Neuropsychiatr Genet* **180**, 428-438 (2019).

65. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415-1424 (2021).

66. Vergne, I. & Deretic, V. The role of PI3P phosphatases in the regulation of autophagy. *FEBS Lett* **584**, 1313-8 (2010).

67. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283 (2013).

68. Ghanbari, M. *et al.* A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Hum Mutat* **35**, 1524-31 (2014).

69. Geng, Z. *et al.* RNA-Seq analysis of obese Pdha1(fl/fl)Lyz2-Cre mice induced by a high-fat diet. *Exp Anim* (2022).

70. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).

71. ClinVar, N.C.f.B.I. VCV001233562.6.

72. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv*, 2021.09.03.21262975 (2021).

73. Hotta, K. *et al.* Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *J Hum Genet* **54**, 727-31 (2009).

TABLES

Table 1. Summary of independent loci reaching genome-wide significance ($P < 5 \times 10^{-9}$) in single variant and internal conditional analyses

CHR	POS (hg38)	Nearest gene	rsID	REF	ALT	ALT Freq	Beta	SE	*P*-value	Known index variant[a]	Novel Locus[b]
Top variant in each locus											
1	177920345	*SEC16B*	rs543874	A	G	20%	0.064	0.006	1.38E-26	Yes	No
2	621558	*TMEM18*	rs939584	C	T	85%	0.058	0.007	1.99E-17	No	No
2	24927427	*ADCY3*	rs10182181	A	G	56%	0.035	0.005	1.76E-11	Yes	No
3	186108951	*ETV5*	rs869400	T	G	82%	0.038	0.006	1.21E-09	No	No
4	45179317	*GNPDA2*	rs12507026	A	T	36%	0.045	0.005	9.55E-19	Yes	No
5	75707853	*POC5*	rs2307111	T	C	55%	-0.032	0.005	7.43E-10	Yes	No
6	50830813	*TFAP2B*	rs2206277	C	T	19%	0.054	0.006	2.05E-18	**Novel**	No
8	76068626	*HNF4G*	rs830463	A	G	47%	0.031	0.005	6.58E-10	No	No
11	27657463	*BDNF*	rs3838785	GT	G	58%	-0.030	0.005	3.14E-09	**Novel**	No
12	49853685	*BCDIN3D*	rs7138803	G	A	30%	0.036	0.005	1.69E-11	Yes	No
13	53533448	*OLFM4*	rs9568868	G	T	14%	0.047	0.007	5.73E-11	No	No
16	53767042	*FTO*	rs1421085	T	C	29%	0.090	0.006	6.11E-59	Yes	No
18	60161902	*MC4R*	rs6567160	T	C	21%	0.053	0.006	8.22E-19	Yes	No
19	47077985	*ZC3H4*	rs28590228	C	T	50%	0.033	0.005	4.75E-10	No	No
22	**29906934**	***MTMR3***	**rs111490516**	**C**	**T**	**4%**	**0.078**	**0.013**	**4.52E-09**	**Novel**	**Yes**
X	31836665	*DMD*	rs1379871	G	C	41%	0.029	0.004	1.35E-11	Yes	No
Secondary signals											
2	422144	*ALKAL2*	rs62107261	T	C	3%	-0.095	0.014	3.83E-12	Yes	No
18	60361739	*MC4R*	rs78769612	G	T	2%	-0.106	0.019	3.53E-08	No	No

Newly identified locus highlighted in bold. *CHR, chromosome; POS, position; REF, reference allele; ALT, alternative allele; ALT Freq, alternative allele frequency; SE, standard error.*

[a] Known index variant 'Yes' indicates previously published index variant from NHGRI-EBI GWAS Catalog; 'No' indicates index variant within 500 kb ± of the published lead variant, not independent of known signal in conditional analysis; 'Novel' indicates new lead variant either not published or conditionally independent.

[b] Novel locus 'Yes' was defined if there is no known index variant within 500 kb ± of the lead variant in current analysis.

FIGURES

Figure 1. Study population group composition.

