

1 **Human knockouts in a cohort with a high rate of consanguinity**

2

3 Danish Saleheen^{1,2,†*}, Pradeep Natarajan^{3,4,†}, Wei Zhao¹, Asif Rasheed², Sumeet
4 Khetarpal⁵, Hong-Hee Won^{3,4}, Konrad J. Karczewski^{4,6}, Anne H. O'Donnell-Luria^{4,6,7},
5 Kaitlin E. Samocha⁶, Namrata Gupta⁴, Mozzam Zaidi², Maria Samuel², Atif Imran²,
6 Shahid Abbas⁸, Faisal Majeed², Madiha Ishaq², Saba Akhtar², Kevin Trindade⁵, Megan
7 Mucksavage⁵, Nadeem Qamar⁹, Khan Shah Zaman⁹, Zia Yaqoob⁹, Tahir Saghir⁹, Syed
8 Nadeem Hasan Rizvi⁹, Anis Memon⁹, Nadeem Hayyat Mallick¹⁰, Mohammad Ishaq¹¹,
9 Syed Zahed Rasheed¹¹, Fazal-ur-Rehman Memon¹², Khalid Mahmood¹³, Naveeduddin
10 Ahmed¹⁴, Ron Do^{15,16}, Daniel G. MacArthur^{4,6}, Stacey Gabriel⁴, Eric S. Lander⁴, Mark J.
11 Daly^{4,6}, Philippe Frossard^{2,†}, John Danesh^{17,18,†}, Daniel J. Rader^{5,19,†}, Sekar
12 Kathiresan^{3,4,†*}

13

14 †Contributed equally

15

16 ¹ Department of Biostatistics and Epidemiology, Perelman School of Medicine at the
17 University of Pennsylvania, Philadelphia, PA, USA

18 ² Center for Non-Communicable Diseases, Karachi, Pakistan

19 ³ Center for Human Genetic Research and Cardiovascular Research Center,
20 Massachusetts General Hospital, Boston, MA, USA

21 ⁴ Broad Institute of Harvard and MIT, Cambridge, MA, USA

22 ⁵ Division of Translational Medicine and Human Genetics, Department of Medicine,
23 Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

- 24 ⁶ Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts
25 General Hospital and Harvard Medical School, Boston, MA
- 26 ⁷ Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
- 27 ⁸ Faisalabad Institute of Cardiology, Faisalabad, Pakistan
- 28 ⁹ National Institute of Cardiovascular Disorders, Karachi, Pakistan
- 29 ¹⁰ Punjab Institute of Cardiology, Lahore, Pakistan
- 30 ¹¹ Karachi Institute of Heart Diseases, Karachi, Pakistan
- 31 ¹² Red Crescent Institute of Cardiology, Hyderabad, Pakistan
- 32 ¹³ The Civil Hospital, Karachi, Pakistan
- 33 ¹⁴ Liaquat National Hospital, Karachi, Pakistan
- 34 ¹⁵ Department of Genetics and Genomic Sciences, Mount Sinai Medical Center, Icahn
35 School of Medicine at Mount Sinai, New York, NY, USA
- 36 ¹⁶ The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at
37 Mount Sinai, New York, NY, USA
- 38 ¹⁷ Department of Public Health and Primary Care, University of Cambridge, UK
- 39 ¹⁸ Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
- 40 ¹⁹ Department of Human Genetics, University of Pennsylvania, USA
- 41
- 42 *Corresponding authors:
- 43 Danish Saleheen, MBBS, PhD
44 Department of Biostatistics and Epidemiology
45 University of Pennsylvania
46 11-134 Translational Research Center

47 3400 Civic Center Boulevard
48 Philadelphia, PA 19104
49 Tel: 215-573-6323
50 Fax: 215-573-2094
51 Email: saleheen@mail.med.upenn.edu
52
53 Sekar Kathiresan, MD
54 Broad Institute and Massachusetts General Hospital
55 CPZN 5.252
56 185 Cambridge Street
57 Boston, MA 02114
58 Tel: 617-643-6120
59 Email: sekar@broadinstitute.org
60
61 Word Count, Summary Paragraph: 253
62 Word Count, Main Text: 1,977

63 **Summary Paragraph**

64 A major goal of biomedicine is to understand the function of every gene in the human
65 genome.¹ Null mutations can disrupt both copies of a given gene in humans and
66 phenotypic analysis of such ‘human knockouts’ can provide insight into gene function.
67 To date, comprehensive analysis of genes knocked out in humans has been limited by the
68 fact that null mutations are infrequent in the general population and so, observing an
69 individual homozygous null for a given gene is exceedingly rare.^{2,3} However,
70 consanguineous unions are more likely to result in offspring who carry homozygous null
71 mutations. In Pakistan, consanguinity rates are notably high.⁴ Here, we sequenced the
72 protein-coding regions of 7,078 adult participants living in Pakistan and performed
73 phenotypic analysis to identify homozygous null individuals and to understand
74 consequences of complete gene disruption in humans. We enumerated 36,850 rare (<1 %
75 minor allele frequency) null mutations. These homozygous null mutations led to
76 complete inactivation of 961 genes in at least one participant. Homozygosity for null
77 mutations at *APOC3* was associated with absent plasma apolipoprotein C-III levels; at
78 *PLAG27*, with absent enzymatic activity of soluble lipoprotein-associated phospholipase
79 *A2*; at *CYP2F1*, with higher plasma interleukin-8 concentrations; and at either *A3GALT2*
80 or *NRG4*, with markedly reduced plasma insulin C-peptide concentrations. After
81 physiologic challenge with oral fat, *APOC3* knockouts displayed marked blunting of the
82 usual post-prandial rise in plasma triglycerides compared to wild-type family members.
83 These observations provide a roadmap to understand the consequences of complete
84 disruption of a large fraction of genes in the human genome.

85 **Main Text**

86 We studied adult participants in the Pakistan Risk of Myocardial Infarction Study
87 (PROMIS) designed to understand the determinants of cardiometabolic diseases in South
88 Asians.⁵ Consanguineous marriages have been common in this region of South Asia for
89 many generations.⁶ In PROMIS, 38.3% of participants reported that their parents were
90 cousins and 38.1% reported themselves being married to a cousin. An expectation from
91 consanguinity is long regions of autozygosity, defined as homozygous loci identical by
92 descent.⁷ Using genome-wide genotyping data available in 17,744 PROMIS participants,
93 we quantified the length of runs of homozygosity, defined as homozygous segments at
94 least 1.5 megabases long. We compared the lengths of runs of homozygosity among
95 PROMIS participants with that seen in other populations from the International HapMap3
96 project. Median length of genome-wide homozygosity among PROMIS participants was
97 6-7 times higher than participants of European (CEU, TSI) ($P = 3.6 \times 10^{-37}$), East Asian
98 (CHB, JPT, CHD) ($P = 5.4 \times 10^{-48}$) and African ancestries (YRI, MKK) ($P = 1.3 \times 10^{-40}$),
99 respectively (**Fig. 1**).

100 In order to characterize the burden of rare homozygous null alleles, we performed
101 whole exome sequencing in 7,078 PROMIS participants (**Table 1**). Across all
102 participants, 1,303,689 protein-coding and splice-site sequence variants in 20,008
103 autosomal genes passed variant quality control metrics. Of these, 47,656 mutations across
104 13,645 autosomal genes were annotated as null (nonsense, frameshift, or canonical
105 splice-site). To increase the probability that mutations annotated as null are *bona fide*, we
106 removed nonsense and frameshift mutations occurring within the last 5% of the transcript
107 and within exons flanked by non-canonical splice sites, splice site mutations at small

108 (<15 bp) introns, at non-canonical splice sites, and where the purported null allele is
109 observed across primates. Common null alleles are less likely to exert strong functional
110 effects as they are less constrained by purifying selection; thus, we limit our analyses in
111 the rest of the manuscript to null mutations with a minor allele frequency (MAF) of <
112 1%.

113 Applying these criteria, we generated a set of 36,850 null mutations across 12,131
114 autosomal genes.⁸ The site-frequency spectrum for these null mutations revealed that the
115 majority was seen only in one or a few individuals (**Fig. 2**).

116 We compared the coefficient of inbreeding (F coefficient) in PROMIS
117 participants with that of 15,249 individuals from outbred populations of European or
118 African American ancestry. The F coefficient estimates the excess homozygosity
119 compared with an estimated outbred ancestor. PROMIS participants had a 4-fold higher
120 median inbreeding coefficient compared to outbred populations (0.016 v 0.0041; $P < 2 \times$
121 10^{-16}) (**Fig. 3A**). Additionally, those in PROMIS who reported that their parents were
122 closely related had even higher median inbreeding coefficients than those who did not
123 (0.024 v 0.013; $P < 2 \times 10^{-16}$).

