

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

3
4 **AUTHORS**

5 Xinruo Zhang^{1*}, Jennifer A. Brody^{2*}, Mariaelisa Graff^{1*}, Heather M. Highland^{1*}, Nathalie Cham^{3*},
6 Hanfei Xu⁴, Zhe Wang³, Kendra Ferrier⁵, Geetha Chittoor⁶, Navya S. Josyula⁶, Xihao Li⁷, Zilin Li⁸,
7 Matthew A. Allison⁹, Diane M. Becker^{10,†}, Lawrence F. Bielak¹¹, Joshua C. Bis², Meher Preethi
8 Boorgula¹², Donald W. Bowden¹³, Jai G. Broome^{14,15}, Erin J. Buth¹⁴, Christopher S. Carlson¹⁶, Kyong-Mi
9 Chang^{17,18}, Sameer Chavan¹², Yen-Feng Chiu¹⁹, Lee-Ming Chuang²⁰, Matthew P. Conomos¹⁴, Dawn L.
10 DeMeo²¹, Margaret Du²², Ravindranath Duggirala²³, Celeste Eng²⁴, Alison E. Fohner²⁵, Barry I.
11 Freedman²⁶, Melanie E. Garrett²⁷, Xiuqing Guo²⁸, Chris Haiman²⁹, Benjamin D. Heavner¹⁴, Bertha
12 Hidalgo³⁰, James E. Hixson³¹, Yuk-Lam Ho³², Brian D. Hobbs^{21,33}, Donglei Hu²⁴, Qin Hui^{34,35}, Chii-Min
13 Hwu³⁶, Rebecca D. Jackson^{37,†}, Deepti Jain¹⁴, Rita R. Kalyani³⁸, Sharon L.R. Kardia¹¹, Tanika N. Kelly³⁹,
14 Ethan M. Lange⁵, Michael LeNoir⁴⁰, Changwei Li³⁹, Loic Le. Marchand⁴¹, Merry-Lynn N. McDonald⁴²,
15 Caitlin P. McHugh¹⁴, Alanna C. Morrison³¹, Take Naseri⁴³, NHLBI Trans-Omics for Precision Medicine
16 (TOPMed) Consortium⁴⁴, Jeffrey O'Connell⁴⁵, Christopher J. O'Donnell^{32,46}, Nicholette D. Palmer¹³,
17 James S. Pankow⁴⁷, James A. Perry⁴⁸, Ulrike Peters⁴⁹, Michael H. Preuss³, D.C. Rao⁵⁰, Elizabeth A.
18 Regan⁵¹, Sefuiva M. Reupena⁵², Dan M. Roden⁵³, Jose Rodriguez-Santana⁵⁴, Colleen M. Sitlani², Jennifer
19 A. Smith^{11,55}, Hemant K. Tiwari⁵⁶, Ramachandran S. Vasani⁵⁷, Zeyuan Wang³⁴, Daniel E. Weeks^{58,59},
20 Jennifer Wessel^{60,61,62}, Kerri L. Wiggins², Lynne R. Wilkens⁴¹, Peter W.F. Wilson^{35,63}, Lisa R. Yanek¹⁰,
21 Zachary T. Yoneda⁶⁴, Wei Zhao^{11,55}, Sebastian Zöllner⁶⁵, Donna K. Arnett⁶⁶, Allison E. Ashley-Koch²⁷,
22 Kathleen C. Barnes¹², John Blangero⁶⁷, Eric Boerwinkle³¹, Esteban G. Burchard⁶⁸, April P. Carson⁶⁹,
23 Daniel I. Chasman^{70,71}, Yii-Der Ida Chen⁷², Joanne E. Curran⁷³, Myriam Fornage^{74,31}, Victor R.
24 Gordeuk⁷⁵, Jiang He³⁹, Susan R. Heckbert^{76,2}, Lifang Hou⁷⁷, Marguerite R. Irvin⁷⁸, Charles Kooperberg¹⁶,
25 Ryan L. Minster⁵⁸, Braxton D. Mitchell⁷⁹, Mehdi Nouraei⁸⁰, Bruce M. Psaty^{2,76,81}, Laura M. Raffield⁸²,
26 Alexander P. Reiner⁷⁶, Stephen S. Rich⁸³, Jerome I. Rotter²⁸, M. Benjamin Shoemaker⁶⁴, Nicholas L.

27 Smith^{84,85,86}, Kent D. Taylor²⁸, Marilyn J. Telen⁸⁷, Scott T. Weiss⁸⁸, Yingze Zhang⁸⁰, Nancy Heard-
28 Costa⁸⁹, Yan V. Sun^{34,35}, Xihong Lin^{7,90}, L. Adrienne Cupples^{4,‡,†}, Leslie A. Lange^{5,‡}, Ching-Ti Liu^{4,‡},
29 Ruth J.F. Loos^{3,91,‡}, Kari E. North^{1,‡}, Anne E. Justice^{6,‡}

30
31 * These authors contributed equally to this work.

32 ‡ These authors equally supervised this work.