A) Pairwise scatter plots of the first three principal components (PCs) by population group. B) This image contains a lollipop chart and a waffle plot illustrating the number and proportion of participants by population group. Our study population was composed of 88,873 participants from 15 population groups, 51% of which are non-European.

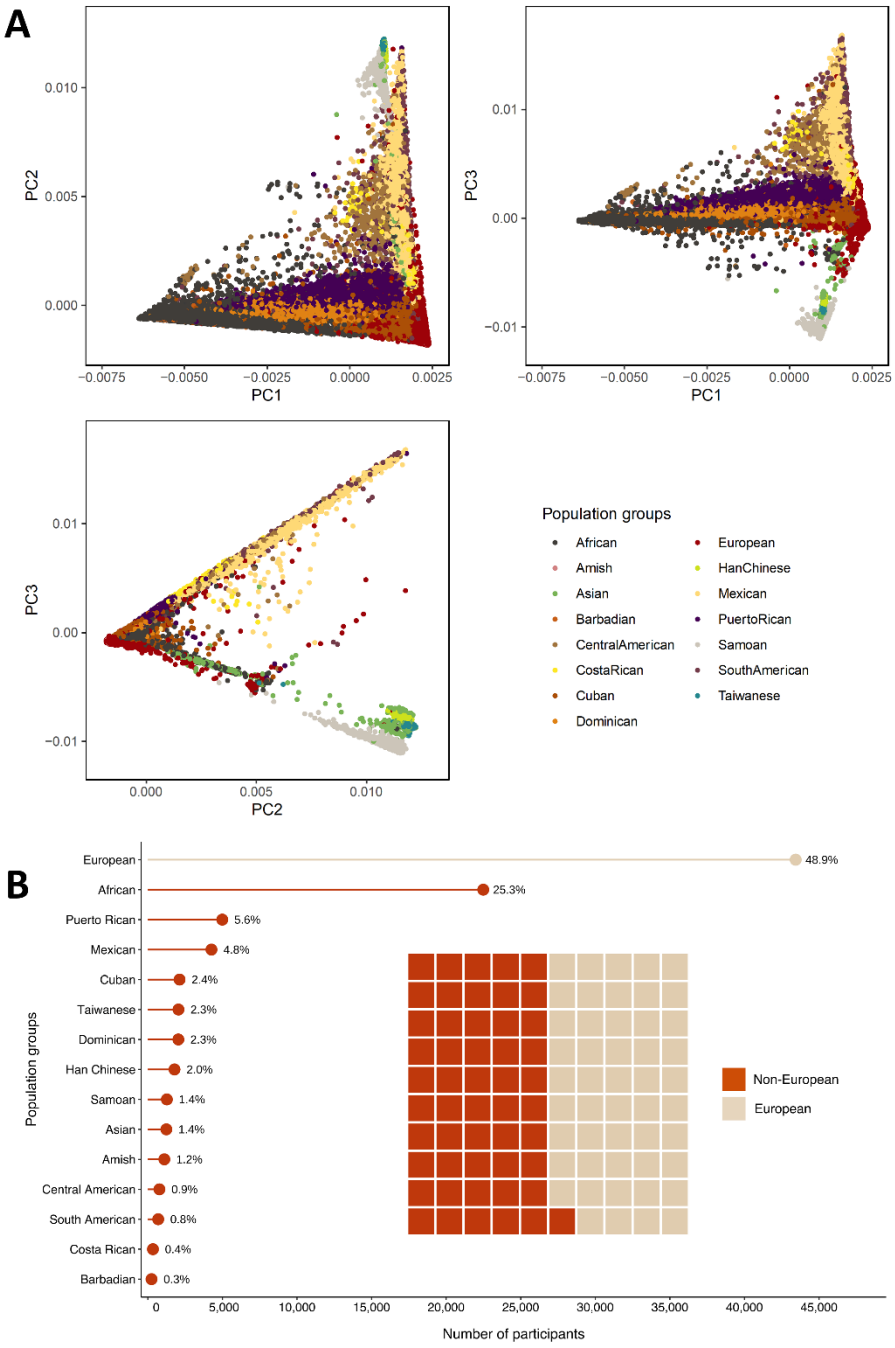

Figure 2. Summary of significant association findings.