124 Across all 7,078 PROMIS participants, both copies of 961 distinct genes were
125 disrupted due to null mutations with MAF < 1%. A full listing of all 961 genes knocked
126 out, the number of knockout participants for each gene, and the specific null mutation(s)
127 are provided in **Supplementary Table 1**. 697 (72.5 %) of the genes were knocked out
128 only in one participant (**Supplementary Figure 1**). About 1 in 5 sequenced participants
129 (1,306 individuals, 18.4 %) had at least one gene knocked by a homozygous null
130 mutation with MAF < 1%. 1,081 of these 1,306 individuals (82.8 %) had complete

131 deficiency for one gene, but a minority of participants were knockouts for more than one
132 gene and one participant had six genes completely inactivated. The F inbreeding
133 coefficient was correlated with the number of homozygous null genes present in each
134 individual. (Spearman $r = 0.29$; $P = 3 \times 10^{-133}$) (**Fig. 3B**).

135 We tested the hypothesis that genes observed in the homozygous null state in
136 PROMIS participants are under less evolutionary constraint. We calculated the
137 probability of being loss-of-function intolerant (at >90% threshold) for each gene (see
138 Methods)^{9,10} and compared this to 961 randomly selected genes. The observed 961
139 homozygous null genes were less likely to be classified as highly constrained (odds ratio
140 0.10; 95% CI 0.095, 0.11; $P < 1 \times 10^{-10}$).

141 We next sought to understand the phenotypic consequences of complete
142 disruption of any of these 961 genes. We applied two approaches. First, for 264 genes
143 where two or more participants were homozygous null, we conducted an association
144 screen against a panel of 201 phenotypic traits (**Supplementary Table 2**). Second, at a
145 single gene, we recalled participants based on genotype across three classes ('wild-type',
146 heterozygous null, and homozygous null) and performed provocative physiologic testing.

147 At 264 genes where two or more participants were homozygous null, we
148 performed association analyses to determine whether homozygous null mutation status
149 was associated with variation in any of 201 traits. For quantitative traits, we compared
150 mean trait values in homozygous null carriers with non-carriers. For dichotomous traits,
151 we performed logistic regression with trait status as the outcome variable and
152 homozygous null carrier status as the predictor variable. Details of covariate adjustments
153 are presented in the Methods. Across quantitative and dichotomous traits, this resulted in

154 the analysis of 15,263 gene-trait pairs and thus, we set Bonferroni-adjusted significance
155 threshold at $P = 3 \times 10^{-6}$.

156 The quantile-quantile plot of expected versus observed association results shows
157 an excess of highly significant results without systematic inflation (**Supplementary**
158 **Figure 2**). Association results surpassed the Bonferroni significance threshold for 14
159 gene-trait pairs (**Supplementary Table 3**). Below, we highlight four associations
160 demonstrating examples of 1) confirmed biochemical deficiency with two independent
161 assays (*PLA2G7*), 2) gene-biomarker association (*CYP2F1*), 3) pure recessive model of
162 association (*A3GALT2*), and 4) confirmed biochemical deficiency with gene-biomarker
163 association (*APOC3*).

164 Lipoprotein-associated phospholipase A2 (Lp-PLA2, encoded by *PLA2G7*)
165 hydrolyzes oxidatively-modified polyunsaturated fatty acids producing
166 lysophosphatidylcholine and oxidized nonesterified fatty acids. Higher soluble Lp-PLA2
167 activity has been correlated with higher risk for cardiovascular disease.¹¹ At *PLA2G7*, we
168 identified two participants homozygous for *PLA2G7* c.663+1G>A. When compared with
169 non-carriers, c.663+1G>A homozygotes have markedly lower Lp-PLA2 enzyme as well
170 as activity (-266 ng/ml, $P = 7 \times 10^{-5}$ for mass; -245 nmol/ml/min, $P = 2 \times 10^{-7}$ for
171 activity) whereas the 102 heterozygotes had an intermediate effect (-107 ng/ml, $P = 4 \times$
172 10^{-26} for mass; -115 nmol/ml/min, $P = 3 \times 10^{-58}$ for activity) (**Supplementary Figure 3a-**
173 **b**).

174 Cytochrome P450 2F1 (encoded by *CYP2F1*) is primarily expressed in the lung
175 and metabolizes pulmonary-selective toxins, such as cigarette smoke, and thus,
176 modulates the expression of environmentally-associated pulmonary diseases.¹² At

177 *CYP2F1*, we identified two participants homozygous for a splice-site mutation, c.1295-
178 2A>G. When compared with non-carriers, c.1295-2A>G homozygotes displayed higher
179 soluble interleukin 8 concentrations (+40.3 %; $P = 3 \times 10^{-6}$) (**Supplementary Figure 4**).
180 *CYP2F1* c.1295-2A>G heterozygotes (n = 3 assayed for interleukin 8) had a more modest
181 effect (+10.7 %; $P = 2 \times 10^{-4}$). Interleukin 8 is a mediator of acute pulmonary
182 inflammation.¹³

183 Alpha-1,3-galactosyltransferase 2 (encoded by *A3GALT2*) catalyzes the formation
184 of the Gal- α 1-3Gal β 1-4GlcNAc-R (α -gal) epitope; the biological role of this enzyme in
185 humans is uncertain.¹⁴ At *A3GALT2*, we identified two participants homozygous for a
186 frameshift mutation, p.Thr106SerfsTer4. Compared with non-carriers, p.Thr106SerfsTer4
187 homozygotes had both reduced fasting insulin C-peptide (-97.4%; $P = 6 \times 10^{-12}$) as well
188 as total fasting insulin concentrations (-91.9%; $P = 1 \times 10^{-4}$). Such an association was
189 only observed in the homozygous state (**Supplementary Figure 5**). *A3galt2*^{-/-} mice and
190 pigs have recently been shown to have glucose intolerance.^{15,16}

191 To understand if the identification of only a single homozygote in PROMIS may
192 still be informative, we performed a complementary analysis, focusing on those with the
193 most extreme standard Z scores ($|Z \text{ score}| > 5$) and requiring that there be evidence for
194 association in heterozygotes as well (see Methods). This procedure highlighted neureglin
195 4 (NRG4), a member of the epidermal growth factor family extracellular ligands which is
196 highly expressed in brown fat, particularly during adipocyte differentiation.^{17,18} At *NRG4*,
197 we identified a single participant homozygous for a frameshift mutation,
198 p.Ile75AsnfsTer23, who had nearly absent fasting insulin C-peptide concentrations (-99.3
199 %; $P = 7 \times 10^{-11}$). When compared with non-carriers, heterozygotes for *NRG4*

200 p.Ile75AsnfsTer23 (n = 7) displayed 54.5 % reduction in insulin C-peptide ($P = 6 \times 10^{-3}$).

201 Mice homozygous deleted for *Nrg4* have recently been shown to have glucose

202 intolerance.¹⁸

203 Apolipoprotein C-III (apoC-III, encoded by *APOC3*) is a major protein

204 component of chylomicrons, very low-density lipoprotein cholesterol, and high-density

205 lipoprotein cholesterol.¹⁹ We and others recently reported that null mutations in

206 heterozygous form lower plasma triglycerides as well as risk for coronary heart

207 disease.^{20,21} In both published studies, no *APOC3* homozygotes were identified despite

208 study of nearly 200,000 participants from the U.S. and Europe. However, in this study of

209 about 7,000 Pakistanis, we identified four participants homozygous for *APOC3*

210 p.Arg19Ter. When compared with non-carriers, p.Arg19Ter homozygotes displayed

211 near-absent plasma apoC-III protein (-89.4 %, $P = 5 \times 10^{-24}$), lower plasma triglyceride

212 concentrations (-61.7 %, $P = 4 \times 10^{-4}$), higher high-density lipoprotein (HDL) cholesterol

213 (+28.1 mg/dL, $P = 6 \times 10^{-9}$); and similar levels of low-density lipoprotein (LDL)

214 cholesterol ($P = 0.11$) (**Fig. 4a-d**).

215 ApoC-III functions as a brake on the metabolism of dietary fat and thus, the

216 complete lack of this protein should promote handling of ingested fat. The availability of

217 humans completely deficient in *APOC3* allowed us to test this hypothesis directly. We re-

218 contacted one homozygous null proband, his wife, and 27 of his first-degree relatives for

219 genotyping and physiologic investigation. Surprisingly, we found that the proband's wife

220 was also a null homozygote, leading to all nine children being obligate homozygotes

221 (**Fig. 4e**). In this family, we challenged homozygotes (n = 6) and non-carriers (n = 7) with

222 a 50 g/m² oral fat load followed by serial blood testing for six hours. *APOC3* p.Arg19Ter

223 carriers had significantly lower post-prandial triglyceride excursions (triglycerides area
224 under the curve 468.3 mg/dL*6 hours vs 1267.7 mg/dL*6 hours; $P = 1 \times 10^{-4}$) (**Fig. 4f**).
225 These data show that complete lack of APOC3 markedly improves clearance of plasma
226 triglycerides after a fatty meal.