33 † Published posthumously

34

35 **AUTHOR INFORMATION**

36

37 **Affiliations**

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

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

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

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

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

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

145

146 **Corresponding Authors**

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

148

149 **Author Contributions**

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

166

167 **Disclosures/Competing Interests**

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

174

175 **ACKNOWLEDGEMENTS**

176 This work was funded in part by National Institutes of Health (NIH) grants (R01 DK122503, T32
177 HL007055, T32 HL129982, R01 HL142825, I01-BX003362, U01 HL120393, R01 HL68959, U01
178 HL072507, K08 HL136928, R01 HL119443, R01 HL055673-18S1, R01 HL92301, R01 HL67348, R01
179 NS058700, R0 AR48797, R01 DK071891, R01 AG058921, F32 HL085989, U01 HL089897, U01
180 HL089856, R01 HL093093, R01 HL133040, I01 BX003340, I01 BX004821, U01 HL072524, R01
181 HL104135-04S1, U01 HL054472, U01 HL054473, U01 HL054495, U01 HL054509, R01 HL055673
182 with supplement -18S1, R01 HL104608, R01 AI132476, R01 AI114555, R01 HL104608-S1, U01
183 HL072507, P20 GM109036, KL2 TR002490 and T32 HL129982, P01 HL132825, R35 CA197449, P01
184 CA134294, U19 CA203654, R01 HL113338, U01 HG009088, R01 HL142302, R01 DK124097, R01
185 DK110113, R01 DK107786, X01 HL134588, R01 HG010297, U01 HG007416, R01 HL105756) and
186 contracts (HHSN268201800001I, HHSN268201500014C,) American Diabetes Association (ADA) grant
187 #1-19-PDF-045, the General Clinical Research Center of the Wake Forest University School of Medicine
188 (M01 RR07122), and a pilot grant from the Claude Pepper Older Americans Independence Center of
189 Wake Forest University Health Sciences (P60 AG10484). A full list of study acknowledgements is
190 detailed in the **Supplementary Note**.

191

192

193 **ABSTRACT**

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

205

206 **INTRODUCTION**

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

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

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

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

233

234 **METHODS**

235

236 **Study Population and Phenotype**

237 Our study population was racially, ethnically, geographically, and ancestrally diverse. We
238 analyzed a multi-population sample of 88,873 adults from 36 studies in the freeze 8 TOPMed and CCDG
239 programs (**Figure 1, Supplementary Data 1**). They belonged to 15 population groups, reflecting the way
240 participants self-identified in each study. For individuals who had unreported or non-specific population
241 memberships (e.g., “Multiple” or “Other”), we applied the Harmonized Ancestry and Race/Ethnicity
242 (HARE) method ¹⁰ to infer their group memberships using genetic data. This imputation was applied to
243 8,015 participants (9% of the overall population), assigning each to one of the existing population groups.
244 In this way, our study population groups were defined based on a combination of self-reported identity

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

247 The 15 population groups were labeled by their self-identified or primary inferred population
248 group (e.g., predominantly African ancestry/admixed African/Black were labeled as “African”). Sample
249 sizes for these groups ranged from 341 to over 43,000 as follows: African (N = 22,488), Amish (N =
250 1,106), Asian (N = 1,241), Barbadian (N = 248), Central American (N = 776), Costa Rican (N = 341),
251 Cuban (N = 2,128), Dominican (N = 2,046), European (N = 43,434), Han Chinese (N = 1,787), Mexican
252 (N = 4,265), Puerto Rican (N = 4,991), Samoan (N = 1,274), South American (N = 695), and Taiwanese
253 (N = 2,053). We refer to analyses involving all 15 population groups as multi-population analysis and
254 group-specific analyses by their primary population group.

255 Among the 88,873 participants, 53,109 (60%) were female and 45,439 (51%) were non-
256 European. The mean (SD) age of the participants was 53.5 (15.1) years. Additional descriptive tables of
257 the participants are presented in **Supplementary Data 2 – 4**. BMI was calculated by dividing weight in
258 kilograms by the square of height in meters. Participants were excluded from analyses if less than 18
259 years of age, had known pregnancy at the time of BMI measurement, had implausible BMI values (above
260 100 kg/m² without corroborating evidence), or did not provide appropriate consent. The mean (SD) of
261 BMI varied by study, ranging from 23.4 (3.1) in GenSALT to 32.7 (6.8) in VAFAR (**Supplementary**
262 **Data 2**), and by population group, ranging from 23.4 (3.1) in Han Chinese to 33.7 (6.8) in Samoans
263 (**Supplementary Data 3**).

264

265 **TOPMed WGS**

266 A detailed description of WGS methods has been reported previously¹¹. Details regarding the
267 laboratory methods, data processing, and quality control are described on the TOPMed website
268 (<https://www.nhlbiwgs.org/methods>). Briefly, ~30X WGS was conducted using Illumina HiSeq X Ten
269 instruments at six sequencing centers. At the Center for Statistical Genetics at University of Michigan,
270 TOPMed sequence data were mapped to the GRCh38 human genome reference sequence in a manner

271 consistent with the joint CCDG/TOPMed functionally equivalent read mapping pipeline¹². Joint
272 genotype calling on all samples in Freeze 8 used the GotCloud pipeline¹³. Variants were filtered using a
273 Support Vector Machine (SVM) implemented in the libsvm software package. Sample-level quality
274 assurance steps included concordance between annotated and genetic sex, between prior SNP array
275 genotyping and WGS-derived genotypes, and between observed and expected relatedness and pedigree
276 information from cleaned sequence data.

277

278 **Common Variant Association Analysis**

279 We performed multi-population WGS association analysis of BMI using GENESIS¹⁴ on the
280 Analysis Commons (<http://analysiscommons.com>)¹⁵ computation platform. Analyses were performed
281 using linear mixed models (LMM). To improve power and control for false positives with a non-normal
282 phenotype distribution, we implemented a fully adjusted two-stage procedure for rank-normalization
283 when fitting the null model¹⁶. The first model was fit by adjusting BMI for age, age squared, sex, study,
284 population group, ancestry-representative PCs generated using PC-AiR¹⁷, sequencing center, sequencing
285 phase and project. A 4th degree sparse empirical kinship matrix (KM) computed with PC-Relate¹⁸ was
286 included to account for genetic relatedness among participants. We also allowed for heterogeneous
287 residual variances across sex by population group (e.g., female European), as it has previously been
288 shown to improve control of genomic inflation¹⁹. Residuals from the first model were rank normal
289 transformed within population group and sex strata. The resulting transformed residuals were used to fit
290 the second stage null model allowing for heterogeneous variances by the population group and sex strata
291 and accounting for relatedness using the kinship matrix. Variants with a MAF of at least 0.5% were then
292 tested individually.