A) Manhattan plot of multi-population, single variant analysis (N = 88,873 individuals). The novel locus *(MTMR3)* is highlighted in red. Previously reported BMI loci are in dark beige. The horizontal dashed line indicates genome-wide significance threshold $P = 5 \times 10^{-9}$. B) Scatterplot showing the minor allele frequency compared to the absolute value of the estimate effect of the index variant at each significant locus. All effect estimates are from the primary analysis conducted across all population groups. Previously reported loci are highlighted in blue, while the novel locus is in red; circles represent the most significant variant at each locus, and triangles show newly reported secondary signals within known loci.

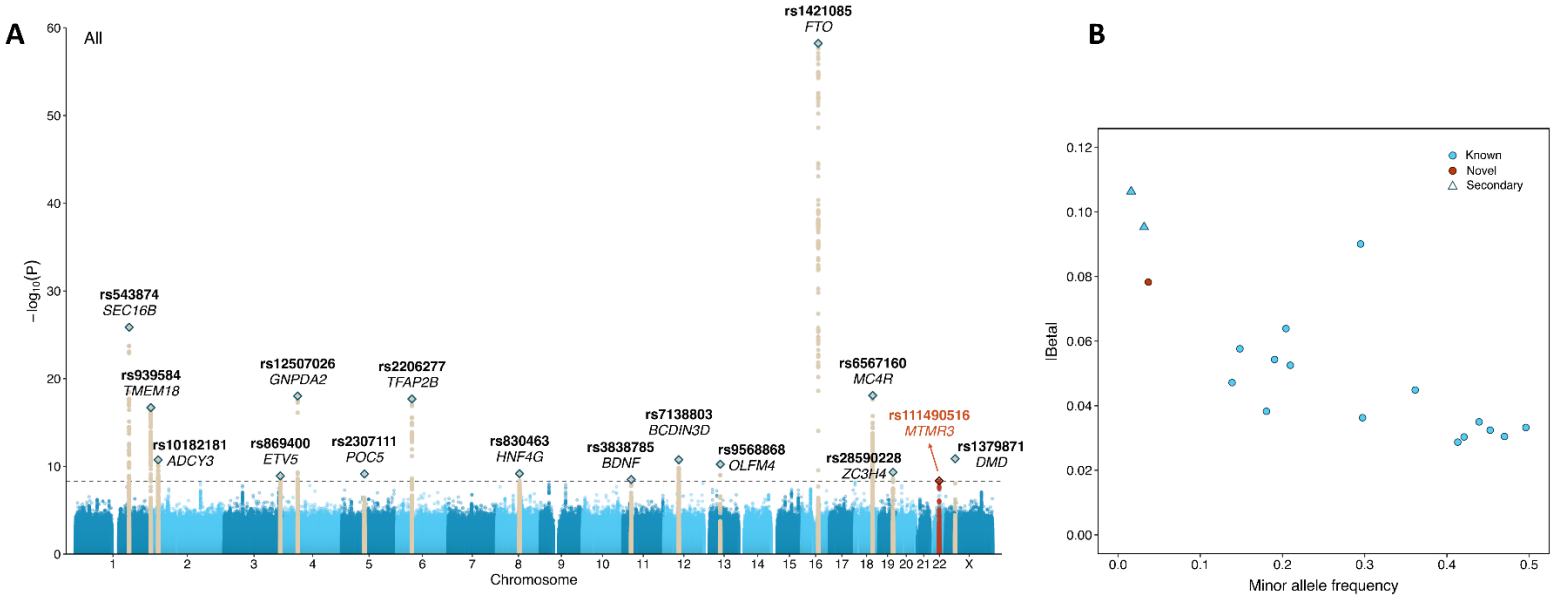

Figure 3. Forest plot of rs111490516 replication. All effect estimates (95% confidence interval) are oriented on the BMI increasing allele and are provided as standard deviation per allele. Actual beta values and *P*-values are in Supplementary Data 7.

.