227 Gene disruption in model organisms followed by phenotypic analysis has been a
228 fruitful approach to understand gene function; here, we extend this concept to the human
229 organism, leveraging naturally-occurring null mutations, consanguinity, and extensive
230 biochemical phenotyping. These results permit several conclusions.

231 First, power to identify human knockouts is improved with the study of
232 populations with high degrees of consanguinity. Using the observed median inbreeding
233 coefficient of sequenced participants, we estimate that with the sequencing of 200,000
234 Pakistanis, about 8,754 genes (95% CI, 8,669-8,834) will be completely knocked out in at
235 least one participant (**Fig. 5**). In contrast, with the sequencing of a similar number of
236 outbred individuals of European, East Asian, or African American ancestries, the
237 expected number of genes knocked out is much less at 1,382 (95% CI, 1,339-1,452),
238 1,423 (95% CI, 1,388-1,457), and 1,822 (95% CI, 1,772-1,871), respectively (minor
239 allele frequency estimates obtained from the Exome Aggregation Consortium, manuscript
240 submitted in parallel). Thus, if a similar number of individuals were sequenced across
241 different ancestries, the number of genes completely knocked out would be nearly six-
242 fold higher in Pakistanis when compared with other outbred populations. For example,
243 among >100,000 participants in Iceland, 1107 homozygous null genes were identified
244 whereas we observe nearly the same number of null genes after analysis of about 7,000
245 Pakistanis.³

246 Second, dense phenotyping can uncover a range of phenotypic consequences from
247 complete disruption of a gene as observed for *PLA2G7*, *CYP2F1*, *A3GALT2* and *NRG4*.
248 Third, recall of complete human knockouts followed by provocative testing may provide
249 physiologic insights. We used this approach to demonstrate that complete lack of
250 apolipoprotein C-III is tolerated and results in both lowered fasting triglyceride
251 concentrations as well as blunted post-prandial lipemia. Finally, to date, most human
252 genetic studies have pursued a phenotype-first (“forward” genetics) approach, beginning
253 with traits of interest followed by genetic mapping. Here, we show that it is now possible
254 to pursue a systematic genotype-first (“reverse” genetics) approach, starting with
255 homozygous null humans followed by methodical examination of a diverse set of traits.
256

257 **Methods**

258 **General overview of the Pakistan Risk for Myocardial Infarction Study (PROMIS).**

259 The PROMIS study is designed to investigate determinants of cardiometabolic diseases
260 in Pakistan. Since 2005, the study has enrolled close to 38,000 participants; the present
261 investigation included 7,078 participants. Participants aged 30-80 years were enrolled
262 from nine recruitment centers based in five major urban cities in Pakistan. Type 2
263 diabetes in the study was defined based on self-report or fasting glucose levels >125
264 mg/dL or HbA1c > 6.5 % or use of glucose lowering medications. The study was
265 approved by the institutional review board at the Center for Non-Communicable Diseases
266 (IRB: 00007048, IORG0005843, FWAS00014490) and all participants gave informed
267 consent.

268

269 **Phenotype descriptions.**

270 Non-fasting blood samples (with the time since last meal recorded) were drawn and
271 centrifuged within 45 minutes of venipuncture. Serum, plasma and whole blood samples
272 were stored at -70°C within 45 minutes of venipuncture. All samples were transported on
273 dry ice to the central laboratory at the Center for Non-Communicable Diseases (CNCD),
274 Pakistan, where serum and plasma samples were aliquoted across 10 different storage
275 vials. Samples were stored at -70°C for any subsequent laboratory analyses. All
276 biochemical assays were conducted in automated auto-analyzers. At CNCD Pakistan,
277 measurements for total-cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, and
278 creatinine were made in serum samples using enzymatic assays; whereas levels of HbA1c
279 were measured using a turbidimetric assay in whole-blood samples (Roche Diagnostics,

280 USA). For further measurements, aliquots of serum and plasma samples were transported
281 on dry ice to the Smilow Research Center, University of Pennsylvania, USA, where
282 following biochemical assays were conducted: apolipoproteins (apoA-I, apoA-II, apoB,
283 apoC-III, apoE) and non-esterified fatty acids were measured through
284 immunoturbidometric assays using kits by Roche Diagnostics or Kamiya; lipoprotein (a)
285 levels were determined through a turbidimetric assay using reagents and calibrators from
286 Denka Seiken (Niigata, Japan); LpPLA2 mass and activity levels were determined using
287 immunoassays manufactured by diaDexus (San Francisco, CA, USA); measurements for
288 insulin, leptin and adiponectin were made using radio-immunoassays by LINCO (MO,
289 USA); levels of adhesion molecules (ICAM-1, VCAM-1, P- and E-Selectin) were
290 determined through enzymatic assays by R&D (Minneapolis, MN, USA); and
291 measurements for C-reactive protein, alanine transaminase, aspartate transaminase,
292 cystatin-C, ferritin, ceruloplasmin, thyroid stimulating hormone, alkaline phosphatase,
293 sodium, potassium, chloride, phosphate, sex-hormone binding globulin were made using
294 enzymatic assays manufactured by Abbott Diagnostics (NJ, USA). Glomerular filtration
295 rate (eGFR) was estimated from serum creatinine levels using the MDRD equation.
296 ApoC-III levels were determined in an autoanalyzer using a commercially available
297 ELISA by Sekisui Diagnostics (Lexington, USA). We also measured the following 52
298 protein biomarkers by multiplex immunoassay using a customised panel on the Luminex
299 100/200 instrument by RBM (Myriad Rules Based Medicine, Austin, TX, USA): fatty
300 acid binding protein, granulocyte monocyte colony stimulating factor, granulocyte colony
301 stimulating factor, interferon gamma, interleukin-1 beta, interleukin 1 receptor,
302 interleukin 2, interleukin 3, interleukin 4, interleukin 5, interleukin 6, interleukin 7,

303 interleukin 8, interleukin 10, interleukin 18, interleukin p40, interleukin p70, interleukin
304 15, interleukin 17, interleukin 23, macrophage inflammatory protein 1 alpha, macrophage
305 inflammatory protein 1 beta, malondialdehyde-modified LDL, matrix metalloproteinase
306 2, matrix metalloproteinase 3, matrix metalloproteinase 9, nerve growth factor beta,
307 tumor necrosis factor alpha, tumor necrosis factor beta, brain-derived neurotrophic factor,
308 CD40, CD40 ligand, eotaxin, factor VII, insulin-like growth factor 1, lecithin-type
309 oxidized LDL receptor 1, monocyte chemoattractant protein 1, myeloperoxidase, N-
310 terminal prohormone of brain natriuretic peptide, neuronal cell adhesion molecule,
311 pregnancy-associated plasma protein A, soluble receptor for advanced glycation end-
312 products, sortilin, stem cell factor, stromal cell-derived factor 1, thrombomodulin, S100
313 calcium binding protein B, and vascular endothelial growth factor.

314

315 **Laboratory methods for array-based genotyping.**

316 As previously described, a genomewide association scan was performed using the
317 Illumina 660 Quad array at the Wellcome Trust Sanger Institute (Hinxton, UK) and using
318 the Illumina HumanOmniExpress at Cambridge Genome Services, UK.²² Initial quality
319 control (QC) criteria included removal of participants or single nucleotide
320 polymorphisms (SNPs) that had a missing rate >5%. SNPs with a MAF <1% and a P-
321 value of 10^{-7} for the Hardy-Weinberg equilibrium test were also excluded from the
322 analyses. In PROMIS, further QC included removal of participants with discrepancy
323 between their reported sex and genetic sex determined from the X chromosome. To
324 identify sample duplications, unintentional use of related samples (cryptic relatedness)

325 and sample contamination (individuals who seem to be related to nearly everyone in the
326 sample), identity-by-descent (IBD) analyses were conducted in PLINK.²³

327

328 **Laboratory methods for exome sequencing.**

329 **Exome sequencing.** Exome sequencing was performed at the Broad Institute.

330 Sequencing and exome capture methods have been previously described.^{24,25} A brief
331 description of the methods is provided below.

332 **Receipt/quality control of sample DNA.** Samples were shipped to the Biological
333 Samples Platform laboratory at the Broad Institute of MIT and Harvard (Cambridge, MA,
334 USA). DNA concentration was determined by PicoGreen (Invitrogen; Carlsbad, CA,
335 USA) prior to storage in 2D-barcoded 0.75 ml Matrix tubes at -20 °C in the SmarTStore
336 (RTS, Manchester, UK) automated sample handling system. Initial quality control (QC)
337 on all samples involving sample quantification (PicoGreen), confirmation of high-
338 molecular weight DNA and fingerprint genotyping and gender determination (Illumina
339 iSelect; Illumina; San Diego, CA, USA). Samples were excluded if the total mass,
340 concentration, integrity of DNA or quality of preliminary genotyping data was too low.