293 Due to the large number of variants tested (N = 90,142,062) in the multi-population analysis, we
294 adopted a significance threshold of 5×10^{-9} as has been used previously²⁰. Group-specific analyses were
295 conducted in the two largest population groups, European and African.

296

297 **Replication Analyses**

298 For the novel single-variant association identified in our discovery analyses, we requested
299 replication from five independent cohorts of similar race, ethnicity, and continental ancestry to our
300 discovery populations ($N_{\text{total}} = 109,748$): Multiethnic Cohort (MEC)²¹, Million Veteran Program (MVP)
301 ^{22,23}, BioMe BioBank Program²⁴, UK Biobank (UKBB)^{25,26}, and the REasons for Geographic And Racial
302 Differences in Stroke (REGARDS) study²⁷. Phenotypes were developed and analyses were carried out
303 under the same protocol as outlined above. We subsequently conducted inverse variance weighted fixed
304 effects meta-analysis in METAL²⁸, using study-specific summary results. Additional details on the parent
305 study design for each replication study are included in the **Supplementary Note**.

306

307 **Conditional Analysis**

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

315 To determine whether association signals in known loci were independent of known signals, we
316 performed conditional analyses using previously published index variants^{5,29-48}. Specifically, we analyzed
317 all genome-wide significant variants that were not the previously reported index variants but located
318 within 500 kb of a known GWAS SNP. Given that these are potential new signals in regions known to
319 influence BMI, index variants were considered independent if the estimated effect (β) value remained \geq
320 90% of the unconditioned β value and $P < 6.25 \times 10^{-3}$ (0.05/8 loci tested). LDlink was used to calculate
321 pairwise LD between potentially independent signals in known loci and produce LD heatmaps using the
322 1000 Genomes Global reference panel⁴⁹.

323

324 **Rare Variant Aggregate Association Analysis**

325 Rare variants with a $MAF \leq 1\%$ were tested in aggregate by gene unit across studies in the multi-
326 population analysis. Variants were grouped into gene units in reference to GENCODE v28, including
327 both coding variants and variants falling within gene-associated non-coding elements. Coding variants
328 included high-confidence loss of function variants (Ensembl Variant Effect Predictor [VEP] LoF = HC),
329 missense variants (MetaSVM score > 0) and in-frame insertion/deletions or synonymous variants
330 (FATHNMM-XF coding score > 0.5). In addition to coding variants, we included variants falling within
331 the promoter of each transcript tested. Promoter regions were defined as falling in the 5 kb region 5' of
332 the transcript and also overlaying a FANTOM5 Cage Peak⁵⁰. In order to identify regulatory elements
333 likely to be acting through the tested gene, we leveraged the GeneHancer database⁵¹. GeneHancer
334 identifies enhancer regions and associates them with the specific genes they are likely to regulate,
335 allowing us to aggregate regulatory regions by the likely target gene. GeneHancer regions were limited to
336 the top 50% scored regions and variants falling in these regulatory elements were further filtered to those
337 most likely to have a functional impact (FATHMM-MKL noncoding score > 0.75). Variants aggregated
338 to gene units were tested using variant set mixed model association tests (SMMAT)⁵². Variants were
339 weighted inversely to their MAF using a beta distribution density function with parameters 1 and 25.
340 Genes were considered significantly associated after Bonferroni correction for the number of genes
341 analyzed ($P < 5 \times 10^{-7}$).

342

343 **Fine-Mapping**

344 In order to identify candidate functional variants underlying association regions, we performed
345 fine-mapping analyses in our multi-population GWAS single variant association summary statistics, using
346 the program PAINTOR⁵³ which integrates the association strength and genomic functional annotation.
347 We used the annotation file from aggregate-based testing described above under “Rare Variant Aggregate
348 Association Analysis” to identify deleterious coding variants, variants within GeneHancer regions, and

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

353

354 **PheWAS**

355 To identify potential novel phenotypic associations with newly discovered variants, we performed
356 a phenome-wide association (PheWAS) in the MyCode Community Health Initiative Study (MyCode), a
357 hospital-based population study in central and northeastern Pennsylvania⁵⁴, and in the Charles Bronfman
358 Institute for Personalized Medicine's BioMe BioBank Program (BioMe) located in New York City⁵⁵;
359 both studies had genetic data linked to electronic health records (EHR). ICD-10-CM and ICD-9-CM
360 codes were mapped to unique PheCodes using the Phecode Map v1.2⁵⁶ from the EHR. Cases were
361 defined if individuals had two or more PheCodes on separate dates, while controls had zero instances of
362 the relevant PheCode. We performed association analyses on PheCodes with $N \geq 20$ cases and 20 controls
363 using logistic regressions, adjusting for current age, sex (for non-sex-specific PheCodes), and the first 15
364 PCs calculated from genome-wide data, and assuming an additive genetic model using the PheWAS
365 package⁵⁷ in R. Analyses were conducted in the overall population as well as in individuals of
366 genetically-informed African ancestry alone (as inferred from k-means clustering of the PCs⁵⁸), given the
367 potential population-specific association of our novel locus. We restricted our analyses to unrelated
368 individuals up to 2nd degree. Association analyses were conducted within each study, followed by inverse
369 variance weighted fixed effects meta-analysis in METAL²⁸. PheCodes were deemed statistically
370 significant after Bonferroni correction for the number of PheCodes analyzed ($P < 0.05/538 = 9.3 \times 10^{-5}$).

371

372 RESULTS

373 Single-variant Analyses

374 Among the 90 million SNPs included in the multi-population analysis, 86% (N = 77,178,487)
375 were rare SNPs with a study-wide MAF of $0.5\% < \text{MAF} \leq 1\%$, and 6% (N = 5,542,150) were low-
376 frequency ($1\% < \text{MAF} \leq 5\%$) SNPs. In the multi-population unconditional analysis, we identified 16 loci
377 that reached the prespecified genome-wide significance threshold of $P < 5 \times 10^{-9}$ (**Table 1, Figure 2,**
378 **Supplementary Figs. 2 – 3**), including one low-frequency (MAF = 4%) and 15 common (MAF 14% –
379 50%) tag SNPs. In general, the low-frequency variant in our primary discovery showed a stronger effect
380 than the common variants, with an estimated effect 2.14 times larger than the average common variants
381 (0.078 vs. 0.037 on average). Of these 16 loci, 15 were in known BMI-associated regions, and one novel
382 locus was identified on chromosome 22 harboring a low-frequency index SNP near *MTMR3*
383 (rs111490516; MAF = 4%, $\beta = 0.078$, SE = 0.013, $P = 4.52 \times 10^{-9}$; **Table 1**). The MAF of this *MTMR3*
384 locus varied widely across population groups, with the highest MAF observed in the African (13%) and
385 Barbadian (13%) population groups, while it ranged from 0% to 5% in other population groups
386 (**Supplementary Data 5**).

387 In the two population-specific analyses, 10 association signals reached genome-wide significance
388 (**Supplementary Data 6, Supplementary Figs 4 – 7**). Two of these signals were also detected in the
389 multi-population analysis. For two loci, *SEC16B* and *FTO*, each population-specific analysis revealed a
390 distinct lead variant compared to the multi-population analysis; however, they were in high LD with ($R^2 =$
391 0.95 and $R^2 = 1.00$, according to TOP-LD⁵⁹; **Supplementary Data 6**) and within 30 kb of the multi-
392 population lead SNPs. Notably, the novel locus in *MTMR3* achieved significance exclusively in the
393 African group. While the most significant SNP in the African population group (rs73396827) differed
394 from that in the multi-population analysis (rs111490516), the two were in strong LD in the TOPMed
395 African population ($R^2 = 1.00$). Both of these SNPs were fixed in the European group (MAF = 0%). In the
396 European group analysis, one SNP in the *ALKAL2* locus on chromosome 2 (rs62107261, $\beta = -0.102$, SE =

397 0.016, $P = 2.08 \times 10^{-10}$, MAF = 5%) was not in LD with the corresponding lead variant in the multi-
398 population analysis ($R^2 = 0.00$, as calculated in the analysis subset), but was a known independent
399 secondary signal at this locus⁴⁰. The remaining SNPs were in the proximity to the index SNPs in the
400 corresponding loci from the multi-population analysis.

401

402 **Replication**

403 The replication sample sizes ranged from 4,413 in BioMe to 79,889 in MVP (**Supplementary**
404 **Data 7**). In the five replication studies of Blacks, Africans, and African Americans, the MAF of
405 rs111490516 in *MTMR3* ranged from 11% to 13%, aligning with the 13% observed in our African and
406 Barbadian groups and contrasting to the 0% to 5% range in our non-African discovery groups
407 (**Supplementary Data 7**). We replicated the novel variant rs111490516, demonstrating directionally
408 consistent associations with BMI across the replication studies and a 68% reduction in the estimated
409 effect when meta-analyzing across replication studies ($\beta = 0.025$, $SE = 0.007$, $P = 4.76 \times 10^{-4}$, MAF =
410 11%) compared to the discovery analysis (**Figure 3, Supplementary Data 7**). In the meta-analysis of
411 198,621 individuals from both discovery and replication studies, the estimated effect was 0.037 with a SE
412 of 0.006 and a P -value of 4.19×10^{-9} (**Figure 3, Supplementary Data 7**).

413 To gain a better understanding of the potential functional consequence of the *MTMR3* locus, we
414 used Ensembl VEP⁶⁰ to annotate all variants in high LD with our top SNP ($R^2 > 0.8$ in the African
415 population group using TOP-LD⁵⁹). Of the 54 variants in high LD, most were intronic or nearby *MTMR3*
416 (**Supplementary Data 8**). Of these, four variants had a moderate CADD (combined annotation dependent
417 depletion) score (scaled CADD > 10) with rs73394881 having the highest relative CADD score⁶¹, three of
418 which lay within a possible enhancer (rs73396896, $R^2 = 0.884$; rs73394881, $R^2 = 0.889$; rs74832232, $R^2 =$
419 0.889).

420

421 **Conditional Analyses**

422 Conditional analysis using the most associated variant at each locus revealed two significant
423 secondary signals after multiple testing correction (**Tables 1, Supplementary Data 9, Supplementary**
424 **Figure 8**). These included a known BMI-associated index variant on chromosome 2 (rs62107261, $\beta = -$
425 0.097, SE = 0.014, $P = 2.06 \times 10^{-12}$, near *ALKAL2*)⁴⁰, which was also the most significant variant at this
426 locus in the European group analysis (**Supplementary Data 6**). We further identified rs78769612 on
427 chromosome 18 ($\beta = -0.100$, SE = 0.019, $P = 2.17 \times 10^{-7}$, near *MC4R*). Although both secondary SNPs
428 were in known BMI-associated loci, rs78769612 near *MC4R* was a new index variant.