341 **Library construction.** Library construction was performed as previously described²⁶,
342 with the following modifications: initial genomic DNA input into shearing was reduced
343 from 3µg to 10-100ng in 50µL of solution. For adapter ligation, Illumina paired end
344 adapters were replaced with palindromic forked adapters, purchased from Integrated
345 DNA Technologies, with unique 8 base molecular barcode sequences included in the
346 adapter sequence to facilitate downstream pooling. With the exception of the palindromic
347 forked adapters, the reagents used for end repair, A-base addition, adapter ligation, and

348 library enrichment PCR were purchased from KAPA Biosciences (Wilmington, MA,
349 USA) in 96-reaction kits. In addition, during the post-enrichment SPRI cleanup, elution
350 volume was reduced to 20 μ L to maximize library concentration, and a vortexing step
351 was added to maximize the amount of template eluted.

352 **In-solution hybrid selection.** 1,973 samples underwent in-solution hybrid selection as
353 previously described²⁶, with the following exception: prior to hybridization, two
354 normalized libraries were pooled together, yielding the same total volume and
355 concentration specified in the publication. 5,263 samples underwent hybridization and
356 capture using the relevant components of Illumina's Rapid Capture Exome Kit and
357 following the manufacturer's suggested protocol, with the following exceptions: first, all
358 libraries within a library construction plate were pooled prior to hybridization, and
359 second, the Midi plate from Illumina's Rapid Capture Exome Kit was replaced with a
360 skirted PCR plate to facilitate automation. All hybridization and capture steps were
361 automated on the Agilent Bravo liquid handling system.

362 **Preparation of libraries for cluster amplification and sequencing.** Following post-
363 capture enrichment, libraries were quantified using quantitative PCR (KAPA Biosystems)
364 with probes specific to the ends of the adapters. This assay was automated using
365 Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were
366 normalized to 2nM and pooled by equal volume using the Hamilton Starlet. Pools were
367 then denatured using 0.1 N NaOH. Finally, denatured samples were diluted into strip
368 tubes using the Hamilton Starlet.

369 **Cluster amplification and sequencing.** Cluster amplification of denatured templates
370 was performed according to the manufacturer's protocol (Illumina) using HiSeq v3

371 cluster chemistry and HiSeq 2000 or 2500 flowcells. Flowcells were sequenced on HiSeq
372 2000 or 2500 using v3 Sequencing-by-Synthesis chemistry, then analyzed using RTA
373 v.1.12.4.2. Each pool of whole exome libraries was run on paired 76bp runs, with and 8
374 base index sequencing read was performed to read molecular indices, across the number
375 of lanes needed to meet coverage for all libraries in the pool.

376 **Read mapping and variant discovery.** Samples were processed from real-time base-
377 calls (RTA v.1.12.4.2 software [Bustard], converted to qseq.txt files, and aligned to a
378 human reference (hg19) using Burrows–Wheeler Aligner (BWA).²⁷ Aligned reads
379 duplicating the start position of another read were flagged as duplicates and not analysed.
380 Data was processed using the Genome Analysis ToolKit (GATK v3).²⁸⁻³⁰ Reads were
381 locally realigned around indels and their base qualities were recalibrated. Variant calling
382 was performed on both exomes and flanking 50 base pairs of intronic sequence across all
383 samples using the HaplotypeCaller (HC) tool from the GATK to generate a gVCF. Joint
384 genotyping was subsequently performed and ‘raw’ variant data for each sample was
385 formatted (variant call format (VCF)). SNVs and indel sites were initially filtered after
386 variant calibration marked sites of low quality that were likely false positives.

387 **Data analysis QC.** Fingerprint concordance between sequence data and fingerprint
388 genotypes was evaluated. Variant calls were evaluated on both bulk and per- sample
389 properties: novel and known variant counts, transition–transversion (TS–TV) ratio,
390 heterozygous–homozygous non-reference ratio, and deletion/insertion ratio. Both bulk
391 and sample metrics were compared to historical values for exome sequencing projects at
392 the Broad Institute. No significant deviation of from historical values was noted.

393

394 **Data processing and quality control of exome sequencing.**

395 **Variant annotation.** Variants were annotated using Variant Effect Predictor³¹ and the
396 LOFTEE⁸ plugin to identify protein-truncating variants predicted to disrupt the respective
397 gene's function with "high confidence." Each allele at polyallelic sites was separately
398 annotated.

399 **Sample level quality control.** We performed quality control of samples using the
400 following steps. For quality control of samples, we used bi-allelic SNVs that passed the
401 GATK VQSR filter and were on genomic regions targeted by both ICE and Agilent
402 exome captures. We removed samples with discordance rate > 10% between genotypes
403 from exome sequencing with genotypes from array-based genotyping and samples with
404 sex mismatch between inbreeding coefficient on chromosome X and fingerprinting. We
405 tested for sample contamination using the verifyBamID software, which examines the
406 proportion of non-reference bases at reference sites, and excluded samples with high
407 estimated contamination (FREEMIX scores > 0.2).³² After removing monozygotic twins
408 or duplicate samples using the KING software³³, we removed outlier samples with too
409 many or too few SNVs (>17,500 total variants for Agilent-captured samples or >18,000
410 for ICE-captured samples; <12,000 total variants; >600 singletons; and >400 doubletons).
411 We removed those with extreme overall transition-to-transversion ratios (>4 or <3) and
412 heterozygosity (heterozygote-to-homozygote ratio >6 or <2). Finally, we removed
413 samples with high missingness (>0.05).

414 **Variant level quality control.** Variant score quality recalibration was performed
415 separately for SNVs and indels use the GATK VariantRecalibrator and
416 ApplyRecalibration to filter out variants with lower accuracy scores. To further reduce

417 the rate of inaccurate variant calls, we further filtered out SNVs with low average quality
418 (quality per depth of coverage (QD) < 2) and a high degree of missingness (> 20 %), and
419 indels also with low average quality (quality per depth of coverage (QD) < 3) and a high
420 degree of missingness (> 20 %).

421

422 **Methods for inbreeding analyses.**

423 **Array-derived runs of homozygosity.** Analyses were conducted in PLINK²³ using
424 genome-wide association (GWAS) data in PROMIS and HapMap 3 populations.
425 Segments of the genome that were at-least 1.5 Mb long, had a SNP density of 1 SNP per
426 20 kb and had 25 consecutive homozygous SNPs (1 heterozygous and/or 5 missing SNPs
427 were permitted within a segment) were defined to be in a homozygous state (or referred
428 as “runs of homozygosity” (ROH)), as described previously.³⁴ Homozygosity was
429 expressed as the percentage of the autosomal genome found in a homozygous state, and
430 was calculated by dividing the sum of ROH length within each individual by the total
431 length of the autosome in PROMIS and HapMap 3 populations respectively. To
432 investigate variability in homozygosity explained by parental consanguinity, the
433 difference in R^2 is reported for a linear regression model of homozygosity including and
434 excluding parental consanguinity on top of age, sex and the first 10 principal components
435 derived from the typed autosomal GWAS data.

436 **Sequencing-derived coefficient of inbreeding.** We compared the coefficient of
437 inbreeding distributions of 7,078 exome sequenced PROMIS participants with 15,248
438 participants (European ancestry = 12,849, and African ancestry = 2,399) who were
439 exome sequenced at the Broad Institute (Cambridge, MA) from the Myocardial Infarction

440 Genetics consortium.²⁵ We extracted approximately 5,000 high-quality polymorphic
441 SNVs in linkage equilibrium present on both target intervals that passed variant quality
442 control metrics based on HapMap 3 data.³⁵ Using PLINK, we estimated the coefficient of
443 inbreeding separately within each ethnicity group.²³ The coefficient of inbreeding was
444 estimated as the observed degree of homozygosity compared with the anticipated
445 homozygosity derived from an estimated common ancestor.³⁶ The Wilcoxon-Mann-
446 Whitney test was used to test whether PROMIS participants had different median
447 coefficients of inbreeding compared to other similarly sequenced outbred individuals and
448 whether the median coefficient of inbreeding was different between PROMIS participants
449 who reported parental relatedness versus not. A two-sided P of 0.05 was the pre-specified
450 threshold for statistical significance.

451

452 **Methods for sequencing projection analysis.**

453 To compare the burden of unique completely inactivated genes in the PROMIS cohort
454 with outbred cohorts of diverse ethnicities, we extracted the minor allele frequencies
455 (maf) of "high confidence" loss-of-function mutations observed in PROMIS, and in
456 European, African, and East Asian ancestry participants from the Exome Aggregation
457 Consortium (ExAC r0.3; exac.broadinstitute.org). For each gene and for each ethnicity,
458 the combined minor allele frequency (cmaf) of rare (maf < 0.1%) "high confidence" loss-
459 of-function mutations was calculated. We then simulated the number of unique
460 completely inactivated genes across a range of sample sizes per ethnicity and PROMIS.
461 The expected probability of observing complete inactivation (two null copies in an
462 individual) of a gene was calculated as $(1 - F) * cmaf^2 + F * cmaf$, which accounts

463 for allozygous and autozygous, respectively, mechanisms for complete genie knockout.
464 F , the inbreeding coefficient, is defined as $F = 1 - (\text{expected heterozygosity rate} /$
465 $\text{observed heterozygosity rate})$. For PROMIS, the median F inbreeding coefficient
466 (0.016) was used for estimation. Down-sampling within the observed sample size for
467 both high-confidence null mutations and synonymous variants did not deviate
468 significantly from the expected trajectory. For a range of sample sizes (0-200,000), each
469 gene was randomly sampled under a binomial distribution ($X \sim B(n, cmaf)$) and it was
470 determined if the gene was successfully sampled at least once. To refine the estimated
471 count of unique genes per sample size, each sampling was replicated ten times.