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

440

441 **Aggregate-based testing**

442 We did not identify any novel gene regions through association tests at the genome-wide level (P
443 $< 5 \times 10^{-7}$) when aggregating variants with MAF $\leq 1\%$. Nevertheless, we successfully replicated previous
444 gene-based associations with the well-known melanocortin 4 receptor (*MC4R*) gene ($P = 8.47 \times 10^{-8}$),
445 with 111 alleles across 37 sites within coding regions, enhancers, and promoters for *MC4R*. The *MC4R*
446 locus was also identified in single-variant analyses.

447

448 **Fine-mapping**

449 To pinpoint the most probable causal variant(s) underlying each of the 16 loci, we subsequently
450 performed fine-mapping using PAINTOR⁵³. Assuming one causal variant per locus, the index variants
451 were the most likely causal variants in 14 loci, with posterior probabilities (PP) ranging from 0.02 and
452 1.00, and seven of them had a PP above 0.50 (**Supplementary Data 11, Supplementary Fig 10**). Two
453 intronic index variants, rs2307111 in *POC5* and rs1379871 in *DMD*, were particularly noteworthy with
454 PP exceeding 0.98. In contrast, variants with the highest PP in *ADCY3* and *ZC3H4* were not the reported
455 index variants, although the highest PP for the *ADCY3* locus was below 0.50, and thus not likely the
456 causal variant underlying this signal. In the *ZC3H4* locus, the highest PP variant (rs55731973, PP = 0.77)
457 was intronic, located in the 5' UTR or upstream of *ZC3H4* depending on alternative transcripts, and
458 resided in probable enhancer regions. Additionally, this variant was a significant cis-eQTL for *SAE1*⁶²,
459 another nearby downstream gene.

460

461 **PheWAS**

462 To explore potential novel pleiotropy, we conducted association tests between the tag variant
463 from our novel locus, rs111490516, and 538 PheCodes available in the MyCode and BioMe studies. No
464 PheCode was significantly associated with rs111490516 following multiple test correction ($P < 9.3 \times 10^{-5}$).
465 However, PheCode 327.3 (Sleep Apnea) and 327.32 (Obstructive Sleep Apnea) ranked among the top
466 associated PheCodes ($P < 0.001$) (**Supplementary Data 12, Supplementary Fig 11**). Perhaps not
467 coincidentally, obesity is one of the strongest risk factors for sleep apnea⁶³.

468

469 **DISCUSSION**

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

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

477 The novel *MTMR3* variant, rs111490516, was most common in our African and Barbadian
478 population groups (MAF = 13%) and of moderate frequency in our Dominican population group (MAF =
479 5%). We further replicated this association in study samples of similar population background. Yet,
480 previous GWAS of BMI focusing on African ancestry individuals failed to identify a significant
481 association in this region. It is not available for lookup in the most recent MVP BMI GWAS²³, although
482 included in our replication. In one of the largest GWAS meta-analyses of imputed genotype data in
483 African ancestry individuals with summary data available publicly, which was conducted by the African
484 Ancestry Anthropometry Genetics Consortium (AAAGC, N up to 42,751)³⁷, this variant was directionally
485 consistent and suggestively associated ($\beta = 0.042$, $P = 1.80 \times 10^{-4}$, MAF = 12%)³⁷. Similarly, in our
486 replication analysis of 109,748 individuals with imputed genotypes, *MTMR3* (rs111490516) was
487 suggestively significant ($\beta = 0.025$, $P = 4.76 \times 10^{-4}$, MAF = 11%). Therefore, the lack of discovery in
488 prior publications is likely not due to insufficient power. As indicated by our fine-mapping results for this
489 novel locus, our index SNP is likely not causal but could be in LD with a causal SNP and also poorly
490 captured in studies relying on imputation. In other words, the causal variant underlying this locus may be
491 nearby, less frequent, and on an LD block more frequent in a population poorly represented in other
492 imputation reference panels, but well represented in our WGS and highly diverse sample (e.g., Caribbean
493 admixed individuals). In this case, one would require sequencing data in a large sample size with the
494 relevant haplotype to detect a significant association that was not able to be identified with imputation in
495 a similar number of people.

496 The SNP rs111490516 lies in an intron of the *MTMR3* (myotubularin related protein 3) gene, with
497 limited evidence of involvement in regulatory or functional protein activity. Other variants mapped to
498 *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 ($\beta =$
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 ($\beta = -$
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 autophagy⁶⁶. 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 individuals⁶⁷. A
507 potential mechanism was proposed later suggesting *MTMR3* may mediate the association between
508 miRNA-4513 and total cholesterol⁶⁸. 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 previously⁴⁷. 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 heterogeneous effects, self-identity (population groups), and ancestry (genotype
545 PCs). As has been shown by others⁴⁷, 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.

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

555

556

557 REFERENCES

558

- 559 1. Collaborators, G.B.D.O. *et al.* Health Effects of Overweight and Obesity in 195 Countries over
560 25 Years. *N Engl J Med* **377**, 13-27 (2017).
- 561 2. Chong, B. *et al.* Trends and predictions of malnutrition and obesity in 204 countries and
562 territories: an analysis of the Global Burden of Disease Study 2019. *EClinicalMedicine* **57**,
563 101850 (2023).
- 564 3. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243-250
565 (2022).
- 566 4. Loos, R.J.F. & Yeo, G.S.H. The genetics of obesity: from discovery to biology. *Nat Rev Genet*
567 **23**, 120-133 (2022).
- 568 5. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology.
569 *Nature* **518**, 197-206 (2015).
- 570 6. Speakman, J.R., Loos, R.J.F., O'Rahilly, S., Hirschhorn, J.N. & Allison, D.B. GWAS for BMI: a
571 treasure trove of fundamental insights into the genetic basis of obesity. *Int J Obes (Lond)* **42**,
572 1524-1531 (2018).

- 573 7. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv*,
574 588020 (2019).
- 575 8. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
576 *bioRxiv*, 563866 (2019).
- 577 9. Abel, H.J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes.
578 *Nature* **583**, 83-89 (2020).
- 579 10. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-
580 wide Association Studies. *Am J Hum Genet* **105**, 763-772 (2019).
- 581 11. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
582 *Nature* **590**, 290-299 (2021).
- 583 12. Regier, A.A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables
584 harmonized variant calling across human genetics projects. *Nat Commun* **9**, 4038 (2018).
- 585 13. Jun, G., Wing, M.K., Abecasis, G.R. & Kang, H.M. An efficient and scalable analysis framework
586 for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25**,
587 918-25 (2015).
- 588 14. Gogarten, S.M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package.
589 *Bioinformatics* **35**, 5346-5348 (2019).
- 590 15. Brody, J.A. *et al.* Analysis commons, a team approach to discovery in a big-data environment for
591 genetic epidemiology. *Nat Genet* **49**, 1560-1563 (2017).
- 592 16. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association
593 studies. *Genet Epidemiol* **43**, 263-275 (2019).
- 594 17. Conomos, M.P., Miller, M.B. & Thornton, T.A. Robust inference of population structure for
595 ancestry prediction and correction of stratification in the presence of relatedness. *Genet*
596 *Epidemiol* **39**, 276-93 (2015).
- 597 18. Conomos, M.P., Reiner, A.P., Weir, B.S. & Thornton, T.A. Model-free Estimation of Recent
598 Genetic Relatedness. *Am J Hum Genet* **98**, 127-48 (2016).

- 599 19. Conomos, M.P. *et al.* Genetic Diversity and Association Studies in US Hispanic/Latino
600 Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am J Hum*
601 *Genet* **98**, 165-84 (2016).
- 602 20. Lin, D.Y. A simple and accurate method to determine genomewide significance for association
603 tests in sequencing studies. *Genet Epidemiol* **43**, 365-372 (2019).
- 604 21. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am*
605 *J Epidemiol* **151**, 346-57 (2000).
- 606 22. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on
607 health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).
- 608 23. Huang, J. *et al.* Genomics and phenomics of body mass index reveals a complex disease network.
609 *Nat Commun* **13**, 7973 (2022).
- 610 24. Abul-Husn, N.S. *et al.* Implementing genomic screening in diverse populations. *Genome Med* **13**,
611 17 (2021).
- 612 25. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range
613 of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
- 614 26. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*
615 **562**, 203-209 (2018).
- 616 27. Howard, V.J. *et al.* The reasons for geographic and racial differences in stroke study: objectives
617 and design. *Neuroepidemiology* **25**, 135-43 (2005).
- 618 28. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide
619 association scans. *Bioinformatics* **26**, 2190-1 (2010).
- 620 29. Wen, W. *et al.* Meta-analysis identifies common variants associated with body mass index in east
621 Asians. *Nat Genet* **44**, 307-11 (2012).
- 622 30. Okada, Y. *et al.* Common variants at CDKAL1 and KLF9 are associated with body mass index in
623 east Asian populations. *Nat Genet* **44**, 302-6 (2012).

- 624 31. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and
625 provides insights into genetic architecture. *Nat Genet* **45**, 501-12 (2013).
- 626 32. Monda, K.L. *et al.* A meta-analysis identifies new loci associated with body mass index in
627 individuals of African ancestry. *Nat Genet* **45**, 690-6 (2013).
- 628 33. Gong, J. *et al.* Fine Mapping and Identification of BMI Loci in African Americans. *Am J Hum*
629 *Genet* **93**, 661-71 (2013).
- 630 34. Wen, W. *et al.* Meta-analysis of genome-wide association studies in East Asian-ancestry
631 populations identifies four new loci for body mass index. *Hum Mol Genet* **23**, 5492-504 (2014).
- 632 35. Winkler, T.W. *et al.* The Influence of Age and Sex on Genetic Associations with Adult Body Size
633 and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet* **11**, e1005378 (2015).
- 634 36. Fernandez-Rhodes, L. *et al.* Trans-ethnic fine-mapping of genetic loci for body mass index in the
635 diverse ancestral populations of the Population Architecture using Genomics and Epidemiology
636 (PAGE) Study reveals evidence for multiple signals at established loci. *Hum Genet* **136**, 771-800
637 (2017).
- 638 37. Ng, M.C.Y. *et al.* Discovery and fine-mapping of adiposity loci using high density imputation of
639 genome-wide association studies in individuals of African ancestry: African Ancestry
640 Anthropometry Genetics Consortium. *PLoS Genet* **13**, e1006719 (2017).
- 641 38. Justice, A.E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking
642 behaviour identifies novel loci for obesity traits. *Nat Commun* **8**, 14977 (2017).
- 643 39. Graff, M. *et al.* Genome-wide physical activity interactions in adiposity - A meta-analysis of
644 200,452 adults. *PLoS Genet* **13**, e1006528 (2017).
- 645 40. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic
646 Signals for Anthropometric Traits. *Am J Hum Genet* **100**, 865-884 (2017).
- 647 41. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index
648 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).