472

473 **Methods for constraint score analysis.**

474 We sought to determine whether the observed homozygous null genes were under less
475 evolutionary constraint by first obtaining constraint loss of function constraint scores
476 derived from the Exome Aggregation Consortium (Lek M et al, in preparation).^{9,10}
477 Briefly, we used the number of observed and expected rare (MAF < 0.1%) loss of
478 function variants per gene to determine to which of three classes it was likely to belong:
479 null (observed variation matches expectation), recessive (observed variation is ~50%
480 expectation), or haploinsufficient (observed variation is <10% of expectation). The
481 probability of being loss of function intolerant (pLI) of each transcript was defined as the
482 probability of that transcript falling into the haploinsufficient category. Transcripts with a
483 $pLI \geq 0.9$ are considered very likely to be loss of function intolerant; those with $pLI \leq 0.1$
484 are not likely to be loss of function intolerant. A list of 961 genes were randomly sampled
485 from a list of sequenced genes 1,000 times and the proportion of loss of function

486 intolerant genes compared to the proportion of the observed homozygous null genes was
487 compared using the chi square test. The likelihood that the distribution of the test
488 statistics deviated from the null was ascertained.

489

490 **Methods for rare variant association analysis.**

491 **Recessive model association discovery.** We sought to determine whether complete loss-
492 of-function of a gene was associated with a dense array of phenotypes. We extracted a list
493 of individuals per gene who were homozygous for a high confidence null allele that was
494 rare (minor allele frequency < 1 %) in the cohort. From a list of 961 genes where there
495 was at least one participant homozygous null and a list of 201 traits, we initially
496 considered 192,960 gene-trait pairings. To reduce the likelihood of false positives, we
497 only considered gene-trait pairs where there were at least two homozygous nulls per gene
498 phenotyped for a given trait yielding 15,263 gene-trait pairs for analysis.

499 For all analyses, we constructed generalized linear models to test whether complete loss
500 of function versus non-carriers was associated with trait variation. A logit link was used
501 for binomial outcomes. Right-skewed continuous traits were natural log transformed.
502 Age, sex, and myocardial infarction status were used as covariates in all analyses. We
503 extracted principal components of ancestry using EIGENSTRAT to control for
504 population stratification in all analyses.³⁷ For lipoprotein-related traits, the use of lipid-
505 lowering therapy was used as a covariate. For glycemic biomarkers, only non-diabetics
506 were used in the analysis. The P threshold for statistical significance was $0.05 / 15,263 =$
507 3×10^{-6} .

508 **Heterozygote association replication.** We hypothesized that some of the associations
509 for homozygous nulls will display a more modest effect for heterozygous nulls. Thus, the
510 aforementioned analyses were performed comparing heterozygous nulls to non-carriers
511 for the fourteen homozygous null-trait associations that surpassed prespecified statistical
512 significance. A P of $0.05 / 13 = 0.004$ was set for statistical significance for these
513 restricted analyses.

514 **Association for single genic homozygotes.** We performed an exploratory analysis of
515 gene-trait pairs where there was only one phenotyped homozygous null. We performed
516 the above association analyses for genes where there was only one homozygous null
517 phenotyped for a given trait and we focused on those with the most extreme standard Z
518 score statistics ($|Z \text{ score}| > 5$) from the primary association analysis and required that
519 there to also be nominal evidence for association ($P < 0.05$) in heterozygotes as well to
520 maximize confidence in an observed single homozygous null-trait association.

521

522 **Methods for recruitment and phenotyping of an *APOC3* p.Arg19Ter proband and**
523 **relatives.**

524 **Methods for Sanger sequencing.** We collected blood samples from a total of 28
525 subjects, including one of the four *APOC3* p.Arg19Ter homozygous participants along
526 with 27 of his family and community members for DNA extraction and separated into
527 plasma for lipid and apolipoprotein measurements. All subjects were consented prior to
528 initiation of the studies (IRB: 00007048 at the Center for Non-Communicable Diseases,
529 Paksitan). DNA was isolated from whole blood using a reference phenol-chloroform
530 protocol.³⁸ Genotypes for the p.Arg19Ter variant were determined in all 28 participants

531 by Sanger sequencing. A 685 bp region of the *APOC3* gene including the base position
532 for this variant was amplified by PCR (Expand HF PCR Kit, Roche) using the following
533 primer sequences: Forward primer CTCCTTCTGGCAGACCCAGCTAAGG, Reverse
534 primer CCTAGGACTGCTCCGGGGAGAAAG. PCR products were purified with Exo-
535 SAP-IT (Affymetrix) and sequenced via Sanger sequencing using the same primers.

536 **Oral fat tolerance test.** Six non-carriers and seven homozygotes also participated in an
537 oral fat tolerance test. Participants fasted overnight and then blood was drawn for
538 measurement of baseline fasted lipids. Following this, participants were administered an
539 oral load of heavy cream (50 g fat per square meter of body surface area as calculated by
540 the method of Mosteller³⁹). Participants consumed this oral load within a time span of 20
541 minutes and afterwards consumed 200 mL of water. Blood was drawn at 2, 4, and 6 hours
542 after oral fat consumption as done previously.^{40,41} All lipid and apolipoprotein
543 measurements from these plasma samples were determined by immunoturbidimetric
544 assays on an ACE Axcel Chemistry analyzer (Alfa Wasserman). A comparisons of area-
545 under-the curve triglycerides was performed between *APOC3* p.Arg19Ter homozygotes
546 and non-carriers using a two independent sample Student's t test; $P < 0.05$ was
547 considered statistically significant.

548

549 **Tables**

550 **Table 1. Baseline characteristics of exome sequenced study participants.**

Characteristic	Value
	(n = 7,078)
Age (yrs) – mean (sd)	50.7 (9.1)
Women – no. (%)	1,249 (17.6 %)
Parents closely related – no. (%)	2,710 (38.3 %)
Spouse closely related – no. (%)	2,697 (38.1 %)
Ethnicity – no. (%)	
Urdu	2,647 (37.4 %)
Punjabi	2,332 (32.9 %)
Sindhi	797 (11.3 %)
Pathan	416 (5.9 %)
Memon	108 (1.5 %)
Gujrati	89 (1.3 %)
Balochi	85 (1.2 %)
Other	604 (8.5 %)
Hypertension – no. (%)[*]	3,936 (55.6 %)
Hypercholesterolemia – no. (%)[†]	1,857 (26.2 %)
Diabetes mellitus – no. (%)[‡]	2,904 (41.0 %)
Coronary heart disease – no. (%)[§]	3,046 (43.0 %)
Smoking – no. (%)	2,814 (39.8 %)

BMI (m/kg²) – mean (sd)	25.9 (4.3)
---	------------

551 *Hypertension defined as systolic blood pressure \geq 140 mmHg, diastolic blood pressure

552 \geq 90 mmHg, or antihypertensive treatment.

553 †Hypercholesterolemia defined as serum total cholesterol >240 mg/dL, lipid lowering

554 therapy or self-report.

555 ‡Diabetes defined as fasting blood glucose \geq 126 mg/dL, or HbA1c >6.5 %, oral

556 hypoglycemics, insulin treatment, or self-report.

557 §Coronary heart disease defined as history of myocardial infarction as determined by

558 clinical symptoms with typical EKG findings or elevated serum troponin I.

559 ||Smoking defined as active current or prior tobacco smoking.

560

561 **Figure Legends**

562 **Fig 1.** Consanguinity leads to regions of genomic segments that are identical by descent
563 and can be observed as runs of homozygosity. Using genome-wide array data in 17,744
564 PROMIS participants and reference samples from the International HapMap 3, the
565 burden of runs of homozygosity (minimum 1.5 Mb) per individual was derived and
566 population-specific distributions are displayed, with outliers removed. This highlights the
567 higher median runs of homozygosity burden in PROMIS and the higher proportion of
568 individuals with very high burdens.

569

570 **Fig 2.** The site-frequency spectrum of synonymous, missense, and high-confidence null
571 mutations is represented. Points represent the proportion of variants within a 1×10^{-4}
572 minor allele frequency bin for each variant category. Lines represent the cumulative
573 proportions of variants categories. The bottom inset highlights that most null variants are
574 often seen in no more than one or two individuals. The top inset highlights that virtually
575 all null mutations are very rare.