- 674 53. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-
675 mapping studies. *PLoS Genet* **10**, e1004722 (2014).
- 676 54. Carey, D.J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-
677 linked biobank for precision medicine research. *Genet Med* **18**, 906-13 (2016).
- 678 55. Abul-Husn, N.S. & Kenny, E.E. Personalized Medicine and the Power of Electronic Health
679 Records. *Cell* **177**, 58-69 (2019).
- 680 56. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and
681 Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).
- 682 57. Carroll, R.J., Bastarache, L. & Denny, J.C. R PheWAS: data analysis and plotting tools for
683 phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-6 (2014).
- 684 58. Dewey, F.E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome
685 sequences from the DiscovEHR study. *Science* **354**(2016).
- 686 59. Huang, L. *et al.* TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome
687 sequence data. *Am J Hum Genet* **109**, 1175-1181 (2022).
- 688 60. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
- 689 61. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic
690 variants. *Nat Genet* **46**, 310-5 (2014).
- 691 62. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
- 692 63. Jehan, S. *et al.* Obstructive Sleep Apnea and Obesity: Implications for Public Health. *Sleep Med*
693 *Disord* **1**(2017).
- 694 64. Hubel, C. *et al.* Genomics of body fat percentage may contribute to sex bias in anorexia nervosa.
695 *Am J Med Genet B Neuropsychiatr Genet* **180**, 428-438 (2019).
- 696 65. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat*
697 *Genet* **53**, 1415-1424 (2021).
- 698 66. Vergne, I. & Deretic, V. The role of PI3P phosphatases in the regulation of autophagy. *FEBS Lett*
699 **584**, 1313-8 (2010).

- 700 67. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**,
701 1274-1283 (2013).
- 702 68. Ghanbari, M. *et al.* A genetic variant in the seed region of miR-4513 shows pleiotropic effects on
703 lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Hum Mutat* **35**, 1524-
704 31 (2014).
- 705 69. Geng, Z. *et al.* RNA-Seq analysis of obese Pdha1(fl/fl)Lyz2-Cre mice induced by a high-fat diet.
706 *Exp Anim* (2022).
- 707 70. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting
708 evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
- 709 71. ClinVar, N.C.f.B.I. VCV001233562.6.
- 710 72. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv*,
711 2021.09.03.21262975 (2021).
- 712 73. Hotta, K. *et al.* Association between obesity and polymorphisms in SEC16B, TMEM18,
713 GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *J Hum Genet* **54**, 727-31 (2009).
- 714
715
716
717
718
719
720
721
722

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	<i>SEC16B</i>	rs543874	A	G	20%	0.064	0.006	1.38E-26	Yes	No
2	621558	<i>TMEM18</i>	rs939584	C	T	85%	0.058	0.007	1.99E-17	No	No
2	24927427	<i>ADCY3</i>	rs10182181	A	G	56%	0.035	0.005	1.76E-11	Yes	No
3	186108951	<i>ETV5</i>	rs869400	T	G	82%	0.038	0.006	1.21E-09	No	No
4	45179317	<i>GNPDA2</i>	rs12507026	A	T	36%	0.045	0.005	9.55E-19	Yes	No
5	75707853	<i>POC5</i>	rs2307111	T	C	55%	-0.032	0.005	7.43E-10	Yes	No
6	50830813	<i>TFAP2B</i>	rs2206277	C	T	19%	0.054	0.006	2.05E-18	Novel	No
8	76068626	<i>HNF4G</i>	rs830463	A	G	47%	0.031	0.005	6.58E-10	No	No
11	27657463	<i>BDNF</i>	rs3838785	GT	G	58%	-0.030	0.005	3.14E-09	Novel	No
12	49853685	<i>BCDIN3D</i>	rs7138803	G	A	30%	0.036	0.005	1.69E-11	Yes	No
13	53533448	<i>OLFM4</i>	rs9568868	G	T	14%	0.047	0.007	5.73E-11	No	No
16	53767042	<i>FTO</i>	rs1421085	T	C	29%	0.090	0.006	6.11E-59	Yes	No
18	60161902	<i>MC4R</i>	rs6567160	T	C	21%	0.053	0.006	8.22E-19	Yes	No
19	47077985	<i>ZC3H4</i>	rs28590228	C	T	50%	0.033	0.005	4.75E-10	No	No
22	29906934	<i>MTMR3</i>	rs111490516	C	T	4%	0.078	0.013	4.52E-09	Novel	Yes
X	31836665	<i>DMD</i>	rs1379871	G	C	41%	0.029	0.004	1.35E-11	Yes	No
Secondary signals											
2	422144	<i>ALKAL2</i>	rs62107261	T	C	3%	-0.095	0.014	3.83E-12	Yes	No
18	60361739	<i>MC4R</i>	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 \pm 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 \pm 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.