576

577 **Fig 3. a,** The distribution of F inbreeding coefficient of PROMIS participants is
578 compared to those of outbred samples of African (AFR) and European (EUR) ancestry.
579 **b,** The burden of homozygous null genes per individual is correlated with coefficient of
580 inbreeding.

581

582 **Fig 4. a.-d.** Among all sequenced participants, apolipoprotein C-III, triglycerides, HDL
583 cholesterol and LDL cholesterol distributions are displayed by *APOC3* null genotype

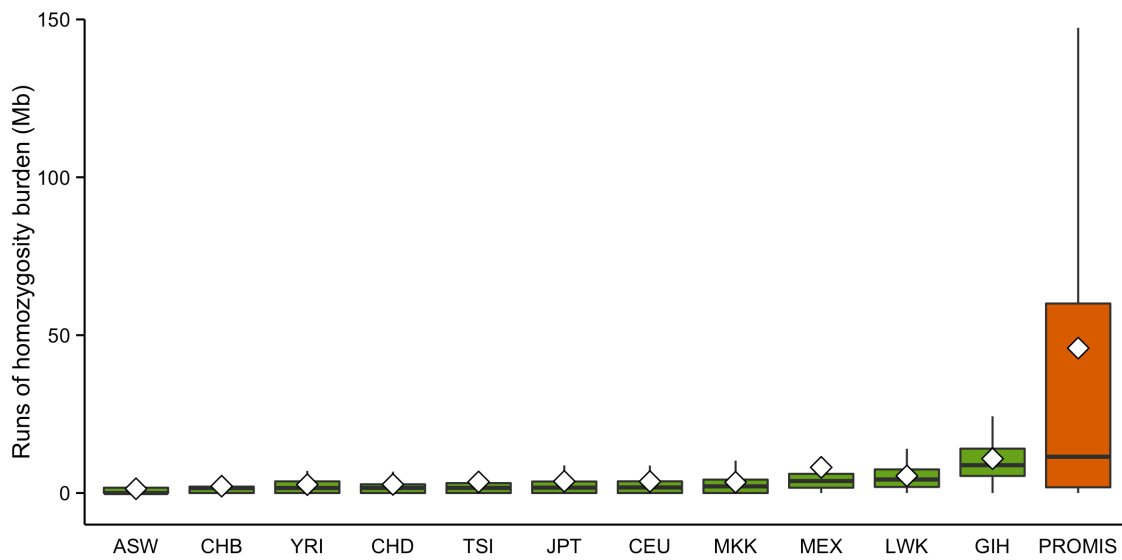
584 status. Apolipoprotein C-III concentration is displayed on a logarithmic base 10 scale. **e.**
585 One of the four *APOC3* null homozygotes and several family members were recruited for
586 genotyping *APOC3* p.Arg19Ter. The proband was married to another null homozygote
587 and has nine obligate homozygote children. Given the extensive first-degree unions, the
588 pedigree is simplified for clarity. **f.** *APOC3* p.Arg19Ter homozygotes and non-carriers
589 within the recruited family participated in an oral 50 g/m² fat tolerance test. Homozygotes
590 had decreased baseline and post-prandial triglyceride concentrations.

591

592 **Fig 5.** The anticipated number of unique homozygous null genes observed with
593 increasing sample sizes sequenced in PROMIS compared with similar African (AFR) and
594 European (EUR) sample sizes using observed allele frequencies and degree of
595 inbreeding.

596 **Figures**

597



598

599

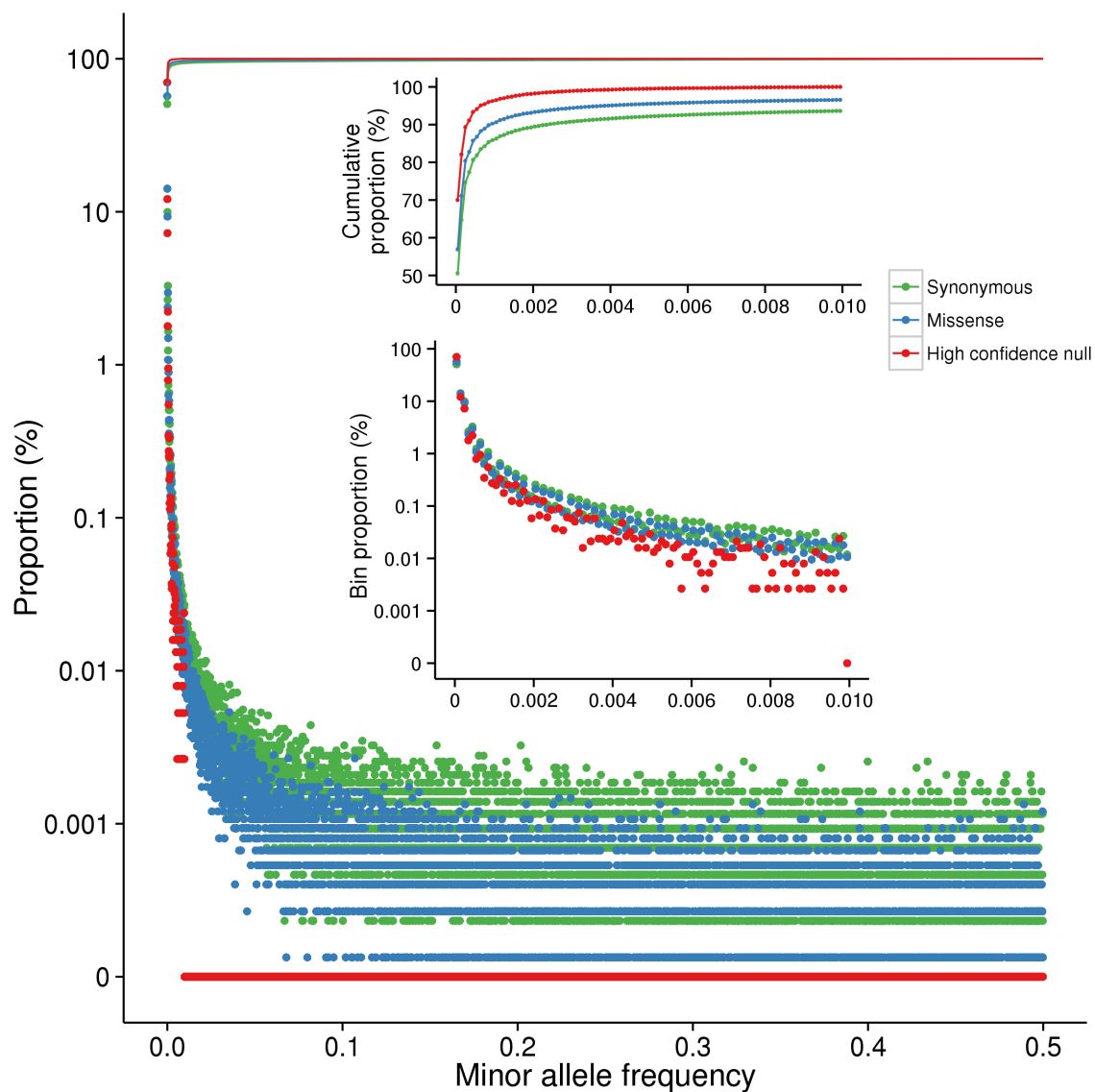
Fig 1. PROMIS participants have an excess burden of runs of homozygosity

600

compared with other populations.

601

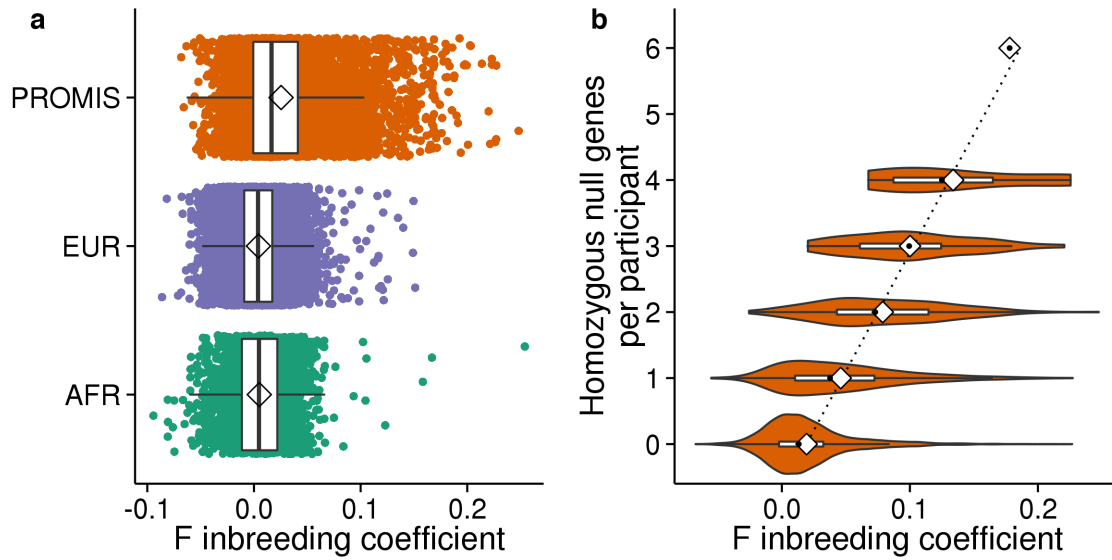
602



603

604 **Fig 2. Null mutations are typically seen in very few individuals.**

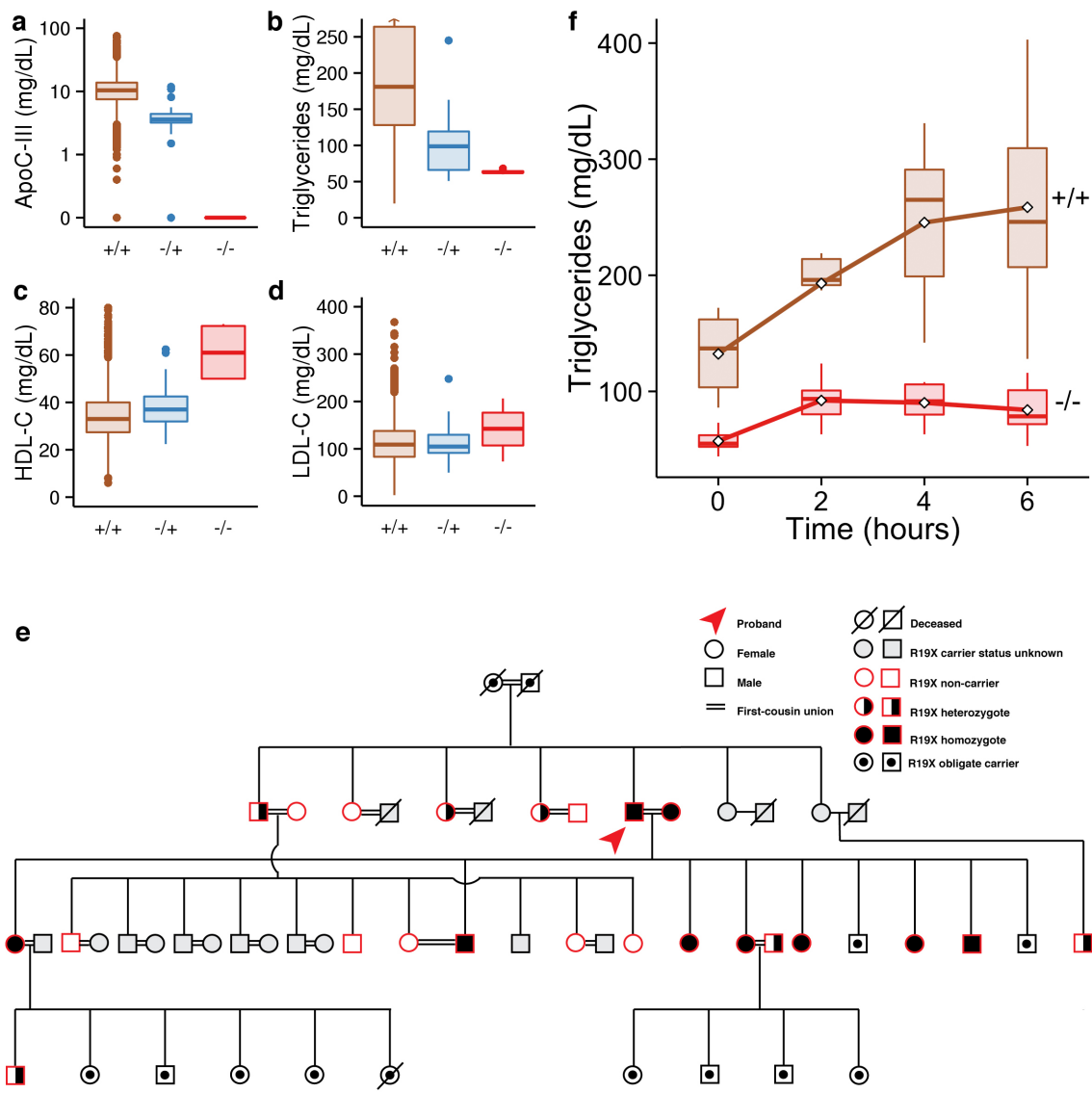
605



606

607 **Fig. 3. Homozygous null burden in PROMIS is driven by excess autozygosity.**

608



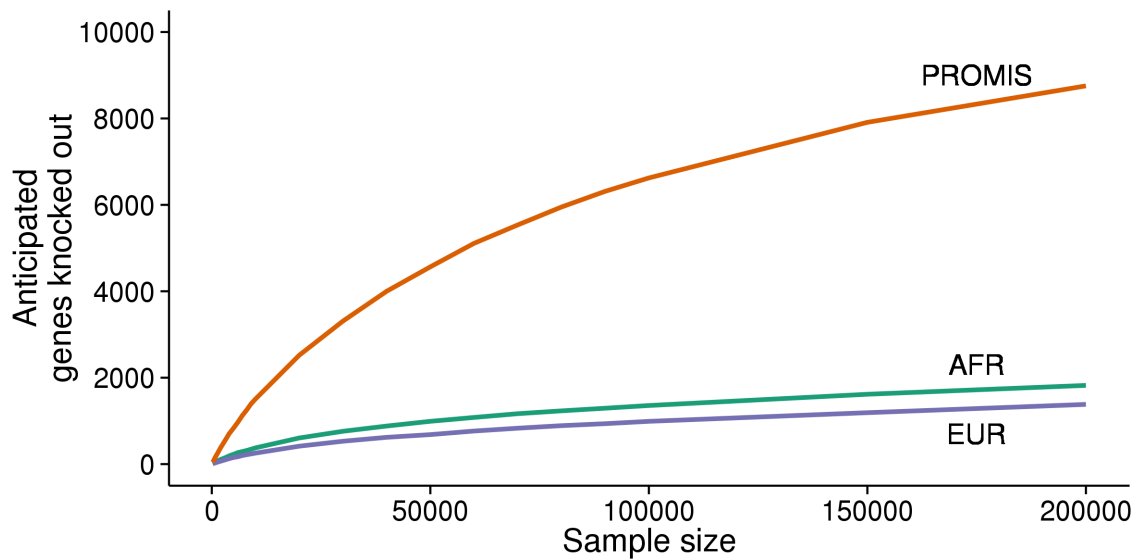
609

610

611 **Fig 4. *APOC3* null homozygotes have diminished fasting triglycerides and blunted**

612 **post-prandial lipemia.**

613



614

615 **Fig 5. Simulations anticipate many more homozygous null genes in the PROMIS**

616 **cohort.**

617 **References**

- 618 1 Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in
619 the post-genomic era. *Nature* **405**, 823-826, doi:10.1038/35015694 (2000).
- 620 2 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human
621 protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).
- 622 3 Sulem, P. *et al.* Identification of a large set of rare complete human knockouts.
623 *Nat Genet*, doi:10.1038/ng.3243 (2015).
- 624 4 Bittles, A. H., Mason, W. M., Greene, J. & Rao, N. A. Reproductive behavior and
625 health in consanguineous marriages. *Science* **252**, 789-794 (1991).
- 626 5 Saleheen, D. *et al.* The Pakistan Risk of Myocardial Infarction Study: a resource
627 for the study of genetic, lifestyle and other determinants of myocardial infarction
628 in South Asia. *European Journal of Epidemiology* **24**, 329-338,
629 doi:10.1007/s10654-009-9334-y (2009).
- 630 6 Modell, B. & Darr, A. Science and society: genetic counselling and customary
631 consanguineous marriage. *Nat Rev Genet* **3**, 225-229, doi:10.1038/nrg754 (2002).
- 632 7 Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human
633 recessive traits with the DNA of inbred children. *Science* **236**, 1567-1570 (1987).
- 634 8 Karczewski, K. J. *LOFTEE (Loss-Of-Function Transcript Effect Estimator)*,
635 <<https://github.com/konradjk/loftee>> (2015).
- 636 9 De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in
637 autism. *Nature* **515**, 209-215, doi:10.1038/nature13772 (2014).
- 638 10 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in
639 human disease. *Nat Genet* **46**, 944-950, doi:10.1038/ng.3050 (2014).