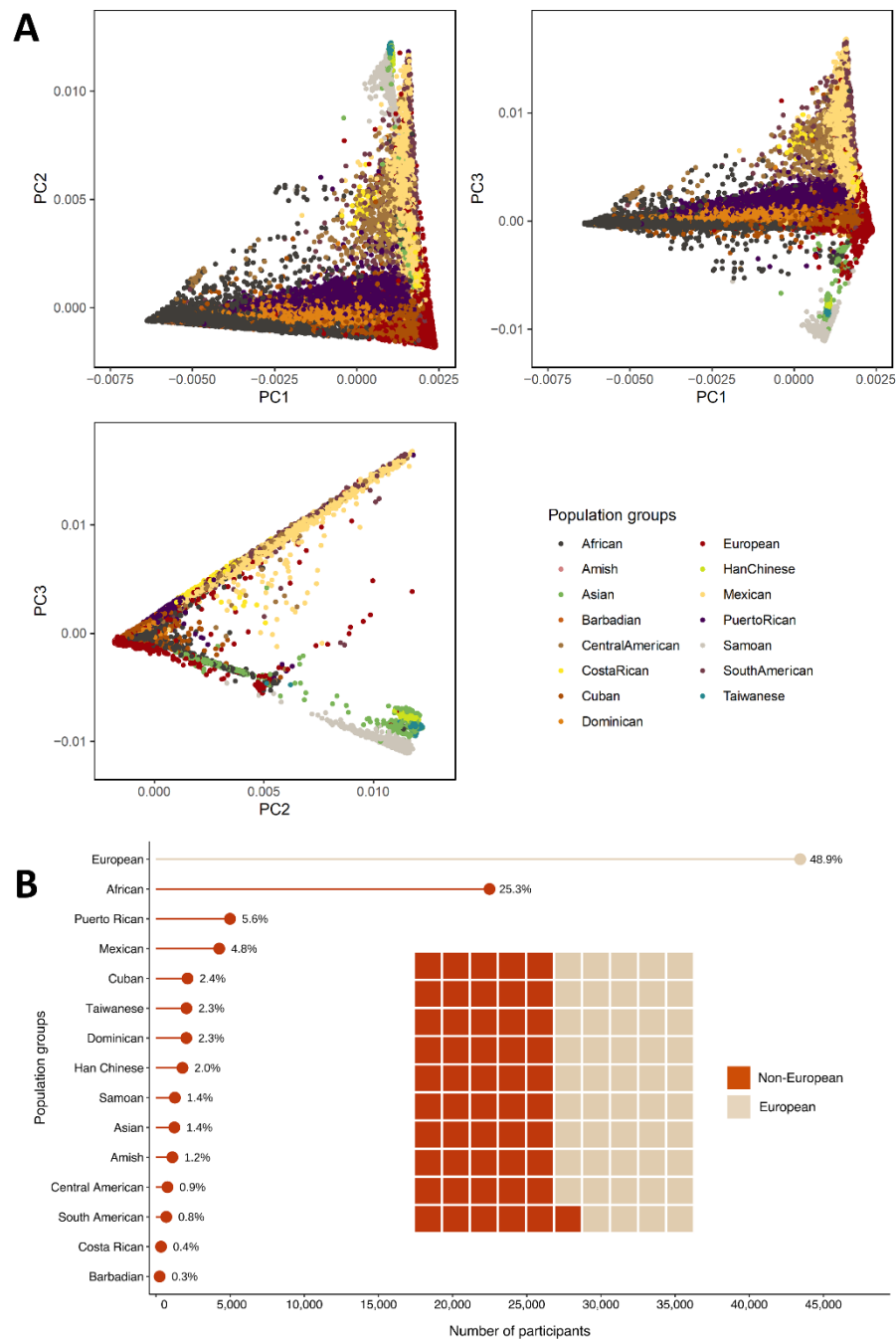


Figure 2. Summary of significant association findings.

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.

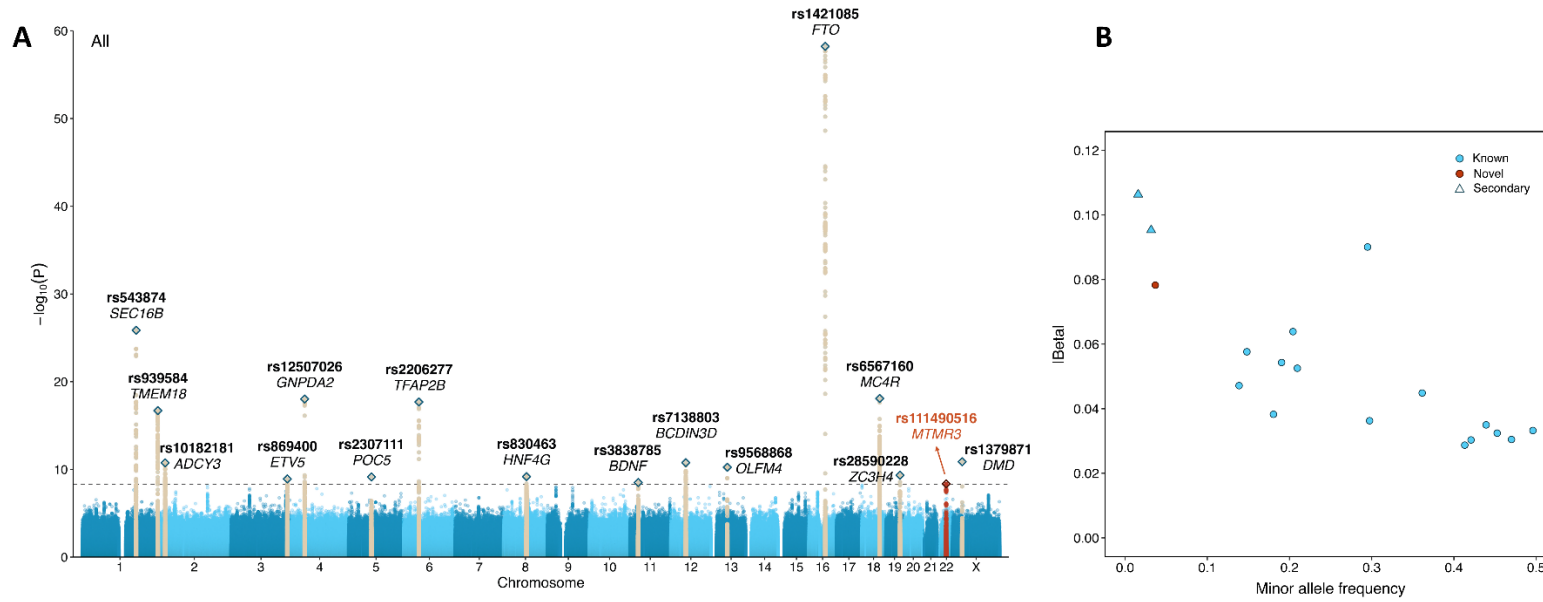


Figure 3. Forest plot of rs111490516 replication. All effect estimates (95% confidence interval) are oriented on the BMI increasing allele and are provided as standard deviation per allele. Actual beta values and *P*-values are in Supplementary Data 7.