- 640 11 Finn, A. V., Nakano, M., Narula, J., Kolodgie, F. D. & Virmani, R. Concept of
641 vulnerable/unstable plaque. *Arterioscler Thromb Vasc Biol* **30**, 1282-1292,
642 doi:10.1161/ATVBAHA.108.179739 (2010).
- 643 12 Carr, B. A., Wan, J., Hines, R. N. & Yost, G. S. Characterization of the human
644 lung CYP2F1 gene and identification of a novel lung-specific binding motif. *The*
645 *Journal of Biological Chemistry* **278**, 15473-15483, doi:10.1074/jbc.M300319200
646 (2003).
- 647 13 Standiford, T. J. *et al.* Interleukin-8 gene expression by a pulmonary epithelial
648 cell line. A model for cytokine networks in the lung. *The Journal of Clinical*
649 *Investigation* **86**, 1945-1953, doi:10.1172/JCI114928 (1990).
- 650 14 Christiansen, D. *et al.* Humans lack iGb3 due to the absence of functional iGb3-
651 synthase: implications for NKT cell development and transplantation. *PLoS*
652 *Biology* **6**, e172, doi:10.1371/journal.pbio.0060172 (2008).
- 653 15 Dahl, K., Buschard, K., Gram, D. X., d'Apice, A. J. & Hansen, A. K. Glucose
654 intolerance in a xenotransplantation model: studies in alpha-gal knockout mice.
655 *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica* **114**,
656 805-811, doi:10.1111/j.1600-0463.2006.apm_393.x (2006).
- 657 16 Casu, A. *et al.* Insulin secretion and glucose metabolism in alpha 1,3-
658 galactosyltransferase knock-out pigs compared to wild-type pigs.
659 *Xenotransplantation* **17**, 131-139, doi:10.1111/j.1399-3089.2010.00572.x (2010).
- 660 17 Schneider, M. R. & Wolf, E. The epidermal growth factor receptor ligands at a
661 glance. *Journal of Cellular Physiology* **218**, 460-466, doi:10.1002/jcp.21635
662 (2009).

- 663 18 Wang, G. X. *et al.* The brown fat-enriched secreted factor Nrg4 preserves
664 metabolic homeostasis through attenuation of hepatic lipogenesis. *Nature*
665 *Medicine* **20**, 1436-1443, doi:10.1038/nm.3713 (2014).
- 666 19 Huff, M. W. & Hegele, R. A. Apolipoprotein C-III: going back to the future for a
667 lipid drug target. *Circ Res* **112**, 1405-1408,
668 doi:10.1161/CIRCRESAHA.113.301464 (2013).
- 669 20 Tg *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary
670 disease. *N Engl J Med* **371**, 22-31, doi:10.1056/NEJMoa1307095 (2014).
- 671 21 Jorgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjaerg-Hansen,
672 A. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *N*
673 *Engl J Med* **371**, 32-41, doi:10.1056/NEJMoa1308027 (2014).
- 674 22 Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian
675 ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984-
676 989, doi:10.1038/ng.921 (2011).
- 677 23 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-
678 based linkage analyses. *American Journal of Human Genetics* **81**, 559-575,
679 doi:10.1086/519795 (2007).
- 680 24 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation
681 from deep sequencing of human exomes. *Science* **337**, 64-69,
682 doi:10.1126/science.1219240 (2012).
- 683 25 Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles
684 conferring risk for myocardial infarction. *Nature*, doi:10.1038/nature13917
685 (2014).

- 686 26 Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-
687 ready human exome targeted capture libraries. *Genome Biology* **12**, R1,
688 doi:10.1186/gb-2011-12-1-r1 (2011).
- 689 27 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
690 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324
691 (2009).
- 692 28 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
693 analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-
694 1303, doi:10.1101/gr.107524.110 (2010).
- 695 29 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using
696 next-generation DNA sequencing data. *Nat Genet* **43**, 491-498,
697 doi:10.1038/ng.806 (2011).
- 698 30 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls:
699 the Genome Analysis Toolkit best practices pipeline. *Current Protocols in*
700 *Bioinformatics* **11**, 11 10 11-11 10 33, doi:10.1002/0471250953.bi1110s43
701 (2013).
- 702 31 McLaren, W. *et al.* Deriving the consequences of genomic variants with the
703 Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070,
704 doi:10.1093/bioinformatics/btq330 (2010).
- 705 32 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in
706 sequencing and array-based genotype data. *American Journal of Human Genetics*
707 **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

- 708 33 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association
709 studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).
- 710 34 Hunter-Zinck, H. *et al.* Population genetic structure of the people of Qatar.
711 *American Journal of Human Genetics* **87**, 17-25, doi:10.1016/j.ajhg.2010.05.018
712 (2010).
- 713 35 Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in
714 schizophrenia. *Nature* **506**, 185-190, doi:10.1038/nature12975 (2014).
- 715 36 Wright, S. Coefficients of inbreeding and relationship. *Am Nat* **56**, 330-338
716 (1922).
- 717 37 Price, A. L. *et al.* Principal components analysis corrects for stratification in
718 genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847
719 (2006).
- 720 38 Sambrook, J. & Russell, D. W. Purification of nucleic acids by extraction with
721 phenol:chloroform. *CSH protocols* **2006**, doi:10.1101/pdb.prot4455 (2006).
- 722 39 Mosteller, R. D. Simplified calculation of body-surface area. *N Engl J Med* **317**,
723 1098, doi:10.1056/NEJM198710223171717 (1987).
- 724 40 Maraki, M. *et al.* Validity of abbreviated oral fat tolerance tests for assessing
725 postprandial lipemia. *Clinical Nutrition* **30**, 852-857,
726 doi:10.1016/j.clnu.2011.05.003 (2011).
- 727 41 Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma
728 lipid profile and apparent cardioprotection. *Science* **322**, 1702-1705,
729 doi:10.1126/science.1161524 (2008).
- 730

731 **Supplementary Information** is linked to the online version of the paper at

732 www.nature.com/nature.

733

734 **Acknowledgements** Dr. Saleheen is supported by grants from the National Institutes of

735 Health, the Fogarty International, the Wellcome Trust, the British Heart Foundation, and

736 Pfizer. Dr. Natarajan is supported by the John S. LaDue Memorial Fellowship in

737 Cardiology from Harvard Medical School. Dr. Kathiresan is supported by grants from the

738 National Institutes of Health (R01HL107816), the Donovan Family Foundation, and

739 Fondation Leducq. Exome sequencing was supported by a grant from the NHGRI

740 (5U54HG003067-11) to Drs. Gabriel and Lander. Dr. MacArthur is supported by a grant

741 from the National Institutes of Health (R01GM104371). In recognition for PROMIS

742 fieldwork and support, we also acknowledge contributions made by the following:

743 Mohammad Zeeshan Ozair, Usman Ahmed, Abdul Hakeem, Hamza Khalid, Kamran

744 Shahid, Fahad Shuja, Ali Kazmi, Mustafa Qadir Hameed, Naeem Khan, Sadiq Khan,

745 Ayaz Ali, Madad Ali, Saeed Ahmed, Muhammad Waqar Khan, Muhammad Razaq Khan,

746 Abdul Ghafoor, Mir Alam, Riazuddin, Muhammad Irshad Javed, Abdul Ghaffar, Tanveer

747 Baig Mirza, Muhammad Shahid, Jabir Furqan, Muhammad Iqbal Abbasi, Tanveer Abbas,

748 Rana Zulfqar, Muhammad Wajid, Irfan Ali, Muhammad Ikhtlaq, Danish Sheikh,

749 Muhammad Imran, Matthew Walker, Nadeem Sarwar, Sarah Venorman, Robin Young,

750 Adam Butterworth, Hannah Lombardi, Binder Kaur and Nasir Sheikh. Fieldwork in the

751 PROMIS study has been supported through funds available to investigators at the Center

752 for Non-Communicable Diseases, Pakistan and the University of Cambridge, UK.

753

754

755 **Author Contributions** Sample recruitment and phenotyping was performed by D.S.,
756 P.F., J.D., A.R., M.Z., M.S., M.F., A.I., N.K.S., S.A., F.M., M.I., S.A., K.T., N.H.M.,
757 K.S.Z., N.Q., M.I., S.Z.R., F.M., K.M., and N.A.. D.S., P.F., J.D., and W.Z. performed
758 array-based genotyping and runs-of-homozygosity analyses. Exome sequencing was
759 coordinated by D.S., N.G., S.G., E.S.L., D.J.R., and S.K.. P.N., W.Z., H.H.W., and R.D.
760 performed exome sequencing quality control and association analyses. P.N., K.J.K.,
761 A.H.O., and D.G.M. performed variant annotation. D.S., S.K. and D.J.R. performed
762 confirmatory genotyping and lipoprotein biomarker assays. D.S. and A.R. conducted
763 recall based studies for the APOC3 knockouts. P.N. and M.J.D. performed bioinformatics
764 simulations. P.N. and K.E.S. performed constraint score analyses. D.S., P.N., and S.K.
765 designed the study and wrote the paper. D.S. and P.N. contributed equally. All authors
766 discussed the results and commented on the manuscript.

767

768 **Author Information** Summaries of all null variants observed in a homozygous are in the
769 online Supplement. They are additionally, with all observed protein-coding variation,
770 publicly available in the Exome Aggregation Consortium browser
771 (exac.broadinstitute.org). Reprints and permissions information is available at
772 www.nature.com/reprints. The authors do not declare competing financial interests.
773 Correspondence and requests for materials should be addressed to D.S. or S.K.
774 (saleheen@mail.med.upenn.edu or sekar@broadinstitute.org